University of Bath

UNIVERSITY OF
BATH

PHD

Uncertainty Quantification in Radiative Transport

Parkinson, Matthew

Award date:
2019

Awarding institution:
University of Bath

Link to publication

Uncertainty Quantification in Radiative Transport

submitted by

M.J. Parkinson

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

September 2018

COPYRIGHT

Signature of Author..

M.J. Parkinson

Summary

We study how *uncertainty in the input data* of the Radiative Transport equation (RTE), affects the distribution of (functionals of) its solution (the *output data*).

The RTE is an integro-differential equation, in up to seven independent variables, that models the behaviour of rarefied particles (such as photons and neutrons) in a domain. Its applications include nuclear reactor design, radiation shielding, medical imaging, optical tomography and astrophysics. We focus on the RTE in the context of nuclear reactor physics where, to design and maintain safe reactors, understanding the effects of uncertainty is of great importance.

There are many potential sources of uncertainty within a nuclear reactor. These include the geometry of the reactor, the material composition and reactor wear. Here we consider uncertainty in the macroscopic cross-sections ('the coefficients'), representing them as correlated spatial random fields. We wish to estimate the statistics of a problem-specific quantity of interest (under the influence of the given uncertainty in the cross-sections), which is defined as a functional of the scalar flux. This is the forward problem of Uncertainty Quantification. We seek accurate and efficient methods for estimating these statistics.

Thus far, the research community studying Uncertainty Quantification in radiative transport has focused on the *Polynomial Chaos* expansion. However, it is known that the number of terms in the expansion grows exponentially with respect to the number of stochastic dimensions and the order of the expansion, i.e. polynomial chaos suffers from the *curse of dimensionality*. Instead, we focus our attention on variants of *Monte Carlo sampling* - studying standard and quasi-Monte Carlo methods, and their multilevel and multi-index variants. We show numerically that the quasi-Monte Carlo rules, and the multilevel variance reduction techniques, give substantial gains over the standard Monte Carlo method for a variety of radiative transport problems. Moreover, we report problems in up to 3600 stochastic dimensions, far beyond the capability of polynomial chaos.

A large part of this thesis is focused towards a rigorous proof that the multilevel Monte Carlo method is superior to the standard Monte Carlo method, for the RTE in one spatial and one angular dimension with random cross-sections. This is the *first rigorous theory of Uncertainty Quantification for transport problems* and the *first rigorous theory for Uncertainty Quantification for any PDE problem which accounts for a path-dependent stability condition*. To achieve this result, we first present an error analysis (including a stability bound on the discretisation parameters) for the combined spatial and angular discretisation of the spatially heterogeneous RTE, which is *explicit in the heterogeneous coefficients*. We can then extend this result to prove probabilistic bounds on the error, under assumptions on the statistics of the cross-sections and provided the discretisation satisfies the stability condition pathwise. The multilevel Monte Carlo complexity result follows.

Amongst other novel contributions, we: introduce a method which combines a direct and iterative solver to accelerate the computation of the scalar flux, by adaptively choosing the fastest solver based on the given coefficients; numerically test an iterative eigensolver, which uses a single source iteration within each loop of a shifted inverse power iteration; and propose a novel model for (random) heterogeneity in concrete which generates (piecewise) discontinuous coefficients according to the material type, but where the composition of materials are spatially correlated.

Acknowledgements

Contents

Chapter 1

Introduction

Contents

Fulfilling the global demand for energy by using clean sources of energy is a difficult task for national governments. Whilst the technology and scalability of renewable energy is still in its infancy, one other option is that of nuclear power. Nuclear power plants create heat energy through (ideally self-sustaining) fission chain reactions. This energy is typically converted to electricity by a turbine.

For centuries, mathematicians and physicists have attempted to predict and emulate such real-life physical phenomena through the use of (mathematical) models, often in the form of differential equations. One such example, and the focus of this thesis, is the *Radiative Transport Equation* (RTE), which can closely model the behaviour of rarefied particles (such as neutrons and photons) around a domain [3]. Accurate and efficient computational modelling of the RTE is particularly important due to the difficulty of experimental work [107]. Our particular focus will be on the application of the RTE to nuclear reactor physics. However, we note that the RTE has many other applications including, but not limited to; atmospheric modelling [68], astrophysics in stellar atmospheres [45], optical tomography (e.g. medical imaging) [61, 168] and radiation exposure in aviation [180].

Its ability to reliably model particle behaviour makes the RTE a central tool in the design and operation of nuclear reactors around the world. Within design, industry want to create schematics for reactors that create as much energy as possible, whilst abiding by strict government safety regulations (such as those set out by the Office of Nuclear Regulation in the UK). Similarly, the

operation of existing reactors also concerns itself with maximising energy output whilst constrained by safety regulations.

We will begin this chapter by introducing the standard assumptions that are made within the radiative transport literature, whilst also describing the underlying physics. We will then formally introduce the radiative transport equation itself, including commonly used boundary conditions, and discuss the two main scenarios of interest - the *fixed source problem* and the *criticality problem*. We then describe some possible solution methods for the fixed source problem, many of which are used in the remaining thesis. Furthermore, we outline a sequence of simpler versions of the RTE that will allow us to show that our methodologies work, and also allow us to do some tractable mathematical theory. Finally, we present a summary of the contributions of this thesis.

1.1 The Radiative Transport Equation

The Radiative Transport equation (RTE, also referred to as Neutron Transport) is a linearised version of the Boltzmann equation. It is a physically derived balance equation (as will be seen later in (1.1.6)) which models the flux of rarefied particles in terms of position, angle, energy and time (in total, seven independent variables). There are a number of different formulations of the RTE, including the integral equation and surface-integral forms [72, 176], but often it is the integro-differential equation formulation which is colloquially known as the RTE. To derive the RTE we make the following two assumptions.

Firstly, we assume that particles are modelled as *point particles*, i.e. they are completely determined by their position and velocity. This means that particles *travel in straight lines*, ignoring quantum and electromagnetic effects. For a discussion on why this assumption is reasonable, we refer the reader to [27, §1.1b].

Secondly, we assume that particle-particle interactions are negligible and hence particles can only interact with larger nuclei. This is also a reasonable assumption because the density of the larger nuclei within a reactor is typically many orders of magnitude bigger than the density of the particles. For example, neutron density is around 10^9 per cubic centimetre, compared to a density of 10^{23} per cubic centimetre for the nuclei [202].

Before we introduce the RTE, let us first introduce some notation and discuss the basic underlying physics. At time $t \in \mathbb{R}^+$ (where $\mathbb{R}^+ := \{t \in \mathbb{R} \mid t \geq 0\}$ denotes the non-negative real numbers), consider a particle with position $\mathbf{r} \in \mathcal{D} \subset \mathbb{R}^3$ travelling in the angular direction $\mathbf{\Theta} \in \mathbb{S}_2$ (where $\mathbb{S}_2 := \{v \in \mathbb{R}^3 \mid |v| = 1\}$ denotes the unit sphere) with some kinetic energy $E \in \mathbb{R}^+$. As time evolves, the particle will eventually undergo a *collision event* with a larger nuclei. There are three possible collision events; absorption (sometimes referred to as capture), scatter and fission.

If the particle undergoes absorption, the larger nuclei absorbs the particle and its energy is lost to the system. Simply put, one particle enters the collision event, no particles leave. This is commonly the case when a particle collides with the nuclei of a control rod or if the boundary of the reactor is designed to absorb neutrons. The rate at which a particle with position $\mathbf{r}$ and energy E undergoes an absorption event is denoted by $\sigma_A(\mathbf{r}, E)$, the *absorption cross-section*.

The second possible collision event is scattering. When a particle with direction $\mathbf{\Theta}'$ and energy E' collides with a nucleus at position $\mathbf{r}$ it can 'bounce off' the nucleus. This changes the particle's direction and energy to $\mathbf{\Theta}$ and E respectively. The rate at which a neutron undergoes such an event is given by $\sigma_S(\mathbf{r}, \mathbf{\Theta}', \mathbf{\Theta}, E', E)$, the *scattering cross-section*. Note that the interaction is rotationally invariant and hence σ_S only depends on the cosine between angles, $(\mathbf{\Theta}' \cdot \mathbf{\Theta})$, so the notation $\sigma_S(\mathbf{r}, \mathbf{\Theta}' \cdot \mathbf{\Theta}, E', E)$ is sometimes used. Typically, particles lose energy during a scattering

collision event, and this is known as *downscatter*. In rarer cases the opposite occurs, so-called upscatter [2].

Perhaps the most complex collision event is that of fission. Here, a particle collides with a nucleus and it is absorbed. Unlike an absorption collision event, this causes a fission reaction to occur within the nucleus, which in turn emits a number of new particles. The average number of new particles emitted at a given energy E is given by $\nu(\cdot, E) \in \mathbb{R}^+$. Moreover, the energies of these particles are given by the distribution $\chi(E) \in \mathbb{R}^+$ (i.e. $\int_{\mathbb{R}^+} \chi(E)\mathrm{d}E = 1$). The released direction for each of the new particles is isotropic in angle. For a particle at $(\mathbf{r}, E)$ in phase space, the rate at which a fission event occurs is given by $\sigma_F(\mathbf{r}, E)$, the *fission cross-section*. Note that sometimes particles are not released immediately after a fission event - we will assume that these delayed particles are negligible [27].

A combination of the discussed cross-sections gives us the *total macroscopic cross-section*, $\sigma(\mathbf{r}, E)$, defined by

$$\sigma(\mathbf{r}, E) := \sigma_A(\mathbf{r}, E) + \frac{1}{4\pi} \int_{\mathbb{R}^+} \int_{\mathbb{S}_2} \sigma_S(\mathbf{r}, \mathbf{\Theta}' \cdot \mathbf{\Theta}, E, E') \, \mathrm{d}\mathbf{\Theta}\mathrm{d}E' + \sigma_F(\mathbf{r}, E) \, . \tag{1.1.1}$$

The total cross-section corresponds to the rate at which any collision event will happen at $(\mathbf{r}, E)$. Inversely, this is commonly known as the *mean-free path*, $\tilde{\lambda} := 1/\sigma$, the average distance travelled from one collision to the next.

At this point we would like to make two notes regarding collision events and their cross-sections. Firstly, the collision events outlined above are somewhat simplified. For example, there are different types of scattering collision event, e.g. elastic and inelastic scattering [27, 102]. We will not discuss these further and assume that any given cross-sections account for this.

Secondly, when we use the term 'cross-section' we are in fact talking about the *macroscopic cross-section*, as opposed to the *microscopic cross-section*. The microscopic cross-section is the rate at which a collision event between a single particle and a single nucleus occurs. Its units are cm^2, although sometimes in the physics literature this is stated in barns (1 barn $= 10^{-24}\mathrm{cm}^2$). The microscopic cross-sections are found experimentally and are stored in nuclear data libraries across the world. Examples of such libraries include; JEFF (used in the UK and internationally) [177], CENDL (China) [75] and ENDF/B (USA) [44].

In comparison, the macroscopic cross-sections σ_A, σ_S and σ_F can be interpreted as the rate at which the respective collision events occur (or probabilities when normalised by σ) between a single particle and a cubic centimetre of the nuclei. Consider a cubic centimetre of material consisting of N types of nucleus, with the ith nucleus having microscopic cross-section $\tilde{\sigma}_{\mathrm{col}}^{(i)}$ (for any of the collision events) and constituting w_i (or $100w_i\%$) of the total volume. Then, the macroscopic cross-section σ_{col} for that collision event and composite material is given by:

$$\sigma_{\mathrm{col}} = \sum_{i=1}^{N} w_i A_d^{(i)} \tilde{\sigma}_{\mathrm{col}}^{(i)} \, , \tag{1.1.2}$$

where $A_d^{(i)}$ denotes the atom density of the ith nucleus (with units in *atoms*/cm^3 and defined in [102, (2-1)]). This implies that the units of the macroscopic cross-sections are cm^{-1}.

The final component of the transport equation is that of a *fixed source* (or forcing) term, $f(\mathbf{r}, \mathbf{\Theta}, E, t) \in \mathbb{R}^+$. The source f corresponds to other sources of particles that are not accounted for by a fission event. For example a radioactive isotope, such as californium-252, loaded into guide tubes within the reactor core will emit neutrons through spontaneous fission reactions [137]. Collectively, we will refer to the source f, the cross-sections σ_A, σ_S and σ_F, and the fission

parameters ν and χ, as the *nuclear input data.*

The derivation of the RTE involves understanding (and balancing) changes in the distribution of particles. The starting point is to consider the *angular density* $N(\mathbf{r}, \boldsymbol{\Theta}, E, t)$; the number of particles at time t, per unit volume (around $\mathbf{r}$) and with the given energy E and angle $\boldsymbol{\Theta}$. Typically, we are more interested in the *angular flux* $\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t)$, which relates to the angular density by the following:

$$\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = v(E)N(\mathbf{r}, \boldsymbol{\Theta}, E, t) , \tag{1.1.3}$$

where $v(E)$ is the particle speed given implicitly by $E = \frac{1}{2}mv^2$ (where m denotes particle mass). The derivation involves considering changes over an infinitesimal subset of the seven independent variables, see for example [27, 52]. We will also be interested in the *scalar flux*, $\phi(\mathbf{r}, E, t)$, the angular average of the angular flux ψ, i.e.

$$\phi(\mathbf{r}, E, t) := (\mathcal{W}\psi)(\mathbf{r}, E, t) := \frac{1}{4\pi} \int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) \, \mathrm{d}\boldsymbol{\Theta} . \tag{1.1.4}$$

The full time-dependent RTE is then: Find the angular flux $\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t)$ satisfying

$$
\begin{aligned}
\frac{1}{v}\frac{\partial}{\partial t}\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = & -\left[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r}, E)\right] \psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) \\
& + \frac{1}{4\pi} \int_{\mathbb{R}^+} \int_{\mathbb{S}_2} \sigma_S(\mathbf{r}, \boldsymbol{\Theta}' \cdot \boldsymbol{\Theta}, E', E)\psi(\mathbf{r}, \boldsymbol{\Theta}', E', t) \, \mathrm{d}\boldsymbol{\Theta}'\mathrm{d}E' \\
& + \frac{\chi(E)}{4\pi} \int_{\mathbb{R}^+} \nu(\mathbf{r}, E')\sigma_F(\mathbf{r}, E') \int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta}', E, t) \, \mathrm{d}\boldsymbol{\Theta}'\mathrm{d}E' \\
& + f(\mathbf{r}, \boldsymbol{\Theta}, E, t) ,
\end{aligned}
\tag{1.1.5}
$$

for all $\mathbf{r} \in \mathcal{D}$, $\boldsymbol{\Theta} \in \mathbb{S}_2$, $E \in \mathbb{R}^+$, $t \in \mathbb{R}^+$, where ∇ denotes the gradient with respect to $\mathbf{r}$ and we introduce some appropriate boundary conditions in Section 1.1.1. Simply put, (1.1.5) is equivalent to, for any $\mathbf{r}$, $\boldsymbol{\Theta}$, E and t

$$
\begin{aligned}
\text{Change in particle numbers in } \psi(\mathbf{r}, \boldsymbol{\Theta}, E, \cdot) = & - \text{Loss of particles in } \psi(\mathbf{r}, \boldsymbol{\Theta}, E, \cdot) \\
& + \text{Gain of particles in } \psi(\mathbf{r}, \boldsymbol{\Theta}, E, \cdot) .
\end{aligned}
\tag{1.1.6}
$$

The interpretation (1.1.6) can be taken further. Consider the operator form of (1.1.5), i.e.

$$\frac{1}{v}\frac{\partial}{\partial t}\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = (-\mathcal{T} + \mathcal{S} + \mathcal{F}) \psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) , \tag{1.1.7}$$

where we define

$$\mathcal{T}\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = \left[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r}, E)\right] \psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) , \tag{1.1.8}$$

$$\mathcal{S}\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = \frac{1}{4\pi} \int_{\mathbb{R}^+} \int_{\mathbb{S}_2} \sigma_S(\mathbf{r}, \boldsymbol{\Theta}' \cdot \boldsymbol{\Theta}, E', E)\psi(\mathbf{r}, \boldsymbol{\Theta}', E', t) \, \mathrm{d}\boldsymbol{\Theta}'\mathrm{d}E' , \tag{1.1.9}$$

$$\mathcal{F}\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = \frac{\chi(E)}{4\pi} \int_{\mathbb{R}^+} \nu(\mathbf{r}, E')\sigma_F(\mathbf{r}, E') \int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta}', E, t) \, \mathrm{d}\boldsymbol{\Theta}'\mathrm{d}E' + f(\mathbf{r}, \boldsymbol{\Theta}, E, t) . \tag{1.1.10}$$

The operator $\mathcal{T}$ on the right hand side of (1.1.7) is a loss term. It comprises of $(\boldsymbol{\Theta} \cdot \nabla\psi)$, which is loss due to particle streaming, and $(\sigma\psi)$ which is particle loss from undergoing a collision event. Particles are obviously lost during absorption and fission events (although in the latter, more particles can be released). One viewpoint of scatter can be that a particle is lost, but a single new particle is generated.

The operator $\mathcal{F}$ on the right hand side of (1.1.7) is a gain term, comprising of gains in neutrons from a fission event and/or a fixed source.

The operator $\mathcal{S}$ on the right of (1.1.7) is an altogether more subtle object. For a given $(\mathbf{r}, \boldsymbol{\Theta}, E, t)$ it corresponds to the gain of particles (note the scattering cross-section is integrated over *all incoming* angles and energies). However, unlike $\mathcal{T}$ and $\mathcal{F}$ the overall neutron numbers in the domain $\mathcal{D} \times \mathbb{S}_2 \times \mathbb{R}^+ \times \mathbb{R}^+$ will remain unchanged. This is because, if there is a scattering collision event where particles are lost from $\psi(\mathbf{r}, \boldsymbol{\Theta}', E', \cdot)$, then there must be a $(\boldsymbol{\Theta}, E)$ where $\psi(\mathbf{r}, \boldsymbol{\Theta}, E, \cdot)$ gains a particle.

1.1.1 Boundary Conditions

As with any differential equation, the RTE (1.1.5) requires a suitable (initial or) boundary condition to ensure well-posedness. In radiative transport, the most commonly considered are the *vacuum* and *reflective* boundary conditions, but there are many others including periodic, rotational and white boundary conditions. For a further discussion on the numerical methods relating to these boundary conditions, see for example [175].

The no-inflow (or vacuum) boundary condition assumes that particles cannot be added/enter from outside of the spatial domain $\mathcal{D}$, i.e. there is zero incoming flux at the boundaries. Mathematically this can be stated as

$$\psi(\mathbf{r}, \boldsymbol{\Theta}, \cdot, \cdot) \ = \ 0 \ , \quad \text{for all} \quad \mathbf{r} \in \partial \mathcal{D}_- \ , \tag{1.1.11}$$

where we define the inflow boundary by

$$\mathcal{D}_- \ := \ \{\mathbf{r} \in \partial \mathcal{D} \mid \boldsymbol{\Theta} \cdot \mathbf{n}(\mathbf{r}) \ < \ 0\} \ , \tag{1.1.12}$$

with $\partial \mathcal{D}$ denoting the boundary of $\mathcal{D}$, and where $\mathbf{n}(\mathbf{r})$ denotes the outward normal vector at $\mathbf{r}$.

On the other hand, reflective boundary conditions state that the (incoming) flux at $\mathbf{r} \in \partial \mathcal{D}_-$, for a given direction $\boldsymbol{\Theta}$, is equal to the (outgoing) flux at the same $\mathbf{r} \in \partial \mathcal{D}_-$, in the reflected direction $\boldsymbol{\Theta}' = \boldsymbol{\Theta} - 2 [\boldsymbol{\Theta} \cdot \mathbf{n}(\mathbf{r})] \mathbf{n}(\mathbf{r})$. That is,

$$\psi(\mathbf{r}, \boldsymbol{\Theta}, \cdot, \cdot) \ = \ \psi(\mathbf{r}, \boldsymbol{\Theta}', \cdot, \cdot) \ , \quad \text{for all} \quad \mathbf{r} \in \partial \mathcal{D}_- \ . \tag{1.1.13}$$

We refer the reader to [178, Fig. 1.1] for an illustration and details. In an industrial context, reflective boundary conditions are often useful when there is a geometric symmetry to the reactor [164]. By modelling one half (or subsequently quarters and eighths) of the reactor (rather than the whole reactor) and imposing reflective boundary conditions, the average flux of particles around the whole domain remains the same - but the cost to compute is greatly reduced.

1.1.2 Transport in Equilibrium

Throughout this thesis we will only consider steady-state problems. Such problems arise when we make the following assumption.

Assumption 1.1.1 *There is no variation in the source term or the boundary condition in time, i.e.* $f(\mathbf{r}, \boldsymbol{\Theta}, E, t) = f(\mathbf{r}, \boldsymbol{\Theta}, E)$, *for all* $\mathbf{r} \in \mathcal{D}$, *and* $\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = \psi(\mathbf{r}, \boldsymbol{\Theta}, E)$, *for all* $\mathbf{r} \in \partial \mathcal{D}$.

This assumption ensures that the angular flux becomes time-independent, i.e. $\psi(\mathbf{r}, \boldsymbol{\Theta}, E, t) = \psi(\mathbf{r}, \boldsymbol{\Theta}, E)$, for all $\mathbf{r} \in \mathcal{D}$, and hence

$$\frac{\partial}{\partial t} \psi(\cdot, \cdot, \cdot, t) = 0 .$$

The time independence of ψ implies that the system is in equilibrium.

Under Assumption 1.1.1, the solution to the transport equation (1.1.5) is equivalent to the solution of the *steady state or time-independent radiative transport equation*: Find $\psi(\mathbf{r}, \boldsymbol{\Theta}, E)$ satisfying

$$
\begin{aligned}
\left[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r}, E)\right] \psi(\mathbf{r}, \boldsymbol{\Theta}, E) = {} & \frac{1}{4\pi} \int_{\mathbb{R}^+} \int_{\mathbb{S}_2} \sigma_S(\mathbf{r}, \boldsymbol{\Theta}' \cdot \boldsymbol{\Theta}, E', E) \psi(\mathbf{r}, \boldsymbol{\Theta}', E') \mathrm{d}\boldsymbol{\Theta}' \mathrm{d}E' \\
& + \frac{\chi(E)}{4\pi} \int_{\mathbb{R}^+} \nu(\mathbf{r}, E') \sigma_F(\mathbf{r}, E') \int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta}', E') \mathrm{d}\boldsymbol{\Theta}' \mathrm{d}E' \qquad (1.1.14) \\
& + f(\mathbf{r}, \boldsymbol{\Theta}, E) ,
\end{aligned}
$$

for all $\mathbf{r} \in \mathcal{D}$, $\boldsymbol{\Theta} \in \mathbb{S}_2$ and $E \in \mathbb{R}^+$, and equipped with suitable boundary conditions (as outlined in Section 1.1.1). Subsequently, (1.1.6) becomes *Loss = Gain* and the operator form (1.1.7) can be re-written as

$$\mathcal{T}\psi(\mathbf{r}, \boldsymbol{\Theta}, E) = (\mathcal{S} + \mathcal{F})\psi(\mathbf{r}, \boldsymbol{\Theta}, E) , \qquad (1.1.15)$$

where the application of the operator $\mathcal{T}$ has been brought to the left hand side.

Two of the main scenarios of interest in particle transport are the so-called *fixed source problem* and the *criticality problem*, which we will discuss below.

1.1.3 Criticality Problem

Assume there is no fixed source, i.e. $f \equiv 0$. Our above discussion assumes that particle gain is exactly balanced with particle loss (or in the time dependent case, only changes due to time evolution). However, given a configuration for the reactor (e.g. material concentrations, cross-sections, position of control rods) it is unlikely that we will achieve an exact balance. In mathematical terminology, (1.1.14) (and (1.1.15)) will not have a (non-trivial) solution. We therefore introduce another parameter $\lambda_{\mathrm{crit}} > 0$ to adjust/re-balance (1.1.14) and this leads us to solve the criticality problem: Find the smallest $\lambda_{\mathrm{crit}} \in \mathbb{R}^+$ and $\psi \not\equiv 0$ such that

$$(\mathcal{T} - \mathcal{S})\psi = \lambda_{\mathrm{crit}}\mathcal{F}\psi , \qquad (1.1.16)$$

with some appropriate boundary conditions. Hence, λ_{crit} is the smallest eigenvalue of the generalised eigenvalue problem, with corresponding eigenfunction ψ. We can assume that λ_{crit} is positive and real as explained in the following remark.

Remark 1.1.2 *It is perhaps surprising that we would expect λ_{crit} to be real and positive. The transport operator $\mathcal{T}$ in (1.1.16) is an example of a non self-adjoint differential operator and therefore we would generally expect that the eigenvalue spectrum of (1.1.16) includes complex numbers.*

However, under the assumptions: (1.1.16) is equipped with the no-inflow boundary conditions (1.1.11); the spatial domain $\mathcal{D}$ is convex; and the cross-sections are homogeneous, isotropic in angle and strictly positive; then it can be shown that there exists a unique eigenpair (λ_{crit}, ψ), with $\lambda_{crit} \in \mathbb{R}^+$ and $\psi \not\equiv 0$, satisfying (1.1.16). We refer the reader to [142, 178, 205] and [52, Chap. XXI, §3.5] for further details, and similar results with weaker assumptions.

The goal of the criticality problem (1.1.16) is to find a measure of *how far the system is from balance*. We can see that this measure is given by λ_{crit} by considering the three possible states of (1.1.16):

- **Subcritical** ($\lambda_{\text{crit}} > 1$). More particles are being lost than gained, the system is safe but inefficient;

- **Critical** ($\lambda_{\text{crit}} = 1$). The gain and loss of particles is balanced and hence (1.1.14) is well-posed;

- **Supercritical** ($\lambda_{\text{crit}} < 1$). More particles are being gained than lost, the system is creating large amounts of energy but is unsafe.

Therefore for the non-critical states, a non-zero fixed source/sink of $(1 - \lambda_{\text{crit}}) \times$ fission contribution is needed to keep a self-sustaining reaction and ensure (1.1.14) is well-posed.

The applications of the criticality problem are confined to the design and management of nuclear reactors. Typically, for criticality problems within industry, the literature refers to *"k-effective"* instead of λ_{crit}, where

$$k_{\text{eff}} := \frac{1}{\lambda_{\text{crit}}} . \tag{1.1.17}$$

To summarise, the criticality problem tells us what state a reactor system is in, given its configuration. Adjustments can then be made to ensure a desired balance of safety and efficiency is achieved.

1.1.4   Fixed Source Problem

For the criticality problem, we are primarily interested in the eigenvalue λ_{crit}, rather than the angular flux ψ of the system. The fixed source problem asks an alternative question. Assuming we have a non-zero fixed source and the system (1.1.14) equipped with a suitable boundary condition, then does a solution of the system uniquely exist? And if so, what is the solution?

Remark 1.1.3 *It can be shown that the problem (1.1.14) has a unique solution, when appended with suitable boundary conditions and under weak assumptions on the cross-sections, source and the spatial domain. For further details, see [52, 138, 203].*

In Section 1.2 we will discuss numerical methods that can are used to estimate the solution.

1.1.5   Challenges of the RTE

It is not difficult to imagine why the problem (1.1.14), in six independent variables (or the problem (1.1.5) in seven independent variables), would be very challenging to solve numerically. For accurate solutions to be achieved by discretisation techniques there will be a huge number of associated degrees of freedom [38]. For time-independent problems for example, [185] reports up to 10^{10} and [148] reports 10^{12} total degrees of freedom.

Moreover, the material properties (and subsequently the values of the cross-sections) can change the underlying behaviour of the RTE [113]. In particular, consider the following notes:

- the transport operator $\mathcal{T}$, on the left-hand side of the transport equation (1.1.14), is a linear first-order differential operator, and as such is *hyperbolic*. It is also an example of a *non self-adjoint differential operator*;
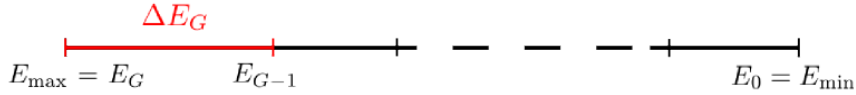
Figure 1-1: Ordering of energy groups from $E_{\max}$ down to $E_{\min}$.

- when σ_S is close to σ (e.g. water) then the RTE is well approximated by the neutron diffusion equation (an *elliptic* PDE) [139]. In fact, this is the reason the diffusion synthetic acceleration (DSA) technique [3, 5, 32] works well in this regime;

- when σ_S is close to σ for the time-dependent RTE, the RTE is well approximated by a time-dependent *parabolic* differential equation [113];

Variations in the cross-sectional values can strongly effect the performance of numerical methods. A prime example of this is the source iteration method [94], which can find a good approximation to the solution to the RTE quickly - unless the system is highly diffusive [178]. We will discuss this further in Section 3.1.2.

1.2 Solution Methods

Now that we have the time-independent RTE (1.1.14) equipped with suitable boundary conditions, we can begin to discuss possible methods for discretising the system and estimating the angular (and scalar) flux. Whilst not the focus of this thesis, we note that to discretise (1.1.5) in time, discrete time steps and finite differences are one option [24, pg.4].

1.2.1 Energy Discretisation

The Multigroup Approximation

The standard method for discretising the RTE over the energy domain is the *multigroup method*. It makes the following assumptions, which we discuss in more detail below:

Assumption 1.2.1 *1. The energy domain is restricted to an interval $[E_{\min}, E_{\max}] \subset \mathbb{R}^+$ (or at least, particles which fall outside this energy range are negligible [27]);*

2. On given sub-intervals of $[E_{\min}, E_{\max}]$, the angular flux can be expressed as the product of a known (piecewise smooth) function in energy with an unknown energy-independent function, see (1.2.2);

3. The source can also be expressed as a similar product of functions, see (1.2.3).

We note that the multigroup method could still be constructed without these assumptions, but it adds substantial complications to the theory, see [27, §1.6*d* and §4] for further discussion.

Consider a strictly decreasing sequence of $G + 1$ points on the energy interval $E_{\max} = E_G > E_{G-1} > \cdots > E_0 = E_{\min}$, which form G groups of size $\Delta E_g = E_g - E_{g-1}$, for $g = 1, \cdots, G$. This is illustrated in Figure 1-1. On each of the energy groups, the multigroup method assigns piecewise constant values (in energy) to the nuclear input data (e.g. cross-sections). We note that cross-sections can have a complex shape in energy (particular in resonance regions, see Figure 1-2) and hence a large G is often required (e.g. [185] reports $G = 150$).
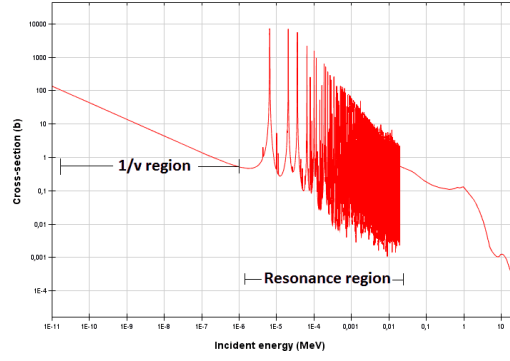
Figure 1-2: Illustrated resonance region for a cross-section of Uranium-238 nuclei, plotted against energy. Taken from [70].

To formulate the multigroup approximation, consider (1.1.14) and re-write $\int_{\mathbb{R}^+} dE' = \sum_{g'=1}^{G} \int_{E_{g'-1}}^{E_{g'}} dE'$. Then integrating (1.1.14) over the interval $[E_{g-1}, E_g]$ gives

$$(\mathbf{\Theta} \cdot \nabla) \int_{E_{g-1}}^{E_g} \psi(\mathbf{r}, \mathbf{\Theta}, E) \mathrm{d}E + \int_{E_{g-1}}^{E_g} \sigma(\mathbf{r}, E) \psi(\mathbf{r}, \mathbf{\Theta}, E) \mathrm{d}E = \int_{E_{g-1}}^{E_g} f(\mathbf{r}, \mathbf{\Theta}, E) \mathrm{d}E \qquad (1.2.1)$$

$$+ \frac{1}{4\pi} \int_{E_{g-1}}^{E_g} \int_{\mathbb{S}^2} \sum_{g'=1}^{G} \int_{E_{g'-1}}^{E_g'} \sigma_S(\mathbf{r}, \mathbf{\Theta}' \cdot \mathbf{\Theta}, E', E) \psi(\mathbf{r}, \mathbf{\Theta}', E') \mathrm{d}E' \mathrm{d}E \mathrm{d}\mathbf{\Theta}'$$

$$+ \int_{E_{g-1}}^{E_g} \frac{\chi(E)}{4\pi} \int_{\mathbb{S}^2} \sum_{g'=1}^{G} \int_{E_{g'-1}}^{E_g'} \nu(\mathbf{r}, E') \sigma_F(\mathbf{r}, E') \psi(\mathbf{r}, \mathbf{\Theta}', E') \mathrm{d}E' \mathrm{d}\mathbf{\Theta}' \mathrm{d}E ,$$

for all $g = 1, \cdots, G$.

Now, Assumption 1.2.1(2.) assumes there exists a *known* and (piecewise smooth) function k_ψ and an unknown function ψ_g (the angular flux on the gth energy group ΔE_g) such that [108]

$$\psi(\mathbf{r}, \mathbf{\Theta}, E) = k_\psi(E) \psi_g(\mathbf{r}, \mathbf{\Theta}) , \quad \text{for all } E_{g-1} \le E \le E_g , \qquad (1.2.2)$$

where k_ψ is normalised such that $\int_{E_{g-1}}^{E_g} k_\psi(E) \mathrm{d}E = 1$, for all $g = 1, \cdots, G$. Likewise, Assumption 1.2.1(3.) assumes there exists a *known* function k_f, normalised such that $\int_{E_{g-1}}^{E_g} k_f(E) \mathrm{d}E = 1$, and f_g (the source on the gth energy group ΔE_g) such that

$$f(\mathbf{r}, \mathbf{\Theta}, E) = k_f(E) f_g(\mathbf{r}, \mathbf{\Theta}) , \quad \text{for all } E_{g-1} \le E \le E_g , \qquad (1.2.3)$$

for all $g = 1, \cdots, G$.

Plugging (1.2.2) and (1.2.3) into (1.2.1) and defining the gth group input data (which can be evaluated as illustrated in [27, §4.5] or [42, pg.11]):

$$\sigma_g(\mathbf{r}) := \int_{E_{g-1}}^{E_g} \sigma(\mathbf{r}, E) k_\psi(E) \, \mathrm{d}E ;$$

$$\sigma_{S,g',g}(\mathbf{r}, \mathbf{\Theta}' \cdot \mathbf{\Theta}) := \int_{E_{g-1}}^{E_g} \int_{E_{g'-1}}^{E_{g'}} \sigma_S(\mathbf{r}, \mathbf{\Theta}' \cdot \mathbf{\Theta}, E', E) k_\psi(E') \, \mathrm{d}E' \mathrm{d}E ; \qquad (1.2.4)$$

$$\chi_g := \int_{E_{g-1}}^{E_g} \chi(E) \, \mathrm{d}E ;$$

$$\nu_{g'}(\mathbf{r}) \sigma_{F,g'}(\mathbf{r}) := \int_{E_{g'-1}}^{E_{g'}} \nu(\mathbf{r}, E') \sigma_F(\mathbf{r}, E') k_\psi(E') \, \mathrm{d}E' ,$$

13

allows us to re-write (1.1.14) as

$$[\boldsymbol{\Theta} \cdot \nabla + \sigma_g(\mathbf{r})]\,\psi_g(\mathbf{r}, \boldsymbol{\Theta}) = \frac{1}{4\pi} \sum_{g'=1}^{G} \int_{\mathbb{S}_2} \sigma_{S,g',g}(\mathbf{r}, \boldsymbol{\Theta}' \cdot \boldsymbol{\Theta})\psi_{g'}(\mathbf{r}, \boldsymbol{\Theta}')\mathrm{d}\boldsymbol{\Theta}'$$

$$+ \frac{\chi_g}{4\pi} \sum_{g'=1}^{G} \nu_{g'}(\mathbf{r})\sigma_{F,g'}(\mathbf{r}) \int_{\mathbb{S}_2} \psi_{g'}(\mathbf{r}, \boldsymbol{\Theta}')\mathrm{d}\boldsymbol{\Theta}' \qquad (1.2.5)$$

$$+ f_g(\mathbf{r}, \boldsymbol{\Theta})\,, \quad \text{for } g = 1, \cdots, G\,.$$

This is a system of equations (with input data $\sigma_g, \sigma_{S,g',g}, \sigma_{F,g'}, \chi_g, \nu_{g'}, f_g$) for the unknowns:

$$\psi_g(\mathbf{r}, \boldsymbol{\Theta}) \;:=\; \int_{E_{g-1}}^{E_g} \psi(\mathbf{r}, \boldsymbol{\Theta}, E')\,\mathrm{d}E'\,, \quad \text{for } g = 1, \cdots, G\,,$$

i.e. the angular flux for the gth energy group. Subsequently the scalar flux for each energy group can be found via:

$$\phi_g(\mathbf{r}) \;:=\; \frac{1}{4\pi} \int_{\mathbb{S}_2} \psi_g(\mathbf{r}, \boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta}\,. \qquad (1.2.6)$$

The boundary conditions (1.1.11), (1.1.13) also hold for the group angular fluxes $\psi_g(\mathbf{r}, \boldsymbol{\Theta})$ in an analogous way.

Throughout this thesis, we will only be considering problems with a single energy group (i.e. $G = 1$, see Assumption 1.3.1(i) below). This is often referred to as the *mono-energetic or the one-speed* transport problem. In this case there is no need to distinguish between different group fluxes (and similar) and we therefore simplify notation by defining $\psi(\mathbf{r}, \boldsymbol{\Theta}) = \psi_g(\mathbf{r}, \boldsymbol{\Theta})$, $\phi(\mathbf{r}) = \phi_g(\mathbf{r})$ and likewise for the group input data (1.2.4), from now on. Moreover, all fission neutrons will have the same energy when released and hence $\chi_1 = 1$. Therefore, for $G = 1$, (1.2.5) becomes

$$[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r})]\,\psi(\mathbf{r}, \boldsymbol{\Theta}) \;=\; \frac{1}{4\pi} \int_{\mathbb{S}_2} \sigma_S(\mathbf{r}, \boldsymbol{\Theta}' \cdot \boldsymbol{\Theta})\psi(\mathbf{r}, \boldsymbol{\Theta}')\,\mathrm{d}\boldsymbol{\Theta}' + \nu(\mathbf{r})\sigma_F(\mathbf{r})\phi(\mathbf{r}) + f(\mathbf{r}, \boldsymbol{\Theta})\,, \quad (1.2.7)$$

where we define

$$\phi(\mathbf{r}) \;:=\; \frac{1}{4\pi} \int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta}\,.$$

1.2.2 Angular Discretisation

We will now consider two possible discretisations with respect to the angular variable $\boldsymbol{\Theta}$, the discrete ordinates (or S_N method) and the spherical harmonics (or P_N method).

Discrete Ordinates (S_N)

Consider the angular variable $\boldsymbol{\Theta} \in \mathbb{S}_2$ in (1.2.5). We can define $\boldsymbol{\Theta}$ in spherical co-ordinates (φ, θ) by

$$\boldsymbol{\Theta}(\theta, \varphi) \;=\; \begin{pmatrix} \boldsymbol{\Theta}^x \\ \boldsymbol{\Theta}^y \\ \boldsymbol{\Theta}^z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}\,, \qquad (1.2.8)$$

where $(\boldsymbol{\Theta}^x, \boldsymbol{\Theta}^y, \boldsymbol{\Theta}^z)$ refer to the direction of the angles in the Cartesian co-ordinate system given by the set of all vectors $\mathbf{r} = (x, y, z)$. The angle $\varphi \in [0, 2\pi]$ represents the azimuthal angle in the xy-plane and $\theta \in [0, \pi]$ is the polar angle in the z-axis. A physical description is given in Figure 1-3 or [27, pg.3].

The discrete ordinates method approximates (1.2.7) by enforcing (1.2.7) to hold for a finite
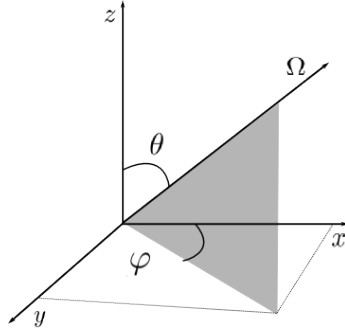
Figure 1-3: Spherical co-ordinates representation of $\boldsymbol{\Theta}$, defined in (1.2.8).

set of N angles $\{\boldsymbol{\Theta}_k\}_{k=1}^N$, where the angular integral is approximated by quadrature. That is, we begin by approximating the scalar flux ϕ by its (N-angle) approximation

$$\phi^N(\mathbf{r}) := \frac{1}{4\pi} \sum_{k'=1}^N w_{k'} \psi_{k'}(\mathbf{r}) . \tag{1.2.9}$$

where we use the notation $\psi_k(\mathbf{r}) = \psi(\mathbf{r}, \boldsymbol{\Theta}_k)$, and where $\{w_k\}$, $\{\boldsymbol{\Theta}_k\}$ denote the weights and directions, respectively, for the specific discrete ordinates method. Then, we have the following system of transport equations along each direction $\boldsymbol{\Theta}_k$:

$$[\boldsymbol{\Theta}_k \cdot \nabla + \sigma(\mathbf{r})] \psi_k(\mathbf{r}) = \sum_{k'=1}^N w_{k'} \frac{\sigma_S(\mathbf{r}, \boldsymbol{\Theta}_{k'} \cdot \boldsymbol{\Theta}_k) + \nu(\mathbf{r})\sigma_F(\mathbf{r})}{4\pi} \psi_{k'}(\mathbf{r}) + f_k(\mathbf{r}) , \tag{1.2.10}$$

for all $k = 1, \cdots, N$, where we have used the notation $f_k(\mathbf{r}) = f(\mathbf{r}, \boldsymbol{\Theta}_k)$.

For an example of the weights w_k and directions $\boldsymbol{\Theta}_k$, for problems (such as (1.2.7)) with angles defined on the sphere $\mathbb{S}_2$, we refer the reader to the T_N quadrature set [200] and the level-symmetric LQ_N set [134, 148].

For problems where the angles are defined on the unit circle, e.g. [110, 10], then one option is the uniformly spaced angles [110, Example 4.1]

$$\boldsymbol{\Theta}_k = \left(\cos \frac{2\pi k}{N}, \sin \frac{2\pi k}{N} \right)^T , \tag{1.2.11}$$

equipped with the (equal) weights $w_k = (2\pi/N)$, for all $k = 1, \cdots, N$.

In 1D problems, where the (cosine of the) angle is defined over the interval $[-1, 1]$, the Gauss-Legendre rules are often used because of their high accuracy (they integrate polynomials of order $2N - 1$ exactly). Moreover, if we consider the *double Gauss rule* (i.e. two separate Gauss-Legendre rules on $[-1, 0)$ and $(0, 1]$ respectively) then the abscissa are well placed to tackle a singularity that arises when the angle becomes perpendicular to the plane (i.e. $\cos \boldsymbol{\Theta}^x = 0$ for a spatially one-dimensional problem in the x-direction). This will be discussed later and is a key feature in Chapter 3 and Chapter 4 of this thesis.

One main disadvantage of the S_N method is the so-called 'ray-effect' [52, 134], i.e. the appearance of unphysical flux oscillations due to the limited number of chosen directions.

Spherical Harmonics (P_N)

One other option for angular discretisation, that goes back to the origins of nuclear reactor theory, is the spherical harmonics (or P_N) method. The P_N method tackles the angular domain by writing

the angular dependence of the flux, cross-sections and source in terms of spherical harmonics [27, §4], [134] of limited degree, and then forces the residual to vanish when integrated against each spherical harmonic.

Spherical harmonics and discrete ordinates have also been combined [68]. Here, discrete ordinates is used in the same way as it is on the left hand side of (1.2.10), but spherical harmonics are used to compute the right hand side of (1.2.10). This method was shown to give good gains over the stand-alone methods.

1.2.3 Spatial Discretisation

Application of the multigroup and discrete ordinates method(s) leads to a system of equations (1.2.10) which have been discretised in energy and angle, but are not discrete in the spatial variable. We now discuss a method for tackling the spatial dimension, the Discontinuous Galerkin finite element method (DG or DG-FEM).

Discontinuous Galerkin

Discontinuous Galerkin (DG) belongs to a family of methods for solving PDEs known as the *finite element methods* (FEM), although strictly speaking it combines ideas from FEM and finite volume schemes. DG was first invented as a solution method for the RTE [166], and it has since been successfully employed across elliptic, hyperbolic and parabolic problems [49, 50, 150, 156]. A unified view and comparison of DG methods for elliptic problems is given in [8] and references therein.

For simplicity, when introducing the DG scheme, we will consider the following problem instead of (1.2.10): For a specific angle $\mathbf{\Theta}_k$, find $\psi(\mathbf{r}, \mathbf{\Theta}_k)$ such that

$$[\mathbf{\Theta}_k \cdot \nabla + \sigma(\mathbf{r})]\, \psi(\mathbf{r}, \mathbf{\Theta}_k) \;=\; f(\mathbf{r}, \mathbf{\Theta}_k)\;, \tag{1.2.12}$$

for all $\mathbf{r}$ in the spatial domain $\mathcal{D} = [0, 1]^3$, and equipped with some appropriate boundary conditions. We note that this problem is similar to the "pure transport problem" that we discuss later in Section 1.3.

We discretise $\mathcal{D} = [0, 1]^3$ by considering the tensor product of 1D meshes $0 = x_0 < x_1 < \cdots < x_{M_x} = 1$, $0 = y_0 < \cdots < y_{M_y} = 1$ and $0 = z_0 < \cdots < z_{M_z} = 1$, where (M_x, M_y, M_z) relates to the refinement of the mesh in each co-ordinate direction. Subsequently, let $\{\mathcal{C}^h\}$ denote a family of (disjoint and open) cuboids D^h (chosen parallel to the axes), with closure $\overline{D^h}$, such that $\overline{\mathcal{D}} = \bigcup_{D^h \in \mathcal{C}^h} \overline{D^h}$, where h denotes the maximum diameter of any cuboid D^h (with boundary ∂D^h, and h depends on the choice of (M_x, M_y, M_z)). Note that we consider cuboids cells D^h for simplicity in this presentation, but many other possibilities exist. These include tetrahedral and hexagonal cells [29, 156], as well as adaptively refined meshes [150].

The first step in constructing our particular numerical method for (1.2.12) can be found by multiplying both sides of (1.2.12) by a test function $v(\mathbf{r})$ (defined on a cuboid D^h), integrating over D^h and then applying Green's identity (integration by parts) to the resulting equation. This gives

$$-\int_{D^h} (\mathbf{\Theta}_k \cdot \nabla v)\psi \; d\mathbf{r} + \int_{D^h} \sigma\psi v \; d\mathbf{r} + \int_{\partial D^h} (\mathbf{\Theta}_k \psi \cdot \mathbf{n})v \; d\mathbf{r} \;=\; \int_{D^h} f_k v \; d\mathbf{r}\;, \tag{1.2.13}$$

for all test functions v, with $\mathbf{n}$ denoting the unit outward normal vector at a cell boundary. Details of this construction are given in Section A.1.

Now, we define the finite element solution space

$$V^h := \{v \in L_2(\mathcal{D}) \mid \forall D^h \in \mathcal{C}^h, \ v\big|_{D^h} \in \mathcal{Q}_K(D^h)\} \ ,$$

where $v\big|_{D^h}$ denotes the restriction of v onto the element D^h and $\mathcal{Q}_K(\cdot)$ denotes the space of polynomials of separate degree K. For the remainder of this discussion we will focus on the case $K = 1$.

Moreover, let $M_f = (2^3)M_x M_y M_z$ denote the total number of degrees of freedom in the 3 spatial dimensions (with each cuboid containing 2^3 degrees of freedom, one in each corner). Then, we define a set of basis functions $\{v_j(\mathbf{r}) \in V^h \mid j = 1, \cdots, M_f\}$ that span the space V^h, such that

$$v_j(\mathbf{r}^{(i)}) := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \ , \quad \text{for } j = 1, \cdots, M_f \ , \qquad (1.2.14)$$

where $\mathbf{r}^{(i)}$ denotes the ith spatial degree of freedom, for $i = 1, \cdots, M_f$. We seek a (piecewise polynomial) function in V^h that is a good approximation to the angular flux for a specific angle (or the scalar flux approximation (1.2.9)). The approximation of the angular flux solution of (1.2.10), for all k, then belongs to $(V^h)^{\otimes N}$, the Nth tensor product space of V^h.

The Discontinuous Galerkin method seeks an approximation of the solution to (1.2.13). That is: Find $\psi_k^{h,N} \in V^h$ such that for all $D^h \in \mathcal{C}^h$,

$$-\int_{D^h} (\boldsymbol{\Theta}_k \cdot \nabla v_j) \psi_k^{h,N} \ d\mathbf{r} + \int_{D^h} \sigma \psi_k^{h,N} v_j \ d\mathbf{r} + \int_{\partial D^h} (\mathbf{F}_k \cdot \mathbf{n}) v_j \ d\mathbf{r} \qquad (1.2.15)$$

$$= \int_{D^h} f_k v_j \ d\mathbf{r} \quad , \text{ for all } k = 1, \cdots, N \ , \text{ and all } v_j \in \mathcal{Q}_1(D^h) \ ,$$

where $\psi_k^{h,N} \approx \psi(\cdot, \boldsymbol{\Theta}_k)$ and where *we choose a numerical flux* $\mathbf{F}_k$ such that $(\mathbf{F}_k \cdot \mathbf{n})$ approximates $(\boldsymbol{\Theta}_k \psi_k^{h,N} \cdot \mathbf{n})$ at the cell boundaries ∂D^h, with $\mathbf{n}$ again denoting the unit outward normal vector at a cell boundary.

The inclusion of the numerical flux is important as, without it, the discontinuities lead $(\boldsymbol{\Theta}_k \psi_k^{h,N} \cdot \mathbf{n})$ to be double-valued at the interior boundaries [96]. Moreover, at the exterior boundaries, the numerical flux ensures that the given boundary conditions are *weakly imposed* - that is, the boundary conditions are satisfied within integral identities. For example, in the case of the no-inflow boundary conditions (1.1.11) and for an element D^h sharing a boundary with $\mathcal{D}_-$ (defined in (1.1.12)), we set $\int_{\partial D_-^h} (\mathbf{F}_k \cdot \mathbf{n}) v_j \ d\mathbf{r} = 0$, where for clarity we re-write the term in (1.2.15) as

$$\int_{\partial D^h} (\mathbf{F}_k \cdot \mathbf{n}) v_j \ d\mathbf{r} = \left(\int_{\partial D_-^h} + \int_{\partial D_+^h} \right) (\mathbf{F}_k \cdot \mathbf{n}) v_j \ d\mathbf{r} = \int_{\partial D_+^h} (\mathbf{F}_k \cdot \mathbf{n}) v_j \ d\mathbf{r} \ , \qquad (1.2.16)$$

where we assume that the angle $\boldsymbol{\Theta}_k$ is not moving directly parallel to an edge of the element D^h, and where we define the inflow and outflow boundaries for an element D^h as (respectively)

$$\partial D_-^h := \{\mathbf{r} \in \partial D^h \mid \boldsymbol{\Theta}_k \cdot \mathbf{n}(\mathbf{r}) < 0\} \quad \text{and} \quad \partial D_+^h := \{\mathbf{r} \in \partial D^h \mid \boldsymbol{\Theta}_k \cdot \mathbf{n}(\mathbf{r}) > 0\} \ . \qquad (1.2.17)$$

Many examples of $\mathbf{F}_k$ are used in the literature, including the (global and local) Lax-Friedrich fluxes [50, 150] and the centered flux [96]. Later, we will apply the so-called *upwind flux* [96] to a variety of spatially two-dimensional applications. That is, we define

$$\mathbf{F}_k(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) := (\boldsymbol{\Theta}_k \cdot \mathbf{n}(\mathbf{r})) \, \psi_k^{h,N,+}(\mathbf{r}) \ , \quad \text{for all } \mathbf{r} \in \partial D_-^h \ , \qquad (1.2.18)$$

for each $k = 1, \cdots, N$, where we also define

$$\psi_k^{h,N,+}(\mathbf{r}) := \lim_{\epsilon \to 0} \psi_k^{h,N}(\mathbf{r} + \epsilon \mathbf{\Theta}_k) , \quad \text{for all } \mathbf{r} \in \partial D_-^h . \tag{1.2.19}$$

The upwind flux is a natural choice for the numerical flux, however so-called *locking*[1] has been observed for systems in optically thick diffusive regimes (i.e. the material is scattering dominated and its mean-free path $\tilde{\lambda} := 1/\sigma$ is small compared with the size of $\mathcal{D}$). We refer the reader to an excellent introduction in [96]. For further discussions, we refer the reader to [49] which gives an excellent survey of DG methods applied to a variety of problems, and to [35] which discusses the stability aspects of certain numerical flux choices.

Finally, to compute a solution to (1.2.15) we write $\psi_k^{h,N}$ as the following basis expansion

$$\psi_k^{h,N} = \sum_{j=1}^{M_f} \Psi_{j,k} v_j \in V^h , \quad \text{for all } k = 1, \cdots, N ,$$

where $\Psi_{j,k}$ is the coefficient corresponding to the jth spatial degree of freedom, for $j = 1, \cdots, M_f$, and the kth angle, for $k = 1, \cdots, N$. We can then solve the resulting linear system to find the vector of coefficients $\mathbf{\Psi} = (\Psi_{j,k})_{j=1,k=1}^{M_f,N}$.

The original analysis of the DG scheme in a two-dimensional spatial domain (for both triangular[2] and rectangular elements) was given in [132]. They considered the pure transport equation (i.e. the right hand side of (1.2.10) is replaced by a generic spatially-dependent function and the resulting equation is studied along a fixed angle, see ahead to Section 1.3), proving $\mathcal{O}(h^K)$ convergence of the angular flux (along the specific direction) in the $L_2(\mathcal{D})$ norm, where h denotes the maximum diameter of the elements. Subsequently for triangular elements, [111] proved $\mathcal{O}(h^{K+1/2})$ convergence in $L_2(\mathcal{D})$, along with similar results on other Lebesgue spaces (but under some uniformity restrictions on the triangulations). The sharpness of this result was shown via a counter example in a later paper [160], although $\mathcal{O}(h^{K+1})$ has been observed in some examples, e.g. [170].

Finite element methods have also been applied to radiative transport in the angular and energy domains [36, 150], instead of the more traditional spatial approach outlined above. This works in much the same way as the spatial DG method, but where the angular and/or energy domain is divided into a mesh, a solution space and piecewise polynomial basis functions are defined, and the weak formulation is constructed (rather than the multigroup method discretisation of the energy domain or the discrete ordinates discretisation of the angular domain, for example). An overall space-angle-energy finite element scheme then considers the tensor product of the individual meshes.

There are also many other methods for angular and spatial discretisation that we have not discussed here. We briefly note a few below;

- reduced-order modelling of angles [38],

- method of characteristics [13] (and short characteristics [195]),

- Crank-Nicolson (or diamond differencing) [5, 93]. We discuss this further in the context of spatially one-dimensional transport problems in Chapter 3,

- corner balance (discontinuous) finite differencing [46].

---

[1]which is defined in [21] by: "a numerical scheme for the approximation of a parameter-dependent problem is said to exhibit *locking* if the accuracy of the approximation deteriorates as the parameter tends to a limiting value."

[2]here, the polynomial space $\mathcal{Q}_K$ is replaced by P_K, which denotes the space of polynomials of *total* degree K

Each of the methods that we have discussed so far, for estimating a solution of the RTE, are *deterministic methods* and require discretisation of the spatial, angular and energy domains.

1.2.4   Analog Monte Carlo Simulation

One further method that does not discretise in space, angle or energy (and does not directly solve the subsequent system of equations) is the analog Monte Carlo (aMC) method [133, §7 − 3]. The idea is to simulate (many) particle histories (i.e. the trajectory of a single particle over time) in a way that replicates the true behaviour of a particle - this includes incorporating random numbers when the behaviour of the particle is statistically uncertain. For this reason, aMC is a standard method used in industry. The method was first devised, in the context of radiative transport, at Los Alamos in 1949 [141]. Since then many developments have been made, of which we discuss a few below. We refer the reader to the texts [127, 192] for further details and [136] for a comprehensive review of acceleration techniques for analog Monte Carlo methods.

The concept of aMC is best explained via an illustrative example. For simplicity assume that we have a homogeneous material over the spatial domain $[0, 1]^3$ and that we are solving a fixed source problem, discussed in Section 1.1.4. Now, consider a single particle with a randomly obtained $\mathbf{r}$, $\boldsymbol{\Theta}$ and E, determined by treating the (normalised) source f over $[0, 1]^3 \times \mathbb{S}_2 \times \mathbb{R}_+$ as a probability density function. A simple example would be a constant source with isotropic scattering and a known energy of the source particles, then we draw $\mathbf{r} \sim \mathcal{U}\left((0, 1)^3\right)$, define $\boldsymbol{\Theta}$ by (1.2.8) with $\theta \sim \mathcal{U}(0, \pi)$, $\varphi \sim \mathcal{U}(0, 2\pi)$, and take E fixed. Here, $\mathcal{U}(\cdot)$ denotes the uniform distribution over a given domain.

The particle now travels in a straight line in the direction $\boldsymbol{\Theta}$ until it undergoes a collision event. The distance the particle travels before the collision event (assuming it does not interact with the boundary) is given by $\mathfrak{d} = -\sigma^{-1} \log y_1$, where $y_1 \sim \mathcal{U}(0, 1)$. A nice justification is given in [178, §4], which uses

$$\mathbb{P}[\text{ distance travelled } \geq \mathfrak{d} \,] \;=\; \exp(-\sigma\mathfrak{d}) \,. \tag{1.2.20}$$

We also refer the reader back to the definition (1.1.1) and the proceeding discussion there. Hence, the new position of the particle is given by

$$\tilde{\mathbf{r}} \;=\; \mathbf{r} - \frac{1}{\sigma} \log y_1 \, \boldsymbol{\Theta} \,. \tag{1.2.21}$$

The type of collision event is determined by splitting the unit interval $[0, 1]$ into disjoint subintervals according to the probabilities of each collision event and then throwing another random number $y_2 \sim \mathcal{U}(0, 1)$ to select a subinterval. That is,

$$\text{Collision Type} \;=\; \begin{cases} \text{fission} & \text{if } y_2 \in \left[0, \frac{\sigma_F}{\sigma}\right) \\ \text{absorption} & \text{else if } y_2 \in \left[\frac{\sigma_F}{\sigma}, \frac{\sigma_A + \sigma_F}{\sigma}\right) \\ \text{scatter} & \text{otherwise} \end{cases} \,. \tag{1.2.22}$$

Once the specific collision event is determined, random numbers can then be generated to establish the latest state of the system. For example, the number of new particles is determined by $\nu(\cdot, E)$ and the corresponding particle energies are distributed according to $\chi(E)$.

Note that in the presence of heterogeneous cross-sections the equation (1.2.21) no longer holds. A number of methods are commonly employed to deal with this issue. One commonly used example is *Woodcock tracking* [207], which artificially homogenises the cross-section(s). This method can be computationally faster (for complex geometries) than other methods (e.g. simply tracking whether

the particle crosses a boundary between materials), but there are issues when there is a large jump in cross-sections between neighbouring materials.

Furthermore, the criticality problem discussed in Section 1.1.3, is more difficult than the fixed source problem, as the source of neutrons relies on the unknown flux (which we are trying to estimate). A common methodology here is to run 'batches' of particle histories and after each batch update the source estimate (until so-called source convergence is achieved [97]). The concept of *superhistory powering* [37] is a particular example.

Extending the simple aMC framework illustrated above allows industrial level codes, such as MCBEND and MONK® , developed by the ANSWERS® software team at the Wood Group, to easily model additional complexities that are not accounted for by (1.1.5), such as delayed neutrons and gamma-rays. Hence, the aMC is often the most accurate technique for transport calculations. However, and as we will see for Monte Carlo sampling later, the convergence in (root-)mean square error (see (2.4.2)) is $\mathcal{O}(n^{-1/2})$, when n particle histories are taken. This is slow and is exacerbated in certain scenarios. For example in shielding calculations, (accurate) tallying at the low-energies is difficult because the particles are often absorbed by the shielding material before they reach low energies [97] and likewise for (accurate) tallying far from the source, for deep penetration problems [97].

Due to the slow convergence of the aMC, a number of so-called variance reduction techniques have been developed to accelerate the convergence. A few of these methods have a similar intuition to that of importance sampling in statistics. The basic idea is to put more onus/effort into simulating (more) particles that are in *more important areas* of the domain (i.e. those with a higher flux) and vice versa. We emphasise that this is not just in space, but in angle and energy also. However, the acquisition of a good importance map requires prior knowledge of the physical processes within the system, e.g. the flux. We refer the reader to the concepts of *Russian roulette* (with splitting) [33, 97] and *source biasing* [33] as examples.

Finally, we emphasise that the aMC method discussed here is *different* to that of the Monte Carlo sampling methods we will discuss later (in Section 2.4.1), with the similarity only arising from the notion of *repeated sampling*.

1.3 Model Problems

As we discussed in Section 1.1.5, finding a solution to (1.1.14) is far from an easy task, even on a supercomputer, due to the coupling over the six independent variables. The mathematical analysis is no easier. Hence, we will consider a sequence of simpler problems that will allow us to show that our methodologies work, and also allow us to do some tractable theory. We emphasise that the following model problems are *not* simplifying approximations of (1.1.14), but are *solutions to* (1.1.14) *under simplifying assumptions on the input data*. This is except for the first assumption that will follow in Assumption 1.3.1, which we already alluded to in Section 1.2.1 and (1.2.5).

We also note that the solution methods discussed in the previous section can easily be applied to any of the problems below, for the most part we will omit the details.

Mono-energetic Radiative Transport

The first model problem we will consider is that of mono-energetic (or one-speed) transport, with isotropic input data. It is given by the following assumptions.

Assumption 1.3.1 *1. ('Mono-energetic') There is only a single energy level E (i.e. $G = 1$);*

2. ('Isotropic input data') *The source f and scattering cross-section σ_S are independent of angle* $\mathbf{\Theta}$.

We remove the notational burden associated with the multigroup approximation, defining $\psi(\mathbf{r}, \mathbf{\Theta}) = \psi_g(\mathbf{r}, \mathbf{\Theta})$ and similarly for the nuclear input data. Under Assumption 1.3.1, then (1.1.14) becomes: Find $\psi(\mathbf{r}, \mathbf{\Theta})$ such that

$$[\mathbf{\Theta} \cdot \nabla + \sigma(\mathbf{r})]\, \psi(\mathbf{r}, \mathbf{\Theta}) \;=\; \frac{\sigma_S(\mathbf{r}) + \nu(\mathbf{r})\sigma_F(\mathbf{r})}{4\pi} \int_{\mathbb{S}_2} \psi(\mathbf{r}, \mathbf{\Theta}')\, \mathrm{d}\mathbf{\Theta}' + f(\mathbf{r}) \;, \qquad (1.3.1)$$

where we note that the isotropic assumption ensures that the coupling in angle is now purely with respect to the angular flux ψ. This formulation also allows us to more easily understand the second quantity that we are interested in, the *scalar flux* $\phi(\mathbf{r})$, defined by

$$\phi(\mathbf{r}) \;:=\; \frac{1}{4\pi} \int_{\mathbb{S}_2} \psi(\mathbf{r}, \mathbf{\Theta}')\, \mathrm{d}\mathbf{\Theta}' \;, \qquad (1.3.2)$$

i.e. the angular average of the angular flux. For flux calculations in a reactor the scalar flux is often more useful - we are typically interested in how many particles are contained with an reactor subdomain, rather than wanting to know the spread over directions.

Radiative Transport in 2D Space

There are two main forms of Radiative Transport in two spatial dimensions, we will refer to them as; the '2D-2D' (or pseudo-3D [150]) problem and the '2D-1D' problem. They are both 2D in the spatial variable, but the '2D-2D' is parameterised by both of the angular variables (θ and φ) in (1.3.1), whereas the '2D-1D' is only parameterised by φ.

Pseudo-3D Radiative Transport

We will show below that the solution to the pseudo-3D transport equation is also the solution to the transport equation in (1.3.1), under the following assumption.

Assumption 1.3.2 *There is no variation in the input data in the z-direction, i.e.*
$\sigma_A(\mathbf{r}) = \sigma_A(x, y)$, $\sigma_S(\mathbf{r}) = \sigma_S(x, y)$, $\sigma_F(\mathbf{r}) = \sigma_F(x, y)$, $\nu(\mathbf{r}) = \nu(x, y)$ *and* $f(\mathbf{r}) = f(x, y)$.

We emphasise this is *not* an approximating assumption. In fact for some parts of the reactor this is often the case, e.g. control rods consist of columns of material and hence have the same cross-sections in the vertical direction.

Now consider an *independent* (to (1.3.1)) spatially 2D problem: Find $\widehat{\psi}^{(2,2)}(x, y, \mathbf{\Gamma})$ such that

$$\left[\mathbf{\Gamma} \cdot \widehat{\nabla} + \sigma(x, y)\right] \widehat{\psi}^{(2,2)}(x, y, \mathbf{\Gamma}) \qquad\qquad\qquad\qquad (1.3.3)$$
$$= \frac{\sigma_S(x, y) + \nu(x, y)\sigma_F(x, y)}{2\pi} \int_{\mathcal{C}_1} \widehat{\psi}^{(2,2)}(x, y, \mathbf{\Gamma}') \left(1 - |\mathbf{\Gamma}'|^2\right)^{-1/2} \, d\mathbf{\Gamma}' \; + \; f(x, y, \mathbf{\Gamma}) \;,$$

with some appropriate boundary conditions, where $\widehat{\nabla} := (\partial/\partial x, \partial/\partial y)$ and $\mathbf{\Gamma}$ is defined by

$$\mathbf{\Gamma}(\theta, \varphi) \;:=\; \begin{pmatrix} \mathbf{\Theta}^x \\ \mathbf{\Theta}^y \end{pmatrix} \;=\; \sin\theta \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \;.$$

The angular domain is the unit disc $\mathcal{C}_1 := \{v \in \mathbb{R}^2 \mid |v| \leq 1\}$. Problems of this form can be seen, for example, in [9] and [191] (with the cylindrical coordinates transport operator, see [27]).

Lemma 1.3.3 *Let* $\widehat{\psi}^{(2,2)}(x,y,\mathbf{\Gamma})$ *be the solution to* (1.3.3) *with some appropriate boundary conditions, and define* ψ *by* $\psi(x,y,z,\mathbf{\Theta}) = \widehat{\psi}^{(2,2)}(x,y,\mathbf{\Gamma})$ *for all* $z \in \mathbb{R}$ *and all* $\Theta_z \in [-1,1]$. *Then, under Assumption 1.3.2 the solution to* (1.3.1) *is* $\psi(x,y,z,\mathbf{\Theta})$.

Proof. First consider the left hand side of (1.3.3), then

$$\left[\mathbf{\Gamma} \cdot \widehat{\nabla} + \sigma(x,y)\right] \widehat{\psi}^{(2,2)}(x,y,\mathbf{\Gamma}) \;=\; [\mathbf{\Theta} \cdot \nabla + \sigma(x,y)]\, \widehat{\psi}^{(2,2)}(x,y,\mathbf{\Gamma}) \;=\; [\mathbf{\Theta} \cdot \nabla + \sigma(\mathbf{r})]\, \psi(\mathbf{r},\mathbf{\Theta})\;,$$

where the first equality holds because $\widehat{\psi}^{(2,2)}$ is independent of z, hence the derivative of $\widehat{\psi}^{(2,2)}$ in the z direction is zero and the choice of Θ_z is irrelevant. The second equality holds due to Assumption 1.3.2 and $\psi = \widehat{\psi}^{(2,2)}$.

Now consider the right hand side of (1.3.1). We are primarily interested in the integral (since it is obvious that $\sigma_S(\mathbf{r}) = \sigma_S(x,y)$ etc.). Since $\sin\theta$ is the Jacobian for spherical co-ordinates, then

$$\int_{\mathbb{S}_2} \psi(\mathbf{r},\mathbf{\Theta}')\, d\mathbf{\Theta}' = \int_0^\pi \int_0^{2\pi} \psi(x,y,z,\sin\theta\cos\varphi,\sin\theta\sin\varphi,\cos\theta)\; \sin\theta\; d\varphi d\theta$$

$$= \int_0^\pi \int_0^{2\pi} \widehat{\psi}^{(2,2)}(x,y,\sin\theta\cos\varphi,\sin\theta\sin\varphi)\; \sin\theta\; d\varphi d\theta\;. \qquad (1.3.4)$$

To integrate this expression, let us first split the integral with respect to θ into the sum of two integrals over $[0,\pi/2]$ and $(\pi/2,\pi]$. Consider the first of these, and make the substitution $r = \sin\theta \in [0,1]$, then

$$dr \;=\; \cos\theta d\theta \;=\; \sqrt{\cos^2\theta}d\theta \;=\; \sqrt{1-\sin^2\theta}d\theta \;=\; \sqrt{1-r^2}d\theta\;,$$

and hence

$$d\theta \;=\; \frac{dr}{\sqrt{1-r^2}}\;.$$

Applying this substitution allows us to re-write (1.3.4) (for $\theta \in [0,\pi/2]$) as

$$\int_0^1 \int_0^{2\pi} \widehat{\psi}^{(2,2)}(x,y,r\cos\varphi,r\sin\varphi)\; r(1-r^2)^{-1/2}\; d\varphi dr\;,$$

and by doing a similar calculation on $(\pi/2,\pi]$, we can combine the two integrals and write

$$\int_0^\pi \int_0^{2\pi} \widehat{\psi}^{(2,2)}(x,y,\sin\theta\cos\varphi,\sin\theta\sin\varphi)\; \sin\theta\; d\varphi d\theta$$

$$= 2\int_0^1 \int_0^{2\pi} \widehat{\psi}^{(2,2)}(x,y,r\cos\varphi,r\sin\varphi)\; r(1-|r|^2)^{-1/2}\; d\varphi dr$$

$$= 2\int_{\mathcal{C}_1} \widehat{\psi}^{(2,2)}(x,y,\mathbf{\Gamma}')\; (1-|\mathbf{\Gamma}'|^2)^{-1/2}\; d\mathbf{\Gamma}'\;,$$

since r is the Jacobian for polar co-ordinates, and we use the definition of $\mathbf{\Gamma}$. This gives us the (integral part of the) right hand side of (1.3.3). Finally, it is fairly simple to show a variety of boundary conditions can also have equivalent representations in this 2D formulation. Hence, the result holds. ∎

2D-1D Radiative Transport

A related problem that has been studied in the literature, e.g. [110, 10, 32], is the two-dimensional radiative transport equation parameterised by a single angle (here referred to as the '2D-1D'
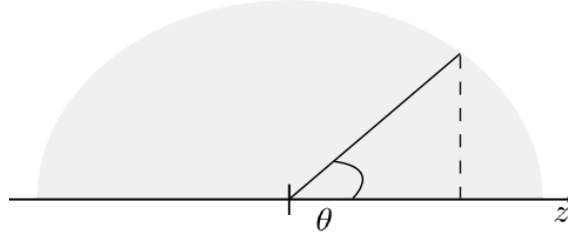
Figure 1-4: Projection of an angle on the unit semi-circle, onto the z-axis.

problem). It is given by: Find $\widehat{\psi}^{(2,1)}(x, y, \boldsymbol{\zeta})$ such that

$$\left[\boldsymbol{\zeta} \cdot \widehat{\nabla} + \sigma(x, y)\right] \widehat{\psi}^{(2,1)}(x, y, \boldsymbol{\zeta}) \;=\; \frac{\sigma_S(x, y) + \nu(x, y)\sigma_F(x, y)}{2\pi} \int_{\mathbb{S}_1} \widehat{\psi}^{(2,1)}(x, y, \boldsymbol{\zeta}') \, d\boldsymbol{\zeta}' \;+\; f(x, y, \boldsymbol{\zeta}) \,,$$
$$(1.3.5)$$

with appropriate boundary conditions, where we define

$$\boldsymbol{\zeta}(\varphi) \;:=\; \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \,,$$

and hence the angular domain of (1.3.5) is the unit circle $\mathbb{S}_1 = \{v \in \mathbb{R}^2 \mid |v| = 1\}$.

1D Slab Geometry

Finally, let us consider the transport equation (1.3.1) one last time. This time we will show that the solution to (1.3.1) is also the solution to a spatially one-dimensional problem, under the following assumption.

Assumption 1.3.4 *There is no variation in the input data in the xy-plane, i.e.*
$\sigma(\mathbf{r}) = \sigma(z)$, $\sigma_S(\mathbf{r}) = \sigma_S(z)$, $\sigma_F(\mathbf{r}) = \sigma_F(z)$, $\nu(\mathbf{r}) = \nu(z)$ *and* $f(\mathbf{r}) = f(z)$.

This assumption is equivalent to assuming that the system is symmetric in the azimuthal angle [53, eq.(5.12)].

Consider the following spatially one-dimensional problem, which is independent of (1.3.1), (1.3.3) and (1.3.5): Find $\widehat{\psi}^{(1,1)}(z, \mu)$ such that

$$\left[\mu \frac{\partial}{\partial z} + \sigma(z)\right] \widehat{\psi}^{(1,1)}(z, \mu) \;=\; \frac{\sigma_S(z) + \nu(z)\sigma_F(z)}{2} \int_{-1}^{1} \widehat{\psi}^{(1,1)}(z, \mu') \, d\mu' \;+\; f(z, \mu) \,, \qquad (1.3.6)$$

with some appropriate boundary conditions, where we define $\mu(\theta) = \boldsymbol{\Theta}^z = \cos\theta$, and hence our angular domain is now $[-1, 1]$. This is equivalent to projecting angles on the unit semi-circle onto the z-axis, as illustrated in Figure 1-4. The equation (1.3.6) is often referred to in the literature as the *1D slab geometry* problem.

Lemma 1.3.5 *Let $\widehat{\psi}^{(1,1)}(z, \mu)$ be the solution to (1.3.6) with some appropriate boundary conditions, and define ψ by $\psi(x, y, z, \boldsymbol{\Theta}) = \widehat{\psi}^{(1,1)}(z, \mu)$, for all $(x, y) \in \mathbb{R}^2$ and all $(\Theta_x, \Theta_y) \in \mathcal{C}_1$ (the unit disc). Then, $\psi(x, y, z, \boldsymbol{\Theta})$ satisfies (1.3.1) under Assumption 1.3.4.*

Proof. First consider the left hand side of (1.3.6). Then,

$$\left[\mu \frac{\partial}{\partial z} + \sigma(z)\right] \widehat{\psi}^{(1,1)}(z, \mu) \;=\; [\boldsymbol{\Theta} \cdot \nabla + \sigma(z)] \widehat{\psi}^{(1,1)}(z, \mu) \;=\; [\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r})] \psi(\mathbf{r}, \boldsymbol{\Theta}) \,,$$

where the first equality holds because $\widehat{\psi}^{(1,1)}$ is independent of x and y and hence the choice of Θ_x and Θ_y is arbitrary (and $\mu = \Theta_z$). The second equality holds because of Assumption 1.3.4 and $\psi = \widehat{\psi}^{(1,1)}$.

For the right hand side of (1.3.6), we are again only really interested in the integral. Applying our assumption gives

$$
\begin{aligned}
\int_{\mathbb{S}_2} \psi(\mathbf{r}, \boldsymbol{\Theta}') \, d\boldsymbol{\Theta}' &= \int_0^\pi \int_0^{2\pi} \psi(x, y, z, \sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta) \, \sin\theta \, d\varphi d\theta \\
&= \int_0^\pi \int_0^{2\pi} \widehat{\psi}^{(1,1)}(z, \cos\theta) \, \sin\theta \, d\varphi d\theta \\
&= -2\pi \int_1^{-1} \widehat{\psi}^{(1,1)}(z, \mu) \, d\mu' \ = \ 2\pi \int_{-1}^1 \widehat{\psi}^{(1,1)}(z, \mu) \, d\mu' \ , \quad\quad (1.3.7)
\end{aligned}
$$

where the third line applies the substitution $\mu = \cos\theta$ (with $d\mu = -\sin\theta$) and then integrates over the (unused) variable φ. Plugging (1.3.7) into the the right hand side of (1.3.1) and using Assumption 1.3.4 gives the right hand side of (1.3.6). ∎

Pure Transport

The final model problem we consider is often referred to as *pure or purely (absorbing) transport*. It is useful for mathematical analysis as it allows us to study the transport operator $\mathcal{T}$, without the additional complexities of angular (and energy) coupling. The pure transport equation is: Find $u(\mathbf{r})$ such that

$$
[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r})] \, u(\mathbf{r}) \ = \ g(\mathbf{r}) \ , \quad\quad (1.3.8)
$$

for all $\mathbf{r} \in \mathcal{D}$, with some suitable boundary conditions. The equation (1.3.8) can be interpreted in two ways, when compared with (1.2.7) - both involve considering the solution for a fixed angle and fixed energy (making them parameters), but they differ in their construction of g: (i) g being a generic right hand side; (ii) system being purely absorbing, i.e. $\sigma_S = \sigma_F = 0$ and $\sigma_A = \sigma$, and hence $g = f$. The analysis holds regardless of the definition.

We prove a closed-form expression for the solution of (1.3.8), subject to the no-inflow boundary conditions, in Section A.2. Subsequently, a useful link between the pure transport problem (1.3.8) and the radiative transport problem (1.3.1) is given in Theorem A.2.1.

We also note that spatially one and two-dimensional versions of the pure transport problem are obvious and we will not construct them here.

1.4 Thesis Contributions and Outline

Within this thesis exist a number of novel contributions. I:

- prove theoretical results on the underlying solution and integral operators of radiative transport in one spatial and one angular dimension, in the case of heterogeneous cross-sections (Section 3.2);

- present an error analysis for the combined spatial and angular discretisation of the spatially heterogeneous radiative transport equation, in one spatial and one angular dimension. The error estimates are explicit in the heterogeneous cross-sections and allow for very low regularity and discontinuities. The discretisation considered is the discrete ordinates method in angle combined with the classical diamond difference scheme in space (Chapter 3);

- prove bounds for the discretisation parameters of the aforementioned discretisation schemes that ensure stability, explicit in the heterogeneous cross-sections (Theorem 3.3.10);

- present an extension of the error estimate to probabilistic bounds, in the case of random cross-sections with certain statistical properties. This includes cross-sections represented by log-normal random fields equipped with the Matérn covariance function (Section 4.1.1);

- theoretically estimate the parameter values in the multilevel Monte Carlo complexity theory (discussed in Section 2.4.3), for a transport equation in one spatial and one angular dimension. This includes the introduction of a sequence of discretisation parameters that are sample dependent in order to ensure stability (Section 4.1.1 - Section 4.1.2);

- present the application of standard and quasi-Monte Carlo, and their multilevel variants, to quantify uncertainty in radiative transport. We apply these techniques to a number of problems, reporting problems in up to 3600 stochastic dimensions (Section 4.3 and Chapter 5);

- present a hybrid direct-iterative algorithm for solving the radiative transport equation in one spatial and one angular dimension (Section 4.2);

- numerically examine an inexact iterative eigensolver which uses a single source iteration within each loop of a (shifted) inverse power iteration. This stops the number of source iterations blowing up, which can occur when the underlying source problem becomes nearly singular due to the variable shifts (Section 5.1.5);

- model the heterogeneity in concrete via Gaussian random fields equipped with a Matérn covariance function. Whilst this isn't novel, the way we model the cross-sections via the random field is - through a sequence of maps. This gives (piecewise) discontinuous cross-sections, where the material composition in the model has spatial correlation according to the Matérn covariance (Section 5.2.5).

Some of these contributions have been presented in the following papers:

- I.G. Graham, M.J. Parkinson, and R. Scheichl. Error Analysis and Uncertainty Quantification for the heterogeneous transport equation in slab geometry. 2018. In preparation for submission, Autumn 2018.

- I.G. Graham, M.J. Parkinson, and R. Scheichl. Modern Monte Carlo variants for Uncertainty Quantification in Neutron Transport. In *Contemporary Computational Mathematics A Celebration of the 80th Birthday of Ian Sloan, J. Dick, F.Y. Kuo, and H. Wozniakowski (Eds.)*, pages 455 – 481. Springer, 2018.

Finally we outline the structure of this thesis. In the next chapter we will introduce the area of Uncertainty Quantification and describe the two key components at its heart - generating realisations of input data and estimating the statistics of the corresponding output data. Firstly, we describe methods to generate realisations of (spatially correlated) random fields, for the nuclear input data. Then, to estimate the statistics of some quantity of interest, we describe Monte Carlo sampling methods (as well as quasi-Monte Carlo, and their multilevel variants).

The third chapter is largely concerned with an error analysis for the combined spatial and angular discretisation of the spatially heterogeneous RTE, in one spatial and one angular dimension. The error estimates are explicit in the heterogeneous coefficients and allow for very low regularity and jumps. We also present a stability bound on the discretisation parameters, explicit in the coefficients. Chapter four then assumes that the cross-sections are random with certain statistical

properties, including the case of log-normal random fields equipped with a Matérn covariance function, and we subsequently prove probabilistic bounds on the error - provided the discretisation satisfies the stability condition pathwise. This leads to novel rigorous complexity estimates for Monte Carlo and multilevel Monte Carlo approaches to uncertainty quantification, that account for the path-dependent stability condition. Moreover, we present a novel hybrid algorithm which combines a direct and iterative solver to accelerate the computation of the solution to the RTE. Numerical results are provided throughout.

In the final chapter we illustrate the techniques described in Chapter 2 to more physically relevant problems in radiative transport (than the spatially one-dimensional fixed source problem considered in Chapters 3 and 4). In particular, we consider a spatially two-dimensional fixed source problem discretised using a Discontinuous Galerkin scheme and discrete ordinates, and a spatially one-dimensional *criticality* problem. We also test an iterative eigensolver which uses a single source iteration within each loop of a shifted inverse power iteration, and propose a model for (random) heterogeneity in concrete - using a sequence of maps combined with a Gaussian random field to generate (piecewise) discontinuous cross-sections, but where the composition of materials in the concrete are spatially correlated.

Chapter 2

Uncertainty Quantification

Contents

The focus of this thesis is on *the theory and application of methods in Uncertainty Quantification* (UQ), to problems in radiative transport. One of the primary aims of UQ is to *accurately and efficiently* estimate the statistical properties of the final state of a model (such as functionals of the solution of a differential equation), given the uncertainty in the input data.

Uncertainty Quantification has many applications outside of radiative transport. Examples include, but are not limited to; finance [41, 80, 106], stochastic differential equations [54, 80], biological systems [167], engineering [16, 40, 67], atmospheric modelling [112], wave phenomena (such as acoustics and seismology) [143, 144] and subsurface groundwater flow [48, 172, 199].

Often uncertainty is grouped into two categories; systematic uncertainty and statistical uncertainty. Systematic uncertainty is 'man-made' uncertainty - in a perfect world there would be none. It is uncertainty that arises because, for example, we have some lack of knowledge of the system or require better measuring equipment. An example in the context of radiative transport are the cross-sections, since nuclear data libraries (such as JEFF [177], CENDL [75] and ENDF/B [44]) report different cross-sectional values for a variety of element/isotope and energy group combinations. On the other hand, statistical uncertainty is inherent to the physical system. For example, the cross-sections relate to the *probability* of each of the possible collision events and so, given the occurrence of a collision event, we cannot say with certainty which will happen (just how likely it is) - recall (1.2.22) and the surrounding discussion.

Whilst the different types of uncertainty are interesting, for our purposes it is not necessary to categorise them in such a way. Instead, we assume that we have a model for the uncertainty in the input data, and that we can generate realisations from this model.

We will begin this chapter by introducing some notation and basic probability theory that are at the heart of Uncertainty Quantification (UQ). We will then introduce the type of UQ problem considered in this text - radiative transport problems with random nuclear input data (i.e. coefficients) represented by spatial (and possibly correlated) random fields. In Section 2.3 we will detail a number of methods used to generate realisations of the random input data, focusing on expansion methods such as the Karhunen-Loéve expansion. Finally, to estimate the statistics of a quantity of interest, we will detail the Monte Carlo (sampling) method and variants thereof, including; quasi-Monte Carlo point sets and multilevel Monte Carlo.

2.1 Random Variables, Random Fields and Moments

To formally describe the random model, we must first present some details of basic probability theory. We will primarily use the texts [95, 116] which cover the discussion below in more detail. For a more radiative transport focussed discussion, we refer the reader to [133]. Two probabilistic quantities of particular importance in this thesis are *random variables* and *random fields*.

Let $(\Omega, \mathcal{G}, \mathbb{P})$ denote a probability space, where $\omega \in \Omega$ denotes an outcome from the sample space Ω, $\mathcal{G}$ denotes a σ-algebra of possible events in Ω, and $\mathbb{P} : \mathcal{G} \mapsto [0, 1]$ denotes an associated probability measure.

A (real) *random variable* Q is a measurable real-valued function $Q : \Omega \mapsto \mathbb{R}$, that assigns numerical values to possible events in the sample space[1]. Subsequently, a *random field* over a given space X is a collection of random variables $\{Q(\cdot, x) \mid x \in X\}$, where the choice of X determines the type of random field, e.g. spatial, temporal, spatio-temporal. The notation Q will interchange between random variables and random fields. Three common examples of random field that we will discuss in this work are the uniform, Gaussian and log-normal random fields.

A random field is defined to be a uniform (respectively, Gaussian) random field on X if, for all $1 \leq n < \infty$ and all combinations $(x_1, \cdots, x_n)$, for $x_1, \cdots, x_n \in X$, the finite collection

$$[Q(\cdot, x_1), \cdots, Q(\cdot, x_n)], \text{ has multivariate uniform (respectively, Gaussian) distribution.} \quad (2.1.1)$$

A random field Q is called log-normal if $\log Q$ is a Gaussian field. Under some technical assumptions that we do not specify, the connection between the random field over a continuum and the finite-dimensional distributions in (2.1.1) are given by Kolmogorovs Extension Theorem [116].

It will turn out that, solutions (and functionals of the solutions) to the radiative transport equation, under the assumed uncertainty in the input data, will be represented by random fields (and random variables) with some unknown underlying distribution(s). *We wish to estimate these distributions.*

An important quantity that is related to the distribution of a random variable Q is its *(statistical) moments* (at least those that exist), where the kth moment is defined by

$$\mathbb{E}[Q^k] := \int_\Omega Q^k(\omega) \, d\mathbb{P}(\omega) = \int_\Omega Q^k(\omega) p(\omega) \, d\omega . \quad (2.1.2)$$

We use $p(\omega)$ to denote the probability density function (pdf) of the random variable Q. Formally,

---

[1] This definition is not rigorous. Strictly speaking, a (real) random variable is defined as a function $Q : \Omega \mapsto \mathbb{R}$, such that for all subsets $M \subset \mathbb{R}$, the pre-image of Q defined by $Q^{-1}(M) := \{\omega \in \Omega \mid Q(\omega) \in M\} \in \mathcal{G}$. We refer the reader to [95, 116] for further details.

the kth moment of a random variable Q exists if $Q \in L_k(\Omega)$, where we define

$$L_k(\Omega) := \left\{ Q : \Omega \mapsto \mathbb{R} \mid \|Q\|_{L_k(\Omega)}^k := \int_\Omega |Q(\omega)|^k \, \mathrm{d}\mathbb{P}(\omega) < \infty \right\} , \qquad (2.1.3)$$

that is $L_k(\Omega)$ denotes the Lebesgue space on Ω with respect to the probability measure $\mathbb{P}$. For random fields, the notion of Lebesgue space(s) extends to the so-called Bochner space(s), $L_k(\Omega; X)$, defined by

$$L_k(\Omega; X) := \left\{ Q : \Omega \mapsto X \mid \|Q\|_{L_k(\Omega;X)}^k = \mathbb{E}\left(\|Q\|_X^k\right) := \int_\Omega \|Q(\omega, \cdot)\|_X^k \, \mathrm{d}\mathbb{P}(\omega) < \infty \right\} , \quad (2.1.4)$$

for some Banach space $(X, \|\cdot\|_X)$. Note the relationship $L_k(\Omega) = L_k(\Omega; \mathbb{R})$.

Whilst the techniques illustrated in this chapter are applicable to estimate all moments (that exist), we will focus on estimating just one - the expected value (the first moment). For a random variable $Q \in L_1(\Omega)$, the expected value $\mathbb{E}[Q]$ is defined by

$$\mathbb{E}[Q] := \int_\Omega Q(\omega) \, \mathrm{d}\mathbb{P}(\omega) < \infty . \qquad (2.1.5)$$

It is important to note that it is not restrictive to focus on estimating $\mathbb{E}[Q]$, and many more complex statistics can be constructed. For example, if we wish to compute the kth moment of Q we can consider another random variable $\xi = Q^k$ and estimate $\mathbb{E}[\xi]$ instead. Similarly, if we want to evaluate variants of moments, e.g. the variance (the centered second moment) defined by

$$\mathbb{V}[Q] := \mathbb{E}\left[(Q - \mathbb{E}[Q])^2\right] = \mathbb{E}[Q^2] - (\mathbb{E}[Q])^2 < \infty , \qquad (2.1.6)$$

where we assume $Q \in L_2(\Omega)$, then we just need to estimate several expected values, i.e. $\mathbb{E}[Q]$ and $\mathbb{E}[\xi]$ with $\xi := Q^2$. In much the same way, we can estimate the pdf of Q by computing a finite number of moments and taking the pdf estimate that maximises the Shannon entropy [23, 31, 34], so-called *moment matching*.

Considering $\xi = \xi(Q, b)$, for another parameter $b \in \mathbb{R}$, allows us to estimate more complex statistics via $\mathbb{E}[\xi]$, e.g. the characteristic function can be estimated by defining $\xi(Q, b) := \exp(ibQ)$, and the cumulative distribution function (cdf) by

$$\mathbb{P}[Q \leq b] = \mathbb{E}[\mathbb{1}_{(-\infty, b]}(Q)] = \mathbb{E}[\xi(Q, b)] , \quad \xi(Q, b) := \mathbb{1}_{(-\infty, b]}(Q) , \qquad (2.1.7)$$

where $\mathbb{1}_S(\cdot)$ denotes the indicator function for some set S.

One exception, which cannot be written simply as the expected value(s) of some quantity, are the quantiles of a distribution. There are a variety of algorithms for computing quantiles, e.g. [66, 118], but we do not deal with the additional technicalities in this text.

2.2 Uncertainty Quantification

As we have already mentioned, the goal of Uncertainty Quantification is to understand how uncertainty in the input data affects the output data of a physical model. In this thesis we will focus on the effects of uncertainty in the nuclear input data on (functionals of) the scalar flux ϕ. Such functionals $\mathfrak{L}$ will produce *quantities of interest* $Q(\omega) \in \mathbb{R}$ by the relationship $Q(\omega) = \mathfrak{L}(\phi(\omega, \cdot))$, for $\omega \in \Omega$. We will assume throughout that $Q \in L_2(\Omega)$, i.e. Q is a random variable with finite variance. We seek an accurate and efficient estimator of $\mathbb{E}[Q]$.

Let us assume that the nuclear input data; $\sigma_S = \sigma_S(\omega, \cdot)$, $\sigma_A = \sigma_A(\omega, \cdot)$, $\sigma_F = \sigma_F(\omega, \cdot)$, $\nu = \nu(\omega, \cdot)$ and $f = f(\omega, \cdot)$ are (possibly dependent or correlated) random fields belonging to some space X, that is problem specific. For notational simplicity we group the input data into a vector of random fields

$$\mathbf{Z}(\omega, \cdot) := [\sigma_S(\omega, \cdot), \sigma_A(\omega, \cdot), \sigma_F(\omega, \cdot), \nu(\omega, \cdot), f(\omega, \cdot)] . \tag{2.2.1}$$

Then, the radiative transport problems introduced in Chapter 1 become integro-differential equations with random coefficients, and subsequently the solution(s) ψ and ϕ (and Q) become random fields (random variables) themselves - we seek their distribution. For example, the transport problem (1.3.6) becomes: Find $\psi(\omega, x, \mu)$ such that

$$\left[\mu\frac{\partial}{\partial x} + \sigma(\omega, x)\right]\psi(\omega, x, \mu) = [\sigma_S(\omega, x) + \nu(\omega, x)\sigma_F(\omega, x)]\phi(\omega, x) + f(\omega, x, \mu) , \tag{2.2.2}$$

for $x \in [0, 1]$, $\mu \in [-1, 1]$, where

$$\phi(\omega, x) := \frac{1}{2}\int_{-1}^{1} \psi(\omega, x, \mu')\,\mathrm{d}\mu' ,$$

with $\sigma(\omega, \cdot) = \sigma_S(\omega, \cdot) + \sigma_A(\omega, \cdot) + \sigma_F(\omega, \cdot)$ and where we satisfy suitable boundary conditions, for almost all realisations $\omega \in \Omega$. The extension of (2.2.2) to higher dimensional problems are obvious.

We note that, as well as being dependent on $\omega \in \Omega$, the scalar flux ϕ (and ψ) is a function of the nuclear input data $\mathbf{Z} = \mathbf{Z}(\omega, \cdot)$. Likewise, $Q = \mathfrak{L}(\phi)$ is a function of $\mathbf{Z}$. We will abuse notation throughout this thesis by considering $Q(\omega)$ and $Q(\mathbf{Z}) = Q(\mathbf{Z}(\omega, \cdot))$ as equivalent, for $\omega \in \Omega$ (also see ahead to Notation 2.3.1).

2.3 Generating Random Fields

Now that we have discussed the underlying probabilistic framework, we can begin to understand how uncertainty propagates through the RTE. As we have already mentioned, we will assume that we have a known model for the uncertainty - we will focus on the case where the uncertainty is represented by a spatial random field. Hence, our first step is to find methods for generating realisations of a random field $\mathbf{Z}$ over the spatial domain $\mathcal{D}$. For simplicity, we will assume that $\mathbf{Z}$ contains only a single random field (unlike the five random fields in (2.2.1)). For further details on this section, we refer the reader to [135, 194].

One of the simplest ways to generate a realisation of a spatial random field is to simulate a random variable, according to some distribution, at each spatial point. For example, drawing from the uniform distribution

$$\mathbf{Z}(\cdot, \mathbf{r}) \sim \mathcal{U}[m - \epsilon, m + \epsilon] , \quad \text{for all} \ \ \mathbf{r} \in \mathcal{D} , \tag{2.3.1}$$

could be an *uncorrelated* model accounting for measurement error ϵ around a mean m. If we draw i.i.d. from a Gaussian distribution, this is typically referred to as *white noise*.

Whilst the model in (2.3.1) is interesting, a more physically relevant model would typically incorporate correlation between spatial points. A simple extension of (2.3.1) is to have perfect correlation on certain (non-overlapping) subdomains of $\mathcal{D}$. That is, consider a sequence of open subdomains $D_i \subset \mathcal{D}$ such that $\overline{\mathcal{D}} = \bigcup_i \overline{D_i}$ and $D_i \cap D_j = \emptyset$, for all $i \neq j$. Then,

$$\text{for each} \ \ D_i , \ \ \text{and for all} \ \mathbf{r} \in D_i , \quad \mathbf{Z}(\cdot, \mathbf{r}) \sim \mathcal{U}[m - \epsilon, m + \epsilon] . \tag{2.3.2}$$
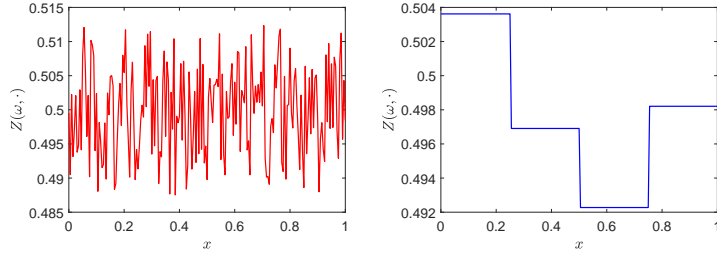
Figure 2-1: Examples of (2.3.1) and (2.3.2) on $\mathcal{D} = [0,1]$, on the left and right respectively, with $m = 0.5$ and $\epsilon = 0.025$. The disjoint domains are $D_1 = [0, 0.25)$, $D_2 = [0.25, 0.5)$, $D_3 = [0.5, 0.75)$ and $D_4 = [0.75, 1]$ for (2.3.2).

Such a piecewise constant model will be used in the numerical results of Section 5.1.6. In Figure 2-1 we give an example of a realisation of (2.3.1) and (2.3.2) in one spatial dimension, with $m = 0.5$ and $\epsilon = 0.025$.

For the rest of this section, we will discuss more advanced models for generating realisations of correlated random fields. Such models rely on the concept of *covariance*, which for two random variables Z_1 and Z_2, is defined by

$$\mathrm{Cov}\,(Z_1, Z_2) \; := \; \mathbb{E}\left[(Z_1 - \mathbb{E}[Z_1])\,(Z_2 - \mathbb{E}[Z_2])\right] \; , \tag{2.3.3}$$

e.g. $Z_1 = \mathbf{Z}(\mathbf{r}_1)$ and $Z_2 = \mathbf{Z}(\mathbf{r}_2)$, for two spatial points $\mathbf{r}_1$ and $\mathbf{r}_2$. The choice of covariance function is dependent upon the problem, but due to (2.3.3) it has to be symmetric positive (semi-)definite. We note that correlation and covariance are proportional to one another, i.e.

$$\mathrm{Corr}(Z_1, Z_2) \;=\; \frac{\mathrm{Cov}\,(Z_1, Z_2)}{\sqrt{\mathbb{V}[Z_1]\mathbb{V}[Z_2]}} \; .$$

2.3.1   Covariance Functions

We will now introduce a family of covariance functions that will be the focus of this thesis. They are known as the Matérn class of covariances and are defined by the function

$$C_\nu(\mathbf{r}_1, \mathbf{r}_2) = \sigma_{var}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(2\sqrt{\nu}\frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\lambda_C}\right)^\nu K_\nu\left(2\sqrt{\nu}\frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\lambda_C}\right) \; , \tag{2.3.4}$$

for two points $\mathbf{r}_1, \mathbf{r}_2 \in \mathcal{D}$ and for some norm $\|\cdot\|$, e.g. the standard ℓ_1 or ℓ_2-norms [47, 172]. The class is parametrised by the smoothness parameter $\nu \geq 0.5$; λ_C is the correlation length, σ_{var}^2 is the variance, Γ is the gamma function and K_ν is the modified Bessel function of the second kind. Note that intuitively, the correlation length λ_C corresponds to the length scale at which two points are correlated and the variance σ_{var}^2 is related to the amplitude of the covariance function. We also note that any Matérn covariance is symmetric (strictly) positive definite, by Bochners theorem. Moreover, we note that it is well known that a Gaussian random field equipped with a Matérn covariance function, such as (2.3.4), is $\lceil\nu\rceil - 1$ times differentiable with respect to the distance $\|\mathbf{r}_1 - \mathbf{r}_2\|$ - this is easy to see by recalling the differentiability of the modified Bessel function of the second kind.

The limiting case of (2.3.4), i.e. when $\nu \to \infty$, corresponds to the Gaussian covariance function (sometimes referred to as the double exponential covariance in the literature), i.e.

$$C_\infty(\mathbf{r}_1, \mathbf{r}_2) \;=\; \sigma_{var}^2 \exp(-\|\mathbf{r}_1 - \mathbf{r}_2\|^2/\lambda_C^2) \; . \tag{2.3.5}$$

The other special case of the Matérn covariance arises when $\nu = 0.5$. This corresponds to the exponential covariance function

$$C_{1/2}(\mathbf{r}_1, \mathbf{r}_2) \;=\; \sigma_{var}^2 \exp(-\|\mathbf{r}_1 - \mathbf{r}_2\|/\lambda_C) \; . \qquad (2.3.6)$$

Other examples of covariance function include the spherical covariance [194] and the rational quadratic covariance [165], but they are not considered in this work.

2.3.2 Karhunen-Loève Expansion

One of the most popular techniques for generating realisations of correlated random fields is to write the random field as an infinite expansion, separating the stochastic dimension(s) from the deterministic dimension(s). There are many available options here, but the methods we will consider differ in their choice of the deterministic component, with the random component taken to be a sequence of (pseudo-)random numbers (distributed according to the choice of random field). An example of such an expansion method, that is extensively used in the literature [60, 78, 93, 123, 181], is the Karhunen-Loève (KL) expansion. The KL expansion takes, as the deterministic component, the (weighted) eigenfunctions of the integral operator with the covariance as its kernel (see ahead to (2.3.8)).

For simplicity, assume each realisation $\mathbf{Z}(\omega, \cdot)$ of a random field (equipped with a continuous covariance function C_ν) is a function on a bounded domain $X \subset \mathbb{R}$. Then by assuming $\mathbf{Z} \in L_2(\Omega, L_2(X))$ (recall (2.1.4)), we can use the KL expansion to sample from $\mathbf{Z}$ by writing

$$\mathbf{Z}(\omega, x) \;=\; \bar{\mathbf{Z}}(x) \;+\; \sum_{i=1}^{\infty} \sqrt{\xi_i}\, \eta_i(x)\, y_i(\omega) \; , \qquad (2.3.7)$$

for all $x \in X$, where $\bar{\mathbf{Z}}(x)$ denotes the mean value of the field at $x \in X$ and $\{y_i\}$ denotes a set of pairwise uncorrelated random variables with mean zero and unit variance [78, 181]. In the special case of a Gaussian random field, $y_i \overset{iid}{\sim} \mathcal{N}(0,1)$, but for all other cases (e.g. $y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$) they are not necessarily independent. The ξ_i and η_i are the eigenvalues and the $L_2(X)$-orthogonal eigenfunctions of the integral operator with kernel given by the chosen covariance function C_ν, e.g. those proposed in Section 2.3.1. That is, we find ξ_i and η_i such that

$$\xi_i \eta_i(\cdot) \;=\; \int_X C_\nu(\cdot, x) \eta_i(x)\, \mathrm{d}x \; , \qquad \text{for } i = 1, 2, \cdots , \qquad (2.3.8)$$

which is an eigenvalue problem, with close relations to the Fredholm integral equation of the second kind. We note that the eigenfunctions are known to be $L_2(X)$-orthogonal by the spectral theorem for a self-adjoint operator.

Notation 2.3.1 *As noted in Section 2.2 the (scalar) flux solution of a specific RTE problem, with random coefficients, is $\phi(\omega, \mathbf{r})$ (for all $\mathbf{r} \in \mathcal{D}$). We can also write this as $\phi(\mathbf{y}(\omega), \mathbf{r})$ to indicate that ϕ obtains its randomness from the d random variables $y_1, \cdots, y_d$ (through the expansion (2.3.7)).*

Likewise, the quantity of interest is defined by

$$Q(\omega) \;:=\; \mathfrak{L}(\phi(\omega, \cdot)) \; ,$$

where $\mathfrak{L}$ is a functional acting on the spatial variable in ϕ. Therefore, we can also write

$$Q(\mathbf{y}(\omega)) \;=\; \mathfrak{L}(\phi(\mathbf{y}(\omega), \cdot)) \; .$$
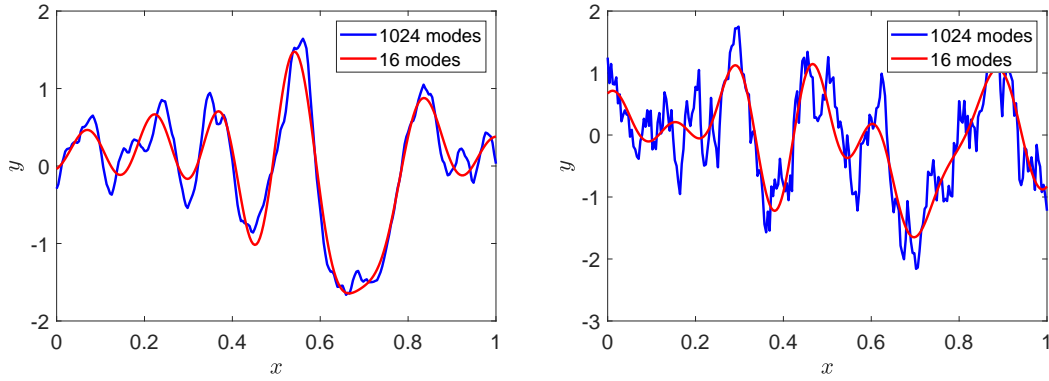
Figure 2-2: Two examples of a Gaussian random field on $[0, 1]$, equipped with Matérn covariance and generated by the truncated KL expansion on a mesh with size $1/256$. The parameters used are $\lambda_C = 0.1$, $\sigma^2_{var} = 1$, $y_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ and; (Left) $\nu = 1.5$, (Right) $\nu = 0.5$.

Moreover, we can consider the expected value of Q as a (deterministic) function of $Q(\mathbf{y})$. For example, if $y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ for all $i = 1, \cdots, d$, then we can write

$$\mathbb{E}[Q] = \frac{1}{2\sqrt{3}} \int_{-\sqrt{3}}^{\sqrt{3}} Q(\mathbf{y}) \; \mathrm{d}\mathbf{y} \; .$$

In practice, the KL expansion needs to be truncated after a finite number of terms, denoted here by d. Hence, (2.3.7) is approximated by

$$\mathbf{Z}(\omega, x) \approx \bar{\mathbf{Z}}(x) + \sum_{i=1}^{d} \sqrt{\xi_i} \; \eta_i(x) \; y_i(\omega) \; . \tag{2.3.9}$$

The accuracy of this truncation depends on the decay of the eigenvalues [135]. In the case of the Matérn class of covariances introduced in Section 2.3.1, it is known that (for $\nu < \infty$) the ith largest eigenvalue satisfies [135]

$$\xi_i \leq c \; i^{-\left(1 + (2\nu/s^{\mathrm{det}})\right)} \; , \tag{2.3.10}$$

for some constant $c > 0$, where s^{det} denotes the dimension of the deterministic field (i.e. $X \subset \mathbb{R}^{s^{\mathrm{det}}}$) and ν is again the smoothness parameter introduced in (2.3.4). That is, for $\nu < \infty$ the decay is algebraic and depends on the smoothness parameter ν. In the case $\nu = \infty$ the decay is exponential.

For the Matérn covariance with $\nu = 0.5$, and for $X \subset \mathbb{R}$, the eigenvalues and eigenfunctions can be estimated numerically as the solution to a transcendental equation [78, 135]. For other choices of X and ν, the eigenpairs can be computed using the Nyström method [155] - see, for example, [62][2]. However, for problems where $X \subset \mathbb{R}^{s^{\mathrm{det}}}$, for $s^{\mathrm{det}} \geq 2$, the cost of generating the KL expansion (particularly finding the eigenpairs via the Nyström method) can be very expensive. There are some methods which reduce the cost of finding the eigenpair, for example a Krylov subspace eigensolver can be accelerated by using a fast multipole method [181], or by using H-matrix techniques [115].

The KL expansion can be used for any given covariance function. However, as the focus of this thesis is on the Matérn class of covariances, we will colloquially refer to the KL expansion for a random field with Matérn covariance as simply the KL expansion. In Figure 2-2, we present two samples of a Gaussian random field on $X = [0, 1] \subset \mathbb{R}$ equipped with different Matérn covariances,

---

[2]We thank Elisabeth Ullmann for allowing us access to her code in `Matlab`

and generated using the KL expansion. In both cases we illustrate the difference between the same sample of a random field, using two different truncation parameters d.

Artificial Covariance Expansion

An alternative method for generating realisations of a random field, on $X \subset \mathbb{R}^2$, was recently introduced in [63, 64, 60]. Here the $\{\eta_i(\cdot)\}$ are chosen to be cosine functions defined on X and the $\{\xi_i\}$ are assumed to (asymptotically) decay like (2.3.10), up to a constant. These artificial choices allow us to generate realisations of a random field, but without any substantial costs - such as those that arise from computing the eigenpairs for the KL expansion.

To sample from $\mathbf{Z}$ at $\mathbf{r} = (r_1, r_2) \in X \subset \mathbb{R}^2$, we can write

$$\mathbf{Z}(\omega, \mathbf{r}) = \overline{Z}(\mathbf{r}) + \sum_{i=1}^{\infty} \sqrt{\xi_i} \cos\left(2\pi\rho_i^1 r_1\right) \cos\left(2\pi\rho_i^2 r_2\right) y_i(\omega) , \tag{2.3.11}$$

where for $i = 1, 2, \cdots$, we define

$$\xi_i := \sigma_{var}^2 \frac{B_i}{\sum_{j=1}^{\infty} B_j} , \qquad B_i := \begin{cases} 1 , & \text{when } i \leq \lambda_C^{-1} \\ \left(i - \frac{1}{\lambda_C}\right)^{-(1+\nu)} , & \text{when } i > \lambda_C^{-1} \end{cases} ,$$

$$\tau_i := \left\lceil -\frac{1}{2} + \sqrt{\frac{1}{4} + 2i} \right\rceil , \qquad \rho_i^1 := i - \frac{1}{2}\tau_i\left(\tau_i + 1\right) , \qquad \rho_i^2 := \tau_i - \rho_i^1 ,$$

with parameters λ_C, σ_{var}^2, ν, and where $\{y_i\}$ denotes a set of pairwise uncorrelated random variables with mean zero and unit variance, e.g. $y_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. We also recall the abuse of notation mentioned in Notation 2.3.1. In Figure 2-3 we give a numerical comparison of the $\{\xi_i\}$ in the KL expansion and the $\{\xi_i\}$ in (2.3.11), where we note that the $\{\xi_i\}$ in (2.3.11) are good approximations to the eigenvalues of the KL expansion when $\lambda_C \approx 1$.

It is useful to observe that $\tau_i \in \mathbb{Z}$, and hence ρ_i^1, $\rho_i^2 \in \mathbb{Z}$, for all $i = 1, 2, \cdots$.

For the remainder of this thesis, all spatially two-dimensional problems will be restricted to the spatial domain $\mathcal{D} = [0, 1]^2$, and hence for simplicity in the proceeding discussion we will restrict to this case now. We note that simple extensions of the comments made below can also be made for many other spatial domains $\mathcal{D}$.

Of course, as with the KL expansion we also have to truncate the expansion (2.3.11) to $d < \infty$ modes. Throughout this work we refer to (2.3.11), truncated to d modes, as the *Artificial Covariance* (AC) expansion. This is to reflect the fact that the $L_2\left[(0, 1)^2\right]$-orthogonality[3] of $\left\{\cos\left(2\pi\rho^1 r_1\right) \cos\left(2\pi\rho^2 r_2\right)\right\}_{\rho^1, \rho^2 = 0}^{\infty}$ implies (2.3.11) is actually the KL expansion (2.3.9) of a random field with an *unknown (artificial) covariance function*. The unknown covariance has eigenfunctions $\eta_i(\mathbf{r}) = \cos\left(2\pi\rho_i^1 r_1\right) \cos\left(2\pi\rho_i^2 r_2\right)$ and eigenvalues ξ_i.

Remark 2.3.2 *We will refer to all random fields for which we use the AC expansion to generate realisations, as random fields equipped with the artificial covariance function.*

Moreover, note that the expansion in (2.3.11) induces a symmetrical structure on $[0, 1]^2$ (with vertical and horizontal lines of symmetry originating from the midpoint of $[0, 1]$ in the x and y axes respectively), because $\cos\left(2m\pi z\right) = \cos\left(2m\pi(1 - z)\right)$, for any $m \in \mathbb{N}$ and any $z \in [0, 1]$. Heterogeneity is unlikely to be symmetric and so to remedy this, and acquire the AC expansion

---

[3] since we can integrate over the r_1 and r_2 components separately and it is well known that $\{\cos\left(2m\pi x\right)\}_{m=0}^{\infty}$ are $L_2(0, 1)$-orthogonal, for $x \in (0, 1)$

(a) $\lambda_C = 1$, $\nu = 1.5$

(b) $\lambda_C = 1$, $\nu = 0.5$

(c) $\lambda_C = 0.25$, $\nu = 1.5$

(d) $\lambda_C = 0.25$, $\nu = 0.5$
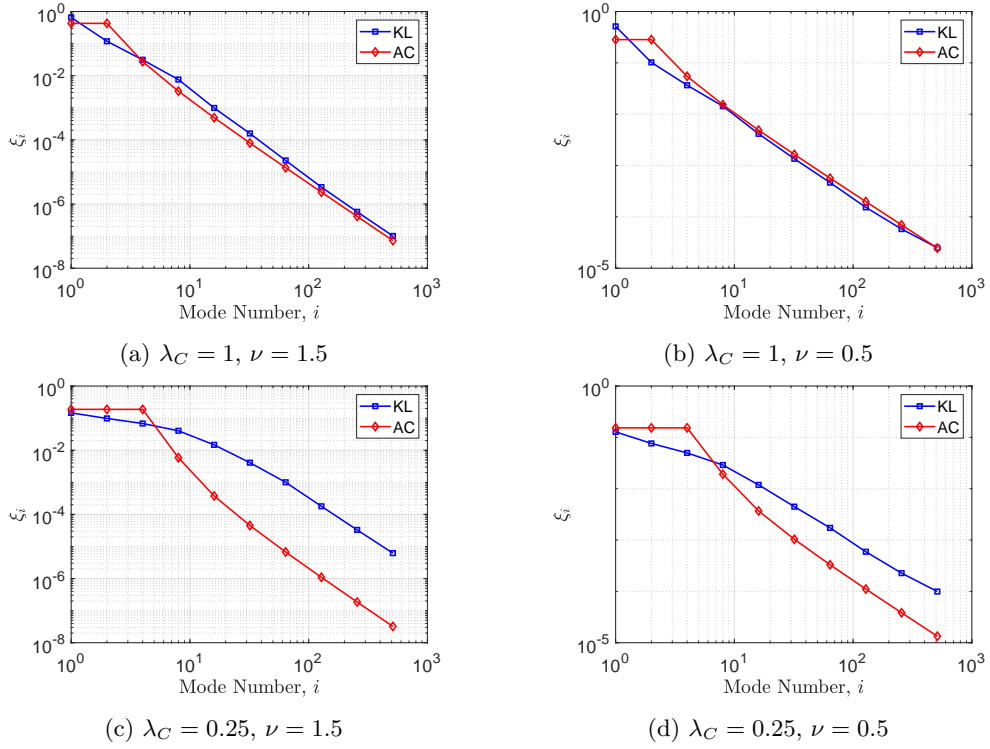
Figure 2-3: A numerical comparison of the $\{\xi_i\}$ for the KL expansion (2.3.9) and the AC expansion (2.3.11), with $\sigma_{var}^2 = 1$ and differing λ_C and ν parameters.

that we will consider within this work, we remove the factor of two from the cosine functions (which still upholds the orthogonality[4] in $L_2\left[(0,1)^2\right]$) to acquire the truncated and non-symmetric AC expansion on $[0,1]^2$:

$$\mathbf{Z}(\omega, \mathbf{r}) \approx \overline{Z}(\mathbf{r}) + \sum_{i=1}^{d} \sqrt{\xi_i} \cos\left(\pi \rho_i^1 r_1\right) \cos\left(\pi \rho_i^2 r_2\right) y_i(\omega) \ . \tag{2.3.12}$$

In Figure 2-4 we present a comparison of samples of a random field, generated using the expansion (2.3.11) and the non-symmetric alternative (2.3.12).

2.3.3 Circulant Embedding

Another popular method for generating realisations of a random field is *Circulant Embedding* [90, 91, 135]. In general, circulant embedding reduces the computational cost of the following method.

Let $\tilde{C}_\nu$ denote the $M \times M$ covariance matrix (the covariance function evaluated at all pairs of M chosen discrete points on X), and factorise $\tilde{C}_\nu$ (i.e. by Cholesky decomposition, which exists because the covariance matrix is symmetric positive semi-definite) into the product of an $M \times M$ lower triangular matrix L_M, and its transpose L_M^T, i.e.

$$\tilde{C}_\nu = L_M L_M^T \ . \tag{2.3.13}$$

Then, we can compute a realisation $\mathbf{Z}(\omega, \cdot)$ of a random field by considering the product of L_M

---

[4]by orthogonality of $\{\cos(m\pi x)\}_{m=0}^{\infty}$
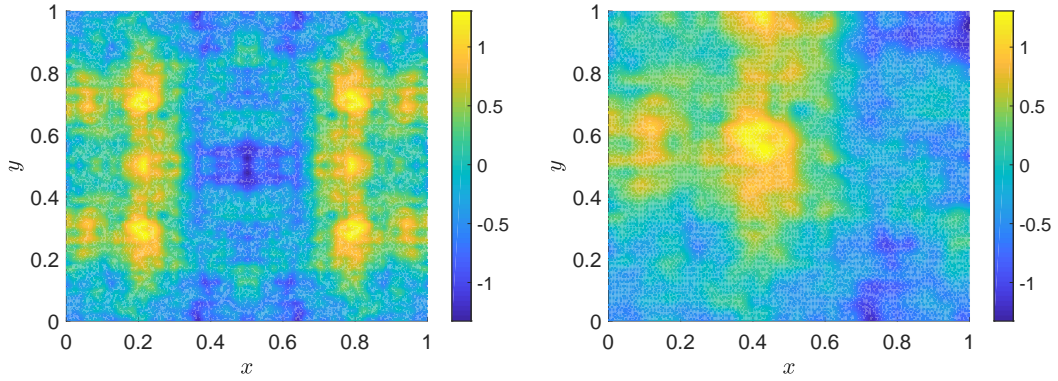
Figure 2-4: Samples of a Gaussian random field on $[0,1]^2$ equipped with the artificial covariance (see Remark 2.3.2), using the expansion methods (2.3.11) and (2.3.12) respectively. We consider a mesh with size $1/256$. The parameters are $\nu = 0.5$, $\lambda_C = 1$, $\sigma_{var}^2 = 1$, $y_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and $d = 1024$ modes.

and an $M \times 1$ vector of pairwise uncorrelated random variables $\mathbf{y}(\omega) = [y_1(\omega), \cdots, y_M(\omega)]$, i.e.

$$\mathbf{Z}(\omega, \mathbf{r}) = L_M \mathbf{y}(\omega) . \tag{2.3.14}$$

The dependence of $\mathbf{Z}$ on the spatial dimension $\mathbf{r} \in X$ is given by the M points on X (and therefore implicitly given by L_M). Moreover, we note that (2.3.14) represents the field at the M discrete points *exactly* (there is no truncation, as with the expansion methods above). However, $\tilde{C}_\nu$ is typically dense [90] and therefore the factorisation step (2.3.13) in the method (2.3.13) – (2.3.14) costs $\mathcal{O}(M^3)$ operations.

Circulant embedding aims to make the factorisation step (2.3.13) more efficient - but requires the M discrete points to be *uniformly spaced* across X. The idea is to embed the covariance matrix $\tilde{C}_\nu$ within a (larger) $N \times N$ *circulant matrix*[5], for $N \geq M$, i.e.

$$\tilde{C}_\nu^{circ} := \begin{pmatrix} \tilde{C}_\nu & A \\ A^T & B \end{pmatrix} , \tag{2.3.15}$$

where $A \in \mathbb{R}^{M \times (N-M)}$ and $B \in \mathbb{R}^{(N-M) \times (N-M)}$ are matrices chosen to ensure that $\tilde{C}_\nu^{circ}$ is symmetric positive definite. The paper [91] gives a theoretical upper bound on the size of A and B which ensures the positive definiteness of $\tilde{C}_\nu^{circ}$; assuming $\tilde{C}_\nu$ is the covariance matrix induced by the Matérn covariance function. Given that $\tilde{C}_\nu^{circ}$ is a positive definite *circulant matrix*, we can then factorise $\tilde{C}_\nu^{circ}$ in $\mathcal{O}(N \log N)$ operations via the *Fast Fourier transform*. We refer the reader to the book [135] for further details. We also note the paper [159] which extends the idea of circulant embedding to allow the M discrete points to be in a block-regular structure (i.e. they are uniformly spaced on sub-domains of X), rather than uniformly spaced over X.

We now have a number of methods which generate realisations of a random field - we will use these methods to generate realisations of some assumed uncertainty in the nuclear *input data*. Therefore, we turn our attention towards methods for estimating the statistics of the *output data*, i.e. the expected value of the quantity of interest Q defined in Section 2.2, given some prescribed uncertainty in the input data.

---

[5]a matrix is circulant if each of its row (shifted by one element to the right and wrapped back around) is the same as the row below

2.4   Computing the Expected Value

In this section, we will discuss methods for estimating the expected value of a quantity of interest $Q := \mathfrak{L}(\phi)$, defined as a functional $\mathfrak{L}$ of the scalar flux ϕ. In order to estimate $\mathbb{E}[Q]$ numerically, we need to make the following approximations:

- **Bias Error:** For a given sample $\omega \in \Omega$, we can only *approximate* the value of $Q(\mathbf{y}(\omega)) \approx Q_h(\mathbf{y}(\omega))$, where h is a parameter relating to the accuracy of the approximation. For our particular example, Q_h is the approximation acquired by applying the functional $\mathfrak{L}(\cdot)$ (or possible approximations of it) to an approximation of ϕ computed using a mesh of size h;

- **Truncation Error:** The expansion methods, discussed in Section 2.3.2, used to generate realisations of the input data are required to be truncated to $d < \infty$ terms. Hence, we are restricted from (the potentially infinite-dimensional) $\mathbf{y} = [y_1, y_2, \cdots] \in \mathbb{R}^{\mathbb{N}}$ to a $d-$dimensional approximation $\mathbf{y}^d = [y_1, \cdots, y_d] \in \mathbb{R}^d$. This also leads to a further approximation of the quantity of interest, i.e. $Q \approx Q_h \approx Q_{h,d}$;

- **Sampling Error:** We can only estimate the (integral in the) expected value $\mathbb{E}[Q_{h,d}]$ by an estimator $\widehat{Q}_{h,d}$, e.g. by Monte Carlo and multilevel Monte Carlo estimators.

We will focus on the case when d is large, and hence for simplicity, we will assume that the truncation error is negligible compared to the bias and sampling errors. Moreover, we will drop the dependence on d, writing $\mathbf{y}^d = \mathbf{y}$, $Q_{h,d} = Q_h$ and $\widehat{Q}_{h,d} = \widehat{Q}_h$. This is sometimes referred to as the *finite-noise* assumption in the literature, e.g. [198].

The first source of error in the approximation is $|\mathbb{E}[Q - Q_h]|$, which arises from estimating $\mathbb{E}[Q]$ by $\mathbb{E}[Q_h]$. The second error source is from the approximation of $\mathbb{E}[Q_h]$ by an estimator $\widehat{Q}_h$. We will quantify the overall accuracy of an estimator by its mean square error (MSE), $e(\widehat{Q}_h)^2$, defined by

$$e(\widehat{Q}_h)^2 \; := \; \mathbb{E}\left[(\mathbb{E}[Q] - \widehat{Q}_h)^2 \right] \; . \tag{2.4.1}$$

If we assume that the estimator $\widehat{Q}_h$ is unbiased, i.e. $\mathbb{E}[\widehat{Q}_h] = \mathbb{E}[Q_h]$, then the MSE (2.4.1) can be expanded as

$$
\begin{aligned}
e(\widehat{Q}_h)^2 \; &= \; \mathbb{E}\left[(\mathbb{E}[Q] - \mathbb{E}[\widehat{Q}_h] + \mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h)^2 \right] \\
&= \; \mathbb{E}\left[\left(\mathbb{E}[Q] - \mathbb{E}[\widehat{Q}_h] \right)^2 \right] + \mathbb{E}\left[\left(\mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h \right)^2 \right] + 2 \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Q] - \mathbb{E}[\widehat{Q}_h] \right) \left(\mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h \right) \right]}_{=0} \\
&= \; \left(\mathbb{E}[Q] - \mathbb{E}\left[\widehat{Q}_h \right] \right)^2 + \mathbb{V}\left[\mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h \right] + \underbrace{\left(\mathbb{E}\left[\mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h \right] \right)^2}_{=0} \\
&= \; (\mathbb{E}[Q - Q_h])^2 \; + \; \mathbb{V}[\widehat{Q}_h] \; , \tag{2.4.2}
\end{aligned}
$$

i.e. the squared bias of the approximation $Q \approx Q_h$, plus the sampling error $\mathbb{V}[\widehat{Q}_h] = \mathbb{E}[(\widehat{Q}_h - \mathbb{E}[Q_h])^2]$. The third term on the second line is zero because $\mathbb{E}[Q]$ and $\mathbb{E}[\widehat{Q}_h]$ are deterministic constants and $\mathbb{E}\left[\left(\mathbb{E}[\widehat{Q}_h] - \widehat{Q}_h \right) \right] = \mathbb{E}[\widehat{Q}_h] - \mathbb{E}[\widehat{Q}_h] = 0$. Using these two facts, the definition of the variance in (2.1.6) and that $\widehat{Q}_h$ is an unbiased estimator i.e. $\mathbb{E}[\widehat{Q}_h] = \mathbb{E}[Q_h]$, then the remaining equalities hold in a similar way.

The rate of convergence of the sampling error $\mathbb{V}[\widehat{Q}_h]$ depends on the specific choice of estimator $\widehat{Q}_h$, and we will discuss a number of possible estimators below. In order to compare the various estimators we measure their effectiveness in terms of computational ϵ-cost $\mathcal{C}_\epsilon$, that is, the number

of (floating point) operations to achieve $e(\widehat{Q}_h)^2 < \epsilon^2$, for a given accuracy $\epsilon > 0$. A sufficient condition for the MSE in (2.4.2) to be less than ϵ^2, is for both the squared bias and the sampling error to be less than $\epsilon^2/2$. This is by no means a necessary condition, and a number of authors have worked on finding the optimal ratio between the two errors, see [51, 100] for examples using Monte Carlo estimators (see ahead to (2.4.8)).

To bound the ϵ-cost for each method, we make the following assumptions on the bias error $|\mathbb{E}\left[Q - Q_h\right]|$ and on the (average) cost to compute a single sample of Q_h, denoted $\mathcal{C}(Q_h)$ (e.g. measured in floating point operations):

$$\left|\mathbb{E}\left[Q - Q_h\right]\right| = \mathcal{O}\left(h^\alpha\right) , \tag{2.4.3}$$

$$\mathbb{E}\left[\mathcal{C}(Q_h)\right] = \mathcal{O}\left(h^{-\gamma}\right) , \tag{2.4.4}$$

for some constants $\alpha, \gamma > 0$. The remainder of this chapter will now be devoted towards different choices of estimators.

One method for estimating Q (and its statistics) which has received significant attention from the UQ community within radiative transport, e.g. [16, 69, 85] and references therein, is the *polynomial chaos expansion*. Most recent research has focussed on using the non-intrusive polynomial chaos approach discussed below, and has largely been concerned with the curse of dimensionality (see ahead to (2.4.7)). Recent advancements include: (adaptive) sparse grid ideas for estimating the coefficients [85]; hybrid mixtures of polynomials [17]; and the high-dimensional model representation [16], i.e. decomposing the quantity of interest into a sum of lower-dimensional representations of the quantity of interest, each of which depends on successively larger subsets of the components of $\mathbf{y} = [y_1, \cdots, y_d]$.

The (non-intrusive) polynomial chaos method estimates the quantity of interest Q by using the following expansion [65]

$$Q(\mathbf{y}(\omega)) \approx \sum_{k=0}^{P} \varphi_k b_k(\mathbf{y}(\omega)) , \tag{2.4.5}$$

where $\{\varphi_k\}$ denote (unknown) coefficients and $\{b_k(\mathbf{y})\}$ denote (known) multi-dimensional orthogonal polynomials, which should be selected based on the type of uncertainty in the input data (we refer the reader to [210, Table 4.1]). We must then estimate the coefficients in (2.4.5), e.g. by using the (non-intrusive) spectral projection (NISP):

$$\varphi_k = \frac{\mathbb{E}\left[Q \, b_k\right]}{\mathbb{E}\left[b_k^2\right]} , \tag{2.4.6}$$

where we estimate the expectation $\mathbb{E}\left[Q \, b_k\right]$ (in the numerator) by computing $Q(\mathbf{y}^{(k)})$ at finitely many selected points $\{\mathbf{y}^{(k)}\}$ and applying quadrature - the denominator $\mathbb{E}\left[b_k^2\right]$ is usually known due to our choice for b_k. To justify (2.4.6), multiply (2.4.5) by an additional orthogonal polynomial b_j, use the orthogonality of $\{b_k\}$ and integrate over Ω.

Polynomial chaos belongs to a larger family of *interpolation* methods, which include the stochastic collocation [209, 20, 65] and stochastic Galerkin methods [78, 30]. The main advantage of these methods is that they can achieve *exponential convergence* in the sampling error [20, 210], provided Q is sufficiently smooth with respect to $\mathbf{y}$ [209]. The drawback is that the cost to achieve a certain accuracy typically grows (sometimes exponentially) with the dimension d, prohibiting their application in high dimensions - this is the *curse of dimensionality* (originally coined by Bellman in 1957 [28]). For example, for the polynomial chaos expansion (2.4.5), the number of polynomials P grows exponentially in the number of stochastic dimensions d and in the maximum order of the

polynomials M_P [78, 18, 16], i.e.

$$P = \frac{(d + M_P)!}{d! \, M_P!} - 1 \, . \tag{2.4.7}$$

The alternative to interpolation methods are the family of *Monte Carlo sampling methods.* Importantly, they *do not suffer from the curse of dimensionality.* We will discuss the standard Monte Carlo method below in Section 2.4.1, before detailing more advanced variants in later sections.

2.4.1 Monte Carlo Sampling

For each $n \in \mathbb{N}$, let $\mathbf{Z}^{(n)}$ denote the nth realisation of a random field (e.g. generated by the expansion methods in Section 2.3.2 using a vector of *independent pseudo-random* numbers $\mathbf{y}^{(n)}$). For our particular example of random nuclear input data within the RTE, e.g. (1.3.1), $\mathbf{Z}^{(n)} := \left[\sigma_S^{(n)}, \sigma_A^{(n)}, \sigma_F^{(n)}, \nu^{(n)}, f^{(n)} \right]$, where $\sigma_S^{(n)}$, $\sigma_A^{(n)}$, $\sigma_F^{(n)}$, $\nu^{(n)}$, $f^{(n)}$ denote the nth realisations of the random fields σ_S, σ_A, σ_F, ν and f respectively.

The (standard) Monte Carlo (MC) estimator for $\mathbb{E}[Q_h]$ is defined by

$$\widehat{Q}_h^{MC} := \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} Q_h(\mathbf{Z}^{(n)}) \, , \tag{2.4.8}$$

where N_{MC} is the number of Monte Carlo samples. The convergence $\widehat{Q}_h^{MC} \to \mathbb{E}[Q_h]$ holds by the law of large numbers (discussed in Appendix B). We note that it is simple to prove that $\widehat{Q}_h^{MC}$ is an unbiased estimator of $\mathbb{E}[Q_h]$, we give a one-line proof in Section B.2.

Furthermore, if we assume that for h sufficiently small

$$\mathbb{V}[Q_h] \approx c \, , \quad \text{for some constant } c > 0 \, , \text{ independent of } h \, , \tag{2.4.9}$$

then the sampling error of the MC estimator (which is given by $\mathbb{V}\left[\widehat{Q}_h^{MC}\right]$, see (2.4.2)) can be shown to be

$$\mathbb{V}\left[\widehat{Q}_h^{MC}\right] = \frac{1}{N_{MC}^2} \mathbb{V}\left[\sum_{n=1}^{N_{MC}} Q_h(\mathbf{Z}^{(n)})\right] = \frac{1}{N_{MC}^2} \sum_{n=1}^{N_{MC}} \mathbb{V}\left[Q_h(\mathbf{Z}^{(n)})\right] = \frac{\mathbb{V}[Q_h]}{N_{MC}} \, , \tag{2.4.10}$$

where the penultimate equality holds by the independence of $\{\mathbf{Z}^{(n)}\}$. This allows us to re-write the MSE (2.4.2), for the specific case of a MC estimator, as:

$$e(\widehat{Q}_h^{MC})^2 = \left(\mathbb{E}\left[Q - Q_h\right]\right)^2 + N_{MC}^{-1} \mathbb{V}[Q_h] \, . \tag{2.4.11}$$

Note that standard Monte Carlo estimators are notoriously slow to converge - $\mathcal{O}(N_{MC}^{-1/2})$ for the root-MSE.

Notation 2.4.1 *For two functions $f, g : X \mapsto \mathbb{R}$, for $X \subset \mathbb{R}$, we use the notation $f \sim g$ to mean that there exists $\tilde{x} \in \mathbb{R}$ such that $|f(x)| \leq c g(x)$, for all $x \geq \tilde{x}$, where $c > 0$ denotes an unspecified constant.*

Due to assumption (2.4.3), a sufficient condition for the squared bias to be less than $\epsilon^2/2$ is $h \sim \epsilon^{1/\alpha}$. Moreover, due to assumption (2.4.9) the sampling error of $\widehat{Q}_h^{MC}$ is less than $\epsilon^2/2$ for $N_{MC} \sim \epsilon^{-2}$. In our numerical computations we will ensure that N_{MC} is taken sufficiently large

by estimating $\mathbb{V}\left[\widehat{Q}_h^{MC}\right]$ using the sample variance, i.e.

$$\mathbb{V}\left[\widehat{Q}_h^{MC}\right] \approx \frac{1}{N_{MC}(N_{MC}-1)} \sum_{n=1}^{N_{MC}} \left(Q_h(\mathbf{Z}^{(n)}) - \widehat{Q}_h^{MC}\right)^2 , \qquad (2.4.12)$$

and increasing N_{MC} until $\mathbb{V}\left[\widehat{Q}_h^{MC}\right] \leq \epsilon^2/2$. A proof that the sample variance is an unbiased estimator of $\mathbb{V}\left[\widehat{Q}_h^{MC}\right]$ is given in Appendix B.2.

With these choices of h and N_{MC}, it follows from Assumption (2.4.4) that the mean ϵ-cost of the standard Monte Carlo estimator is

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\right] = \mathbb{E}\left[\sum_{n=1}^{N_{MC}} \mathcal{C}(Q_h(\mathbf{Z}^{(n)}))\right] = N_{MC}\,\mathbb{E}\left[\mathcal{C}(Q_h)\right] = \mathcal{O}\left(\epsilon^{-2}h^{-\gamma}\right) = \mathcal{O}\left(\epsilon^{-2-\frac{\gamma}{\alpha}}\right) .$$
$$(2.4.13)$$

The ϵ^{-2} arises from the slow convergence of the standard Monte Carlo estimator. In our application, where each sample $Q_h(\mathbf{Z}^{(n)})$ involves the numerical solution of an integro-differential equation, the MC estimator quickly becomes impractical. The alternative Monte Carlo approaches that we will now present aim to improve this situation in two complimentary ways. Firstly we will discuss methods that find well-distributed samples in high dimensions (opposed to the random samples of MC), such as quasi-Monte Carlo (QMC) sampling methods. Thereafter we discuss variance reduction techniques, and in particular multilevel variants of Monte Carlo. These methods use a hierarchy of numerical approximations to the integro-differential equation to shift the bulk of the computations to cheap, inaccurate coarse models whilst providing the required accuracy with only a handful of expensive, accurate model solves. They are also complimentary to the QMC methods we will discuss first.

2.4.2 Improved Sampling: Quasi-Monte Carlo

One approach to reduce the computational ϵ-cost in (2.4.13), by improved sampling, is based on the use of quasi-Monte Carlo (QMC) rules, which replace the random samples in (2.4.8) by *carefully chosen deterministic samples* (leading to an estimator $\widehat{Q}_h^{QMC}$). Initially interest in QMC points arose within number theory in the 1950's, and the theory is still at the heart of good QMC point construction today. The primary aim of such theory nowadays (we revisit this point below) is to find point sets that achieve $\mathcal{O}(N_{QMC}^{-\lambda})$ convergence of the (square root of the) sampling error, for some $\lambda > 1/2$, with the *hidden constant being independent of d*. For example, for a particular PDE problem with (log-normal and uniform respectively) random coefficients, [89, 124] show dimension independent $\mathcal{O}\left(N_{QMC}^{-1+\delta}\right)$ convergence, for any $\delta > 0$, by proving that the quantity of interest they consider (i.e. $Q = Q(\mathbf{y})$) belongs to some appropriate weighted Sobolev space (i.e. Q has bounded mixed derivatives, with respect to $\mathbf{y}$). Such an analysis is still an open question for transport problems and we do not attempt it here.

More recently, so-called *higher-order QMC rules* have been developed that can achieve $\mathcal{O}(N_{QMC}^{-\lambda})$ convergence for $\lambda > 1$, provided the quantity of interest belongs to an appropriate *higher-order weighted Sobolev space* (see for example [74, §2.2]). One example is the *interlaced polynomial lattice rule (IPL)* which has been studied in [57, 86, 87]. Despite the name, the IPL rule is both a lattice rule (discussed below) and a type of digital net [74].

In the 1950's and 1960's, the analysis of QMC rules within the number theory community focused on the convergence of QMC rules with respect to the number of samples N_{QMC}, whilst ignoring the dependence on d. This meant that convergence rates of the form $\mathcal{O}\left((\log N_{QMC})^{d-1}N_{QMC}^{-1}\right)$

(unless other assumptions were made) were often proven. As the demand for tackling problems with large d increased, this was deemed a fundamental flaw of QMC methods as (for large d) the logarithmic factor could dominate the improved $\mathcal{O}(N_{QMC}^{-1})$ convergence rate. It wasn't until the paper [41] that QMC rules were more widely believed to be effective for high-dimensional problems (in this case a 360-dimensional integral was estimated) [122]. Subsequently, recent analysis has been more focused towards *dimension-independent* convergence rates.

A notable example has come in the context of PDEs with random coefficients. Here the theory for (dimension-independent) quasi-Monte Carlo estimators (and multilevel variants, see ahead to Section 2.4.3) has been developed, where the coefficients are represented by a uniform random field [124, 125], or the (more difficult[6]) case where the coefficients are represented by a log-normal random field [89, 123]. A recent paper has also analysed QMC rules for an eigenvalue problem with uniform random coefficients [79]. Reviews of the recent developments of QMC theory, within high-dimensional quadrature and (elliptic) PDEs with random coefficients, are given in [58, 122] respectively.

Beyond theoretical considerations, there have been many applications of QMC rules to problems in this field, e.g. [90, 93, 172]. Other research areas where QMC rules have been successfully applied include; stochastic differential equations [84], maximum likelihood estimation [121], and (more recently) estimation of (functionals of) Bayesian posterior distributions [56, 179]. We will focus on the practical implementation of QMC rules and will not discuss the underlying theory any further - we refer the interested reader to the books [152, 131] for good introductory references.

Before we get into the details of certain QMC rules, it will be useful to observe that the problem of estimating $\mathbb{E}[Q]$ is equivalent to estimating a d-dimensional integral with integrand $Q(\mathbf{y})$ - where we recall that we assumed $\mathbf{y} = [y_1, \cdots, y_d] \in \mathbb{R}^d$. For example, when ω takes values from Ω according to the uniform probability measure i.e. $\mathrm{d}\mathbb{P}(\omega) = \mathrm{d}\omega$, we can re-write

$$\mathbb{E}[Q] = \int_\Omega Q(\mathbf{y}(\omega))\, \mathrm{d}\omega = c \int_{[0,1]^d} Q(\tilde{\mathbf{y}})\, \mathrm{d}\tilde{\mathbf{y}}\,, \tag{2.4.14}$$

where $\tilde{\mathbf{y}} = \mathcal{M}\mathbf{y}(\omega)$ is defined by a suitable (linear) transformation $\mathcal{M}: \Omega \mapsto [0,1]^d$ and c is an appropriate scaling constant. Another example is when ω takes values from $\mathbb{R}^d$ according to the product Gaussian measure, i.e. $\mathrm{d}\mathbb{P}(\omega) = \prod_{j=1}^d \varphi(\omega_j)\mathrm{d}\omega$, then

$$\mathbb{E}[Q] = \int_\Omega Q(\mathbf{y}(\omega)) \prod_{j=1}^d \varphi(\omega_j)\, \mathrm{d}\omega = \int_{(0,1)^d} Q(\tilde{\mathbf{y}})\, \mathrm{d}\tilde{\mathbf{y}}\,, \tag{2.4.15}$$

where $\tilde{\mathbf{y}} = \mathcal{M}\mathbf{y} := [\Upsilon(y_1), \cdots, \Upsilon(y_d)]$. Here Υ denotes the univariate standard normal cumulative distribution function[7] and φ denotes the standard normal probability density function, i.e.

$$\Upsilon(z) := \int_{-\infty}^z \varphi(t)\, \mathrm{d}t\,, \qquad \varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad \text{for any } z \in \mathbb{R}\,. \tag{2.4.16}$$

We note that for the infinite-dimensional case (i.e. $d \to \infty$) it is substantially more difficult to prove (2.4.15). We refer the reader to the extensive discussion in [89, eq.(1.6)-(1.7) and §3.3].

The choice of QMC point sets can be split into two categories: lattice rules and nets [152]. We will discuss (randomised) rank-1 lattice rules below.

---

[6]for example, because we no longer have (uniform) bounds from above and below on the random coefficient(s) - hence Lax-Milgram cannot be used

[7]in the literature this is usually denoted Φ, but we reserve that notation for a quantity relating to an approximation of the scalar flux ϕ

Rank-1 Lattice Rules

Lattice rules correspond to, as the name suggests, point sets where the points are distributed in a lattice structure. A lattice rule (in d-dimensions) is determined by a generating vector $\mathbf{z} \in \mathbb{Z}^d$ and the difficulty, for large d, comes in attempting to construct a generating vector that produces well-distributed points. Specifically we seek a point set, from which the estimator $\widehat{Q}_h^{QMC}$ is constructed, which minimises the worst-case error:

$$\sup_{\|Q_h\|_{\mathfrak{F}} \leq 1} |\mathbb{E}[Q_h] - \widehat{Q}_h^{QMC}| \, , \tag{2.4.17}$$

for some (given and possibly weighted) function space $\mathfrak{F}$.

The construction of generating vectors in high-dimensions first started with the Korobov construction [58] and since then several methods have been devised. A notable example is the component-by-component (CBC) algorithm [186, 187], a greedy algorithm which minimises a variant of the worst-case error (2.4.17) at each step. Subsequently, [154] presented a fast-CBC algorithm, which accelerates the computation of the worst-case error (2.4.17) by using the Fast Fourier Transform, see also [74, eq.(12),(15)] and [57]. For the reader concerned with a purely practical implementation of QMC methods, a repository of QMC rules including pre-computed generating vectors can be found at [73, 120]. In theory, the generating vector $\mathbf{z}$ should be chosen problem specific [89]. However, standard generating vectors such as those available at [73, 120], seem to also work well (and better than MC samples).

Once we have a suitable generating vector $\mathbf{z} \in \mathbb{Z}^d$, we can construct $P \geq 2$ rank-1 lattice points $\{\mathbf{y}^{(p)}\}$ by using the simple formula

$$\mathbf{y}^{(p)} \; = \; \mathrm{frac}\left(\frac{p}{P}\mathbf{z}\right) \, , \qquad p = 1, \ldots, P \, , \tag{2.4.18}$$

where 'frac$(\cdot)$' denotes the fractional part function applied component-wise to a vector. We note that (2.4.18) is a rank-1 lattice rule because it contains *one* generating vector. There have been some works into higher-rank lattice rules, e.g. [59], but we do not consider these here.

However, the rank-1 lattice points (2.4.18) were chosen deterministically and are *not independent* (hence there would be an additional covariance term in the QMC equivalent of (2.4.10), see [129, eq.(5)]) and the resulting estimator will be biased. Fortunately, by *randomising* the rank-1 lattice points - which is achieved by 'adding[8]' a uniformly distributed shift to (2.4.18) - we can make the subsequent QMC estimator (see ahead to (2.4.20)) *unbiased*. Note that it is important that the shift preserves the lattice structure, otherwise the randomised QMC points become standard Monte Carlo points and no benefit will be observed.

Consider $S \in \mathbb{N}$ independent and uniformly distributed random shifts $(\Delta_s)_{s=1}^S$ in $[0,1)^d$. Then, we can construct $N_{QMC} = S\,P$ randomised rank-1 lattice points by

$$\mathbf{y}^{(p,s)} \; = \; \mathrm{frac}\left(\frac{p}{P}\mathbf{z} + \Delta_s\right) \, , \qquad p = 1, \ldots, P, \;\; s = 1, \ldots, S \tag{2.4.19}$$

where the number of random shifts S is fixed. Observe that the random shift ensures $\mathbf{y}^{(p,s)} \sim \mathcal{U}\left([0,1)^d\right)$ and it preserves the lattice structure between all P points (for the sth shift).

The randomised lattice points in (2.4.19) are used to generate realisations $\mathbf{Z}^{(p,s)} = \mathbf{Z}(\mathbf{y}^{(p,s)})$ of a random field, in the same way as the MC samples. Then, we have the (randomised) QMC

---

[8]the definition of 'adding' changes dependent on the type of QMC rule considered

estimator

$$\widehat{Q}_h^{QMC} \;:=\; \frac{1}{S}\sum_{s=1}^{S}\widehat{Q}_h^{\Delta_s} \;=\; \frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}Q_h\left(\mathbf{Z}^{(p,s)}\right) \;=\; \frac{1}{N_{QMC}}\sum_{n=1}^{N_{QMC}}Q_h\left(\mathbf{Z}^{(n)}\right) \;, \qquad (2.4.20)$$

where we have abused notation by writing $\mathbf{Z}^{(p,s)} = \mathbf{Z}^{(n)}$, for a 2-tuple (p,s) which uniquely maps to each $n = 1, \cdots, N_{QMC}$, and where we define the one-shift estimator (taking P samples) by

$$\widehat{Q}_h^{\Delta_s} \;:=\; \frac{1}{P}\sum_{p=1}^{P}Q_h\left(\mathbf{Z}^{(p,s)}\right) \;. \qquad (2.4.21)$$

We now prove that the randomised lattice points produce an *unbiased* QMC estimator (2.4.20):

$$\begin{aligned}
\mathbb{E}_\Delta[\widehat{Q}_h^{QMC}] &= \mathbb{E}_\Delta\left[\frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}Q_h\left(\mathbf{Z}(\mathbf{y}^{(p,s)})\right)\right] \\
&= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}_\Delta\left[Q_h\left(\mathbf{Z}(\mathbf{y}^{(p,s)})\right)\right] \\
&= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}\int_{[0,1)^d}Q_h\left(\mathbf{Z}\left[\mathrm{frac}\left(\frac{p}{P}\mathbf{z}+\Delta_s\right)\right]\right)\,\mathrm{d}\Delta_s \\
&= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}\int_{[0,1)^d}Q_h\left(\mathbf{Z}(\mathbf{y}^{(p,s)})\right)\,\mathrm{d}\mathbf{y}^{(p,s)} \\
&= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}[Q_h] \\
&= \mathbb{E}[Q_h] \;,
\end{aligned}$$

where the fourth line follows from $\mathbf{y}^{(p,s)} \stackrel{d}{=} \Delta^s \sim \mathcal{U}\left([0,1)^d\right)$ (as noted below (2.4.19)), and where we use $\stackrel{d}{=}$ to denote equality in distribution and we use $\mathbb{E}_\Delta[\cdot]$ to denote the expectation with respect to the random shift $\Delta \sim \mathcal{U}\left([0,1)^d\right)$, i.e. for any function $f = f(\Delta)$, then

$$\mathbb{E}_\Delta\left[f\right] \;:=\; \int_{[0,1)^d}f\left(\Delta\right)\,\mathrm{d}\Delta \;. \qquad (2.4.22)$$

Therefore, even a single shift leads to an *unbiased* QMC estimator (2.4.20). Moreover, we note that the randomisation also gives us a practical error estimate of the sampling error - the variance of the estimator with respect to the shifts [89], i.e.

$$\mathbb{V}_\Delta\left(\widehat{Q}_h^{QMC}\right) \;:=\; \mathbb{E}_\Delta\left[\left(\widehat{Q}_h^{QMC} - \mathbb{E}[Q_h]\right)^2\right] \;. \qquad (2.4.23)$$

Let us now make the following assumption [90], which can be justified based on certain assumptions being satisfied (see [89, 125] for further details):

$$\mathbb{V}_\Delta\left[\widehat{Q}_h^{QMC}\right] \;\leq\; c\,(N^{QMC})^{-1/\lambda} \;, \qquad (2.4.24)$$

for $\lambda \in (1/2, 1]$ and a constant $c > 0$. We note that the Monte Carlo rate is recovered as $\lambda \to 1$. Therefore, we can re-write the MSE (2.4.2), for the specific case of a QMC estimator using a

randomised rank-1 lattice rule, as:

$$e(\widehat{Q}_h^{QMC})^2 \; \le \; (\mathbb{E}_\Delta\,[Q-Q_h])^2 \; + \; c(N^{QMC})^{-1/\lambda} \; . \tag{2.4.25}$$

Again, assumption (2.4.3) ensures a sufficient condition for the squared bias in (2.4.25) to be less than $\epsilon^2/2$ is $h \sim \varepsilon^{1/\alpha}$. Moreover, due to assumption (2.4.24) the sampling error of $\widehat{Q}_h^{QMC}$ is less than $\epsilon^2/2$ for $P \sim N_{QMC} \sim \epsilon^{-2\lambda}$ (for a fixed number of shifts S). In our numerical computations we will ensure that N_{QMC} is taken sufficiently large by estimating $\mathbb{V}_\Delta\left[\widehat{Q}_h^{QMC}\right]$ by the (unbiased [89]) sample variance (with respect to the number of shifts S), i.e.

$$\mathbb{V}_\Delta\left(\widehat{Q}_h^{QMC}\right) \; \approx \; \frac{1}{S(S-1)}\sum_{s=1}^{S}\left(\widehat{Q}_h^{\Delta_s}-\widehat{Q}_h^{QMC}\right)^2 \; , \tag{2.4.26}$$

and increasing N^{QMC} until the sample variance in (2.4.26) is less than $\epsilon^2/2$. We recall that $\widehat{Q}_h^{\Delta_s}$ is defined in (2.4.21). The sample variance in (2.4.26) is concerned with[9] the squared difference between the QMC estimator over all shifts, $\widehat{Q}_h^{QMC}$, and the QMC estimator for a single shift Δ_s, $\widehat{Q}_h^{\Delta_s}$, for all $s = 1, \cdots, S$.

Therefore, by using $h \sim \varepsilon^{1/\alpha}$, $P \sim N_{QMC} \sim \epsilon^{-2\lambda}$ and (2.4.4), we can show that the mean computational ϵ-cost of the QMC estimator satisfies

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}^{QMC})\right] \; = \; \mathbb{E}\left[\sum_{n=1}^{N_{QMC}}\mathcal{C}(Q_h(\mathbf{Z}^{(n)}))\right] \; = \; N_{QMC}\,\mathbb{E}\left[\mathcal{C}(Q_h)\right] \; = \; \mathcal{O}\left(\epsilon^{-2\lambda-\frac{\gamma}{\alpha}}\right) \; . \tag{2.4.27}$$

When $\lambda \to \frac{1}{2}$, this is essentially a reduction in the ϵ-cost by a whole order of ϵ, compared with the standard MC estimator (see (2.4.13)). In the case of non-smooth random fields, we typically have $\lambda \approx 1$ and the ϵ-cost grows with the same rate as that of the standard MC estimator. However, in our experiments and in experiments for diffusion problems [90], the absolute cost is always reduced.

We note that it is important to select S sufficiently large in order to guarantee that the estimate in (2.4.26) is accurate. However (for a fixed computational budget), we also want the number of points $P \sim \epsilon^{-2\lambda}$ to be large in order to reduce the sampling error in (2.4.24). In the literature, $S \in [8, 20]$ is often suggested.

Latin Hypercube Sampling

One other alternative to MC and QMC sampling is *Latin Hypercube sampling*, a type of stratified sampling first introduced in [140]. Here, the samples are chosen to satisfy a Latin Hypercube condition [157, Chapter 10] - for $d = 2$ this means that, when the space is divided into N_{LH}^2 disjoint squares, only a single sample is permitted in each row and each column of the squares. Once a hypercube (or square for $d = 2$) is selected for the sample, the sample position is then drawn from a uniform distribution over that hypercube. Formally, the nth Latin Hypercube sample is given by

$$\mathbf{y}^{(n)} \; := \; \frac{\boldsymbol{\pi}^{(n)}-\mathbf{U}^{(n)}}{N_{LH}} \; , \quad \text{for } \mathbf{U}^{(n)}\sim\mathcal{U}(0,1)^d \; , \quad \text{for all } n = 1,\cdots,N_{LH} \; , \tag{2.4.28}$$

---

[9]in comparison, (2.4.12) is concerned with the squared difference between the MC estimator over all samples, $\widehat{Q}_h^{MC}$, and the single sample $Q_h(\mathbf{Z}^{(n)})$, for all $n = 1, \cdots, N_{MC}$
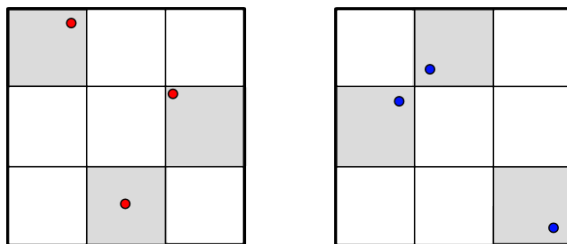
Figure 2-5: Two examples of Latin Squares in 2D, with $N_{LH} = 3$ samples. The selected Latin squares are shaded grey and the samples are uniformly distributed over these squares.
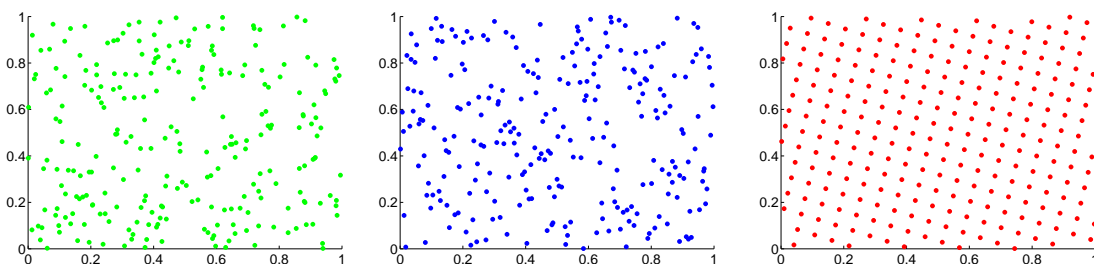
Figure 2-6: Comparison of 256 (Left) Monte Carlo, (Centre) Latin Hypercube and (Right) quasi-Monte Carlo samples on $[0, 1]^2$. MC points drawn uniformly on $[0, 1]^2$, Latin Hypercube samples drawn using the `lhsdesign` function in `Matlab` and the QMC points are drawn according to an (non-randomised) extensible rank-1 lattice rule found at [120].

where each $\boldsymbol{\pi}^{(n)}$ denotes a vector of elements chosen randomly from the set $\{1, \cdots, N_{LH}\}$ such that, for all $j = 1, \cdots, d,$

$$\left(\boldsymbol{\pi}^{(1)}\right)_j \, , \, \left(\boldsymbol{\pi}^{(2)}\right)_j \, , \, \cdots \, , \, \left(\boldsymbol{\pi}^{(N_{LH})}\right)_j \, ,$$

is a random permutation of the set $\{1, \cdots, N_{LH}\}$. We present two simple illustrations of $N_{LH} = 3$ Latin Hypercube samples (for $d = 2$) in Figure 2-5.

The Latin Hypercube estimator then takes the same form as (2.4.8), but with Latin Hypercube samples $\mathbf{y}^{(n)}$. It can be proven that [140, Thm. 1 and pg.244], [157, Theorem 10.1], under certain conditions on Q (which are always satisfied as $N_{LH} \to \infty$ [193]), the Latin Hypercube estimator is unbiased and has lower variance than the Monte Carlo estimator (2.4.8). However, the convergence rate (for the sampling error) is the same as the MC estimator [193, Corollary 1], but with a better constant [119, 193]. We refer the reader to [196, 208] for more advanced Latin Hypercube sampling procedures.

We illustrate the difference, for $d = 2$, between the Monte Carlo, Latin Hypercube and (rank-1 lattice) quasi-Monte Carlo samples in Figure 2-6. We also note that a numerical comparison of the accuracy of Monte Carlo, Latin Hypercube and a (net based) QMC method is given in [119], whom concluded that QMC outperformed both methods, in general.

2.4.3 Variance Reduction: Multilevel Monte Carlo

The 'improved sampling' techniques previously discussed, such as the QMC method, aim to reduce the sampling error in the MSE by finding samples that are better distributed than MC samples. We will now consider a second category of methods, the *variance reduction* methods, which seek to find estimators with lower variance when using the same initial sampling. Many of these techniques

have been developed and we mention just a few here: importance sampling, control variates [81], Monte Carlo with least-squares [151] and Richardson extrapolation [80, 130]. The focus for the remainder of this chapter will be on multilevel and multi-index Monte Carlo, two examples using control variates.

The multilevel Monte Carlo (MLMC) method uses a hierarchy of discrete models of increasing cost and accuracy, corresponding to a sequence of decreasing discretisation parameters $h_0 > h_1 > \cdots > h_L = h$. Here, only the most accurate model on level L is designed to give a bias error of $\mathcal{O}(\epsilon)$ by choosing $h_L = h \sim \epsilon^{1/\alpha}$ as above (for MC). The bias error of the other models can be significantly higher.

MLMC methods were first proposed in an abstract way for high-dimensional quadrature by Heinrich [103] and then popularised in the context of stochastic differential equations in mathematical finance by Giles [80]. They were first applied in uncertainty quantification in [26, 48] for elliptic PDEs and quickly gained popularity. They have since been further developed and applied in a variety of other problems, including parabolic problems [83], hyperbolic problems [93, 143, 144], variational inequalities [117] and Kalman filters [105].

MLMC methods exploit the linearity of the expectation, writing

$$\mathbb{E}[Q_h] \;=\; \mathbb{E}[Q_L] \;=\; \sum_{\ell=0}^{L} \mathbb{E}[Y_\ell]\,, \qquad \text{where} \quad Y_\ell \;:=\; Q_\ell - Q_{\ell-1} \quad \text{and} \quad Q_{-1} := 0\,,$$

where we have abused notation by writing $Q_\ell = Q_{h_\ell}$. Each of the expected values $\mathbb{E}[Y_\ell]$ on the right hand side is then estimated separately. In particular, in the case of a standard MC estimator with N_ℓ samples used to estimate each $\mathbb{E}[Y_\ell]$, we obtain the MLMC estimator

$$\widehat{Q}_h^{MLMC} \;:=\; \sum_{\ell=0}^{L} \widehat{Y}_\ell^{MC} \;=\; \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} Y_\ell(\mathbf{Z}^{(\ell,n)})\,. \tag{2.4.29}$$

Here, $\{\mathbf{Z}^{(\ell,n)}\}_{n=1}^{N_\ell}$ denotes the samples of the random field corresponding to i.i.d. MC samples on level ℓ and chosen independently from the samples on the other levels. An illustration of the overall algorithm is given in Figure 2-7.

Since each $\mathbb{E}[Y_\ell]$ is estimated independently (here using MC estimators) then we can write

$$\mathbb{V}[\widehat{Q}_h^{MLMC}] \;=\; \mathbb{V}\left[\sum_{\ell=0}^{L} \widehat{Y}_\ell^{MC}\right] \;=\; \sum_{\ell=0}^{L} \mathbb{V}\left[\widehat{Y}_\ell^{MC}\right]\,, \tag{2.4.30}$$

and hence by using (2.4.10) we have that $\mathbb{V}\left[\widehat{Y}_\ell^{MC}\right] = N_\ell^{-1}\mathbb{V}[Y_\ell]$. Subsequently the MSE (2.4.2) for the MLMC estimator (2.4.29) can be re-written as:

$$e(\widehat{Q}_h^{MLMC})^2 \;=\; (\mathbb{E}\left[Q - Q_h\right])^2 \;+\; \sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Y_\ell]\,. \tag{2.4.31}$$

We want to find the sequence $\{N_\ell\}$ that achieves $\mathbb{V}[\widehat{Q}_h^{MLMC}] = \sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Y_\ell] \leq \epsilon^2/2$, for as little computational cost as possible. This result was first given in [80], and we detail the proof in Appendix C by using the method of Lagrange multipliers.

Lemma 2.4.2 *Consider the MLMC estimator $\widehat{Q}_h^{MLMC}$, defined in (2.4.29), where $\mathbb{E}[Y_\ell]$ is estimated with N_ℓ Monte Carlo samples, for all $\ell = 0, \cdots, L$. Then the (optimal) choice of $\{N_\ell\}$, in the sense that it minimises the cost of the MLMC estimator whilst achieving $\mathbb{V}[\widehat{Q}_h^{MLMC}] \leq \epsilon^2/2$,*
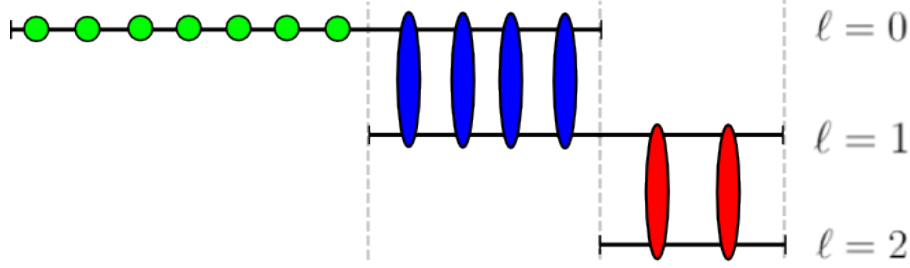
Figure 2-7: Illustration of the MLMC algorithm for three levels. (Green) circles denote samples estimating $\mathbb{E}[Q_0]$; (Blue/Red) ellipses denote samples estimating the difference $\mathbb{E}[Y_\ell]$. Note that estimating the expectation of the difference $Y_\ell = Q_\ell - Q_{\ell-1}$ requires using the same sample on two different levels.

is given by

$$N_\ell \;=\; \left\lceil 2\epsilon^{-2} \left(\sum_{\ell=0}^{L} \sqrt{\mathbb{V}[Y_\ell]\mathcal{C}_\ell} \right) \sqrt{\frac{\mathbb{V}[Y_\ell]}{\mathcal{C}_\ell}} \right\rceil \;, \tag{2.4.32}$$

where $\mathcal{C}_\ell := \mathbb{E}\left[\mathcal{C}(Y_\ell)\right]$ denotes the average cost of a single sample of $Y_\ell = Q_\ell - Q_{\ell-1}$, and $\lceil \cdot \rceil$ denotes the ceiling function.

In practice it is necessary to estimate $\mathbb{V}[Y_\ell]$ and $\mathcal{C}_\ell$ in (2.4.32) from the computed samples, updating N_ℓ as the simulation progresses.

The key idea in MLMC is to avoid estimating $\mathbb{E}[Q_h] = \mathbb{E}[Q_L]$ directly. Instead, the expectation $\mathbb{E}[Y_0] = \mathbb{E}[Q_0]$ of a possibly strongly biased, but cheap approximation of Q_h is estimated. The bias of this coarse model is then estimated by a sum of correction terms $\mathbb{E}[Y_\ell]$ using increasingly accurate and expensive models. Since the Y_ℓ represent small corrections between the coarse and fine models, it is reasonable to conjecture that there exists $\beta > 0$ such that

$$\mathbb{V}[Y_\ell] \;=\; \mathcal{O}(h_\ell^\beta) \;, \tag{2.4.33}$$

i.e. the variance of Y_ℓ decreases as $h_\ell \to 0$. Such a condition holds if Q_ℓ converges to Q pathwise. This is verified for a diffusion problem in [47], and for a certain transport problem in Chapter 4 and Chapter 5. Therefore the number of samples N_ℓ to achieve a prescribed accuracy on level ℓ can be gradually reduced (according to (2.4.32)), leading to a lower overall cost of the MLMC estimator. More specifically, we have the following cost savings.

Remark 2.4.3 *(i) On the coarsest level, using (2.4.4), the cost per sample is reduced from $\mathcal{O}(h^{-\gamma})$ to $\mathcal{O}(h_0^{-\gamma})$. Provided $\mathbb{V}[Q_0] \approx \mathbb{V}[Q_L]$ and h_0 can be chosen independently of ϵ (i.e. (2.4.9)), the cost of estimating $\mathbb{E}[Q_0]$ to an accuracy of ϵ in (2.4.29) is reduced to $\mathcal{O}(\epsilon^{-2})$;*

(ii) On the finer levels, the number of samples N_ℓ to estimate $\mathbb{E}[Y_\ell]$ to an accuracy of ϵ in (2.4.29) is proportional to $\mathbb{V}[Y_\ell]\epsilon^{-2}$. Now, provided $\mathbb{V}[Y_\ell] = \mathcal{O}(h_\ell^\beta)$, for some $\beta > 0$ (i.e. (2.4.33)), which is guaranteed if Q_ℓ converges almost surely to Q pathwise, then we can reduce the number of samples as $h_\ell \to 0$. Depending on the actual values of α, β and γ, the cost to estimate $\mathbb{E}[Y_L]$ on the finest level can, in the best case, be reduced to $\mathcal{O}(\epsilon^{-\gamma/\alpha})$.

We emphasise that each sample of $Y_\ell(\mathbf{Z}) := Q_\ell(\mathbf{Z}) - Q_{\ell-1}(\mathbf{Z})$ is found by computing Q_ℓ and $Q_{\ell-1}$ for *the same realisation* $\mathbf{Z}$. To achieve variance reduction (i.e. (2.4.33)) it is very important that $Q_\ell(\mathbf{Z})$ is *(strongly) positively correlated* with $Q_{\ell-1}(\mathbf{Z})$, for all $1 \leq \ell \leq L$. To explore this point

further, recall that we assumed (in (2.4.9)) that $\mathbb{V}[Q_\ell] \approx c$, a positive constant independent of h_ℓ, for all $\ell = 0, \cdots, L$. Then, we can write

$$\mathbb{V}[Y_\ell] \;=\; \mathbb{V}[Q_\ell] + \mathbb{V}[Q_{\ell-1}] - 2\mathrm{Cov}\,(Q_\ell, Q_{\ell-1}) \;\approx\; 2\,(\mathbb{V}[Q_\ell] - \mathrm{Cov}\,(Q_\ell, Q_{\ell-1}))\;,$$

where $\mathrm{Cov}\,(\cdot, \cdot)$ is defined in (2.3.3). To be able to estimate $\mathbb{E}[Y_\ell]$ in fewer samples than $\mathbb{E}[Q_\ell]$, but to achieve the same overall prescribed accuracy, we require that $\mathbb{V}[Y_\ell] < \mathbb{V}[Q_\ell]$. A necessary condition for this to hold is $\mathrm{Cov}\,(Q_\ell, Q_{\ell-1}) \geq (1/2)\mathbb{V}[Q_\ell]$, but assuming that Q_ℓ converges to Q pathwise then $\lim_{\ell \to \infty} \mathrm{Cov}\,(Q_\ell, Q_{\ell-1}) = \mathbb{V}[Q_\ell]$. This variance reduction (i.e. $\lim_\ell \mathbb{V}[Y_\ell] = 0$) is fundamental to the gains of MLMC.

Assuming (2.4.33) holds, [48] proved the following theorem on the average computational ϵ-cost of the MLMC estimator ([80] first proved a similar result). We present a sketch of the proof in Appendix C.

Theorem 2.4.4 *Assume that (2.4.3), (2.4.33) and (2.4.4) hold with $\alpha, \beta, \gamma > 0$ and $\alpha \geq \frac{1}{2}\min\{\beta, \gamma\}$. Then, for any $\epsilon < \exp(-1)$, there exists an $L \sim \log(\epsilon^{-1})$ and a sequence $\{N_\ell\}_{l=0}^L$ such that $e(\widehat{Q}_h^{MLMC})^2 \leq \epsilon^2$ and*

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\right] \;=\; \mathcal{O}\left(\epsilon^{-2-\max\{0,\,(\gamma-\beta)/\alpha\}}\right), \quad \text{for } \beta \neq \gamma\,. \tag{2.4.34}$$

For $\beta = \gamma$, we can achieve $\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\right] = \mathcal{O}\left(\epsilon^{-2}\log(\epsilon)^2\right)$.

Theorem 2.4.4 states that, provided (2.4.33) holds for some $\beta > 0$, the MLMC always achieves a gain of $\mathcal{O}\left(\epsilon^{-\min\{\beta,\gamma\}/\alpha}\right)$ over standard Monte Carlo. For $\beta > \gamma$, the cost of the MLMC method is $\mathcal{O}\left(\epsilon^{-2}\right)$ (i.e. the dominant computational cost is on the coarsest levels, see Remark 2.4.3(i)). This fact can be exploited to design unbiased multilevel estimators of the expected value of the exact quantity of interest $\mathbb{E}[Q]$ with cost $\mathcal{O}\left(\epsilon^{-2}\right)$ [169]. On the other hand, if $\gamma > \beta = 2\alpha$, the cost of the MLMC method is

$$\mathcal{O}\left(\epsilon^{-2-\frac{\gamma-\beta}{\alpha}}\right) \;=\; \mathcal{O}\left(\epsilon^{-2-\frac{\gamma}{\alpha}+\frac{2\alpha}{\alpha}}\right) \;=\; \mathcal{O}\left(\epsilon^{-\gamma/\alpha}\right)\;,$$

(i.e. the dominant computational cost is on the finest levels, see Remark 2.4.3(ii)). This is optimal, in the sense that it is equivalent (up to a constant) to the cost of computing a *single (standard) Monte Carlo sample* to $\mathcal{O}(\epsilon)$ accuracy - since (2.4.3) requires $h \sim \epsilon^{1/\alpha}$ to ensure $\mathcal{O}(\epsilon)$ accuracy and the mean cost of a single sample is given by (2.4.4), hence

$$\mathbb{E}[\mathcal{C}(Q_h)] \;=\; \mathcal{O}\left(h^{-\gamma}\right) \;=\; \mathcal{O}\left(\epsilon^{-\gamma/\alpha}\right)\;.$$

For certain problems (e.g. hyperbolic PDEs), a stability condition on the discretisation parameter h must be satisfied otherwise instabilities in the numerical scheme can lead to a large bias error in the approximation $Q \approx Q_h$. The appearance of such stability conditions is a particular issue for MLMC, which relies upon the estimates Q_h being (strongly) positively correlated for different values of h. The author is not aware of any theoretical results for MLMC when the discretised problem requires a stability constraint, despite the increased interest in such problems - see for example [25, 143, 144]. The detailed analysis of the affect of the stability condition on the MLMC theory, for a spatially one-dimensional radiative transport problem, is given in our paper [92]. We will discuss this further in Chapter 4.

One other area of active research within MLMC, is its application to the computation of cdfs and pdfs, e.g. [31, 67, 82, 118]. Here there is the additional difficulty of justifying (2.4.33) when

the quantity of interest is discontinuous on its domain. For example, consider the cumulative distribution function of Q and write

$$\mathbb{P}[Q \leq b] \;=\; \mathbb{E}[\mathbb{1}_{(-\infty,b]}(Q)] \;=\; \mathbb{E}\left[\xi(Q,b)\right] \;, \quad \xi(Q,b) := \mathbb{1}_{(-\infty,b]}(Q) \;, \quad \text{for all } b \in \mathbb{R} \;, \quad (2.4.35)$$

where $\mathbb{1}_S(\cdot)$ denotes the indicator function for some set S. Then, we can use the MLMC estimator (2.4.29) to estimate $\mathbb{E}[\xi] = \mathbb{P}[Q \leq b]$, by considering a sequence of approximations $\{\xi_\ell\}_{\ell=0}^L$ as before, where we define $\xi_\ell(\omega,b) := \xi\left(Q_\ell(\omega),b\right)$, for all $\ell = 0,\cdots,L$. However in general, there exists a $b \in \mathbb{R}$ such that, for all $\omega \in \Omega$ (and assuming $Q_\ell(\omega) \neq Q_{\ell-1}(\omega)$) then

$$\begin{aligned} \text{either:} \quad & \xi_\ell(\omega,b) \;=\; 1 \;, \quad \text{and} \quad \xi_{\ell-1}(\omega,b) \;=\; 0 \;; \\ \text{or} \quad & \xi_\ell(\omega,b) \;=\; 0 \;, \quad \text{and} \quad \xi_{\ell-1}(\omega,b) \;=\; 1 \;. \end{aligned} \qquad (2.4.36)$$

That is, the estimates of the quantity of interest fall on opposite sides of the discontinuity, depending on the chosen level. This counteracts the correlated samples that MLMC relies upon, see [82, Remark 5.1].

One option to remedy this is to consider a *smooth approximation* to the discontinuous function [82]. The problem with this smoothing approach is that the smoothing parameter should depend on the required tolerance of the problem, which leads to difficult tuning of MLMC. For the specific case of estimating densities and distributions, other options have been suggested which move away from formulas such as (2.4.35). For example, [31] uses the moment matching method where (finitely many) statistical moments of the quantity of interest are estimated via MLMC, and then an estimate of the pdf is constructed using the moment estimates.

Finally we note that the multilevel approach is not restricted to standard MC estimators and can also be used in conjunction with QMC estimators [84, 125, 123] or with stochastic collocation [198]. A comprehensive review is given in [81]. We discuss the multilevel variant of QMC next.

Multilevel Quasi-Monte Carlo

An important observation, first made in [84], is that the gains of quasi-Monte Carlo are *complimentary* to the gains of MLMC. Consider samples $\mathbf{Z}^{(\ell,n)}$ of a random field corresponding to the (randomised) rank-1 lattice points (2.4.19), for each level $\ell = 0,\cdots,L$ (instead of the i.i.d. MC samples in (2.4.29)) - with the same abuse of notation as in (2.4.20). Then, we can define the multilevel quasi-Monte Carlo (MLQMC) estimator by

$$\widehat{Q}_h^{MLQMC} \;:=\; \sum_{\ell=0}^L \widehat{Y}_\ell^{QMC} \;=\; \sum_{\ell=0}^L \frac{1}{N_\ell^{QMC}} \sum_{n=1}^{N_\ell^{QMC}} Y_\ell(\mathbf{Z}^{(\ell,n)}) \;, \qquad (2.4.37)$$

where $N_\ell^{QMC} = S\,P_\ell$ denotes the total number of QMC samples on level ℓ, with S uniform shifts of P_ℓ rank-1 lattice points, for each $\ell = 0,\cdots,L$.

A theoretical result on the computational ϵ-cost of $\widehat{Q}_h^{MLQMC}$ can be proven, which is analogous to Theorem 2.4.4, see [197, 125, 123]. We present the result below.

Theorem 2.4.5 *Assume that (2.4.3) and (2.4.4) hold with $\alpha,\gamma > 0$, and that there exists $\lambda \in (\frac{1}{2},1]$ and $\beta > 0$ such that $\alpha \geq \frac{1}{2}\min\{\beta,\lambda^{-1}\gamma\}$ and*

$$\mathbb{V}_\Delta\left[\widehat{Y}_\ell^{QMC}\right] \;=\; \mathcal{O}\left(\left(N_\ell^{QMC}\right)^{-1/\lambda} h_\ell^\beta\right) \;. \qquad (2.4.38)$$

Moreover, fix the number of random shifts on each level as S. Then, for any $\epsilon < \exp(-1)$, there exists an $L \sim \log(\epsilon^{-1})$ and a sequence $\{N_\ell^{QMC}\}_{\ell=0}^L$ such that $e\left(\widehat{Q}_h^{MLQMC}\right) \leq \epsilon^2$ and

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLQMC})\right] = \mathcal{O}\left(\epsilon^{-2\lambda - \max\{0, \frac{\gamma - \beta\lambda}{\alpha}\}}\right) . \tag{2.4.39}$$

For $\beta\lambda = \gamma$, we can achieve $\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLQMC})\right] = \mathcal{O}\left(\epsilon^{-2\lambda}\log(\epsilon^{-1})^{1+\lambda}\right)$.

The convergence rate $\lambda \in (\frac{1}{2}, 1]$ can be further improved by using the higher-order QMC rules that we previously mentioned, e.g. [57, 86, 87], but we do not consider this here.

The complexity result in Theorem 2.4.5 (and Theorem 2.4.4) can also be extended to include the truncation error that arises from the use of the (d_ℓ-truncated) expansion methods in Section 2.3.2. We refer to [123, 125] for further details.

The question still remains as to what the optimal values for N_ℓ^{QMC} are in (2.4.37). In an analogous way to (2.4.32), we can show that they are given by

$$N_\ell^{QMC} = \left\lceil 2^\lambda \epsilon^{-2\lambda} \left(\sum_{\ell=0}^L \mathcal{C}_\ell^{\frac{1}{\lambda+1}} \mathbb{V}_\Delta[Y_\ell]^{\frac{\lambda}{\lambda+1}}\right)^\lambda \left(\frac{\mathbb{V}_\Delta[Y_\ell]}{\mathcal{C}_\ell}\right)^{\frac{\lambda}{\lambda+1}} \right\rceil . \tag{2.4.40}$$

Note that (2.4.40) depends strongly on the value of $\lambda \in (\frac{1}{2}, 1]$, i.e. the rate of (sampling error) convergence of the (randomised) rank-1 lattice rule from (2.4.38).

In practice it is difficult to accurately estimate λ, because a pre-asymptotic rate $\lambda_{\text{eff}} > \lambda$ is typically observed [123]. We will give a practically more useful approach for computing (on-the-fly) quasi-optimal N_ℓ^{QMC} below, see (2.4.42) and the surrounding algorithm. First let us give details on the justification behind it (i.e. (2.4.41)) - we refer to [123] for further details.

Let us assume that $\mathbb{V}_\Delta\left[\widehat{Y}_\ell^{QMC}\right] \approx v_\ell \left(N_\ell^{QMC}\right)^{-1/\lambda}$ (ignoring higher-order terms), for some $\lambda > 0$ and for some level-dependent constant $0 < v_\ell \leq ch_\ell^\beta$, for all $\ell = 0, \cdots, L$. Moreover, assume $\mathcal{C}_\ell \approx w_\ell N_\ell^{QMC}$ (ignoring lower order terms) for a level-dependent constant w_ℓ which is independent of N_ℓ^{QMC}, for all $\ell = 0, \cdots, L$. Then, if we set up the same constrained optimisation problem that is used in the computation of (2.4.40) (and is similar to that outlined in the proof of (2.4.32)), it follows that

$$\frac{\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]}{\mathcal{C}_\ell} \approx \frac{\lambda}{\mu} , \quad \text{for all } \ell = 0, \cdots, L , \tag{2.4.41}$$

where μ is the Lagrange multiplier (a level-independent constant) used to compute the solution of the optimisation problem (further details are given in [123, § 3.3]). This means that the theoretically optimal values of N_ℓ^{QMC} from (2.4.40) ensure that $\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]\mathcal{C}_\ell^{-1}$ is approximately constant on all levels. Hence, building an algorithm that (on-the-fly) ensures $\mathbb{V}_\Delta[\widehat{Y}_0^{QMC}]\mathcal{C}_0^{-1} \approx \mathbb{V}_\Delta[\widehat{Y}_1^{QMC}]\mathcal{C}_1^{-1} \approx \cdots \approx \mathbb{V}_\Delta[\widehat{Y}_L^{QMC}]\mathcal{C}_L^{-1}$ will give us quasi-optimal values for N_ℓ^{QMC} - without needing to compute them a-priori and without relying on an accurate estimate of λ. This leads us to the following adaptive procedure to choose N_ℓ^{QMC}, as suggested in [84]. We also note that the algorithm can easily be adapted to estimate quasi-optimal N_ℓ for MLMC. We use this adaptive procedure for all numerical experiments, instead of (2.4.32) and (2.4.40).

Starting with an initial number of samples on all levels, we alternate the following two steps until $\mathbb{V}[\widehat{Q}_h^{MLQMC}] \leq \epsilon^2/2$:

(i) Estimate $\mathcal{C}_\ell$ and $\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]$ by using all of the current N_ℓ^{QMC} samples, for all $\ell = 0, \cdots, L$.

(ii) Compute

$$\ell^* = \arg\max_{\ell=0}^{L} \left(\frac{\mathbb{V}_\Delta[\widehat{Y}_\ell^{QMC}]}{\mathcal{C}_\ell} \right) , \tag{2.4.42}$$

and double the number of samples on level ℓ^*.

We will discuss an extension of the multilevel Monte Carlo framework, known as Multi-Index Monte Carlo, in Appendix D.

In the next two chapters we propose and analyse efficient multilevel Monte Carlo methods (discussed in Section 2.4.3) for quantifying the effect of uncertainty in the *nuclear input data* on the *output variable* ϕ (or a related quantity of interest). There is a growing recent interest in this question in the more general context of kinetic equations (e.g. [143, 144, 213]). In the particular case of nuclear applications, our work is relevant to the assessment of how material fluctuations can affect the uncertainty of flux computations.

The analysis of multilevel Monte Carlo methods for PDE problems has so far been restricted to the case of ODE and coercive elliptic PDE models, where it has generated a lot of interest (e.g. [80], [48]). As far as we are aware, the results in the next two Chapters are the first which consider this question for hyperbolic integro-differential equations of the form (1.1.14).

52

Chapter 3

Deterministic Error Analysis for Heterogeneous Transport

Contents

The results of this chapter describe how (deterministic) heterogeneity in the material coefficients manifests itself in the operators underlying the RTE, and consequently in the error estimate for the numerical method. That is, this chapter is concerned with an error estimate for the deterministic problem - this result is key to the analysis of the multilevel Monte Carlo method in Chapter 4.

To allow the first results to be established, we make the simplifying assumption of one spatial and one angular dimension, the so-called "1D slab-geometry" case in reactor theory (which we previously introduced in Section 1.3). We discretise with the classical discrete ordinates method, using a certain Gauss like rule with $2N$ quadrature points in angle and classical diamond differencing (or Crank-Nicolson) on a mesh with step-size h in the spatial variable. The resulting approximation of the scalar flux ϕ is denoted $\phi^{h,N}$.

We assume that the spatial domain can be partitioned into subintervals, on each of which the input data σ_S, σ_A and f belongs to the Hölder[1] space C^η, for some $\eta \in (0,1)$. This allows for data with low smoothness and permits jumps in material properties across interfaces. We denote this space by C^η_{pw} and equip it with the norm $\|\cdot\|_{\eta,pw}$. The main result of this chapter, and our first error estimate, is Theorem 3.3.11 which shows that there are constants $\mathcal{R}, \mathcal{R}'$, both dependent on

---

[1] we refer the reader ahead to Notation 3.1.1 for details on the notation used in this introduction

σ, σ_S such that, when

$$N^{-1} \; + \; h \log N \; + \; h^{\eta} \; \leq \; \mathcal{R}(\sigma, \sigma_S)^{-1} \; , \tag{3.0.1}$$

we have the error estimate for the scalar flux:

$$\|\phi - \phi^{h,N}\|_{\infty} \; \leq \; \mathcal{R}'(\sigma, \sigma_S) \left(N^{-1} \; + \; h \log N \; + \; h^{\eta} \right) \|f\|_{\eta, pw} \; , \tag{3.0.2}$$

Both $\mathcal{R}, \mathcal{R}'$ are independent of f and their dependence on σ, σ_S is known explicitly, but they blow up, e.g., if $\sigma_S/\sigma \to 1$ anywhere in the domain or if $\|\sigma\|_{\infty}\|\sigma^{-1}\|_{\infty} \to \infty$. This is natural as, for example, it is known that when σ_S is close to σ the transport equation degenerates to a diffusion equation and hence we expect subsequent difficulties. We presented an overview of the required steps to achieve (3.0.2) in our paper [93, §3], without many details and with certain assumptions removed. The proof of (3.0.1), (3.0.2) is obtained by generalising the theory of the integral equation reformulation of (1.1.5) to the heterogeneous case; the homogeneous case having been studied in detail in [163].

The appearance of the $h \log N$ term in (3.0.2) reflects the fact that the transport equation in slab geometry has a singularity in its angular dependence (explained in Section 3.1). This imposes a compatiblity constraint (i.e. (3.0.1)), which implies that the angular discretisation cannot be indefinitely refined if the spatial discretisation is kept fixed. The appearance of this term in the error estimate means that the accuracy of the method, measured in $\| \cdot \|_{\infty}$, can be no better than $\mathcal{O}(h)$, even if the cross-sections are very smooth. A faster rate is possible if one uses a higher order method or measures the error in L_p norms, the latter proved for constant cross-sections in [163]. However, we will not pursue this further and thus limit our analysis to less smooth data ($\eta < 1$).

The numerical analysis of the RTE (and related integro-differential equation problems) dates back at least as far as the work of H.B. Keller [114]. After a huge growth in the mathematics literature in the 1970's and 1980's, progress has been slower since. This is perhaps surprising, since discontinuous Galerkin (DG) methods (i.e. Section 1.2.3) have enjoyed a massive recent renaissance and the neutron transport problem was one of the motivations behind the original introduction of DG [166].

The fundamental paper on the analysis of the discrete ordinates method for the transport equation is [163], where a full analysis of the combined effect of angular and spatial discretisation is given, under the assumption that the cross-sections are constant. The delicate relation between spatial and angular discretisation parameters required to achieve stability and convergence is described there, and is also seen again in the present work (see the $h \log N$ term in (3.0.2)). Later research e.g. [11], [12] produced analogous results for models of increasing complexity and in higher dimensions, but the proofs were mostly confined to the case of cross-sections that are constant in space. A separate and related sequence of papers (e.g. [128], [204], and [7]) allow for variation in cross-sections, but error estimates explicit in this data are not available there.

We also note that there is a growing literature in the numerical analysis of kinetic equations, of which the RTE is a particular example, with an emphasis on "asymptotic preserving" schemes, which retain accuracy as the scattering ratio σ_S/σ approaches unity. Interest in this question in the deterministic case goes back a long way, e.g. [109], which has led to recent work on UQ in this context (e.g. [213]). For further general discussion on the transport equation, see [52, 133].

The structure of this chapter is as follows. In Section 3.1, we introduce the model problem; the Radiative Transport equation in the 1D slab geometry with spatially heterogeneous cross-sections, and its discretisation. To set up the error analysis, Section 3.2 describes the classical integral

equation reformulation of the RTE under very weak smoothness assumptions on the cross-sections. From here we can prove results relating to the underlying operators and their regularity - that are explicit in the cross-sections. In Section 3.3, the elements are brought together to prove (3.0.1) and (3.0.2). Numerical results are presented alongside, which show that our *error estimate is sharp*. Chapter 4 then extends this analysis to include a probabilistic error estimate (see ahead to (4.0.1)) and a rigorous analysis of the (multilevel) Monte Carlo method.

3.1 The Model Problem

We study the *mono-energetic 1D slab geometry problem*, for the angular flux $\psi(x,\mu)$:

$$\mu\frac{\partial\psi}{\partial x}(x,\mu) \; + \; \sigma(x)\psi(x,\mu) \; = \; \sigma_S(x)\phi(x) \; + \; f(x) \;, \quad x \in (0,1), \quad \mu \in [-1,1], \tag{3.1.1}$$

$$\text{where} \qquad \phi(x) \; = \; \frac{1}{2}\int_{-1}^{1}\psi(x,\mu')\;d\mu' \;, \tag{3.1.2}$$

denotes the scalar flux, subject to zero incoming flux:

$$\psi(0,\mu) \; = \; 0, \;\; \text{for} \;\; \mu > 0 \;\; \text{and} \;\; \psi(1,\mu) = 0, \;\; \text{for} \;\; \mu < 0 \;. \tag{3.1.3}$$

The total cross-section $\sigma(x)$ is given by $\sigma = \sigma_S + \sigma_A$. The problem (3.1.1) – (3.1.3) was shown to be equivalent to (1.3.1) equipped with no-inflow boundary conditions, when the input data is constant in two of the spatial dimensions (here assumed to be y and z), see Section 1.3. Note that (3.1.1) degenerates at $\mu = 0$, which corresponds to particles moving perpendicular to the x-direction.

Notation 3.1.1 *When working on the spatial domain $(0,1)$, for $1 \leq p \leq \infty$, we will denote the standard Lebesgue spaces as L_p with norm $\|\cdot\|_p$, i.e. for $1 \leq p < \infty$*

$$L_p(0,1) \; := \; \left\{ g : (0,1) \mapsto \mathbb{R} \; | \; \|g\|_{L_p(0,1)}^p \; := \; \int_{(0,1)} |g(x)|^p \; dx \; < \; \infty \right\} \;.$$

When $p = \infty$, $\|g\|_\infty := ess\, sup_{x\in[0,1]}|g(x)|$. For any interval $I \subset [0,1]$, we denote by $\mathrm{C}(I)$ the space of uniformly continuous functions on I, equipped with norm $\|\cdot\|_\infty$. For $0 < \xi \leq 1$, we let $\mathrm{C}^\xi(I)$ denote the space of Hölder continuous functions on I with Hölder exponent $\xi \in (0,1]$ and with norm

$$\|g\|_{\mathrm{C}^\xi(I)} \; := \; \|g\|_\infty + \sup_{\substack{x,y\in I \\ x\neq y}} \frac{|g(x) - g(y)|}{|x-y|^\xi} \;.$$

(The space $\mathrm{C}^1(I)$, for $\xi = 1$, is also the space of Lipschitz continuous functions on I.) When $I = [0,1]$ we write for short $\mathrm{C} = \mathrm{C}(I)$, $\mathrm{C}^\xi = \mathrm{C}^\xi(I)$ and $\|g\|_\xi = \|g\|_{\mathrm{C}^\xi(I)}$. Finally, for any normed spaces X and Y, we write $\|\cdot\|_{X\mapsto Y}$ to denote the operator norm of an operator mapping $X \mapsto Y$.

In what follows, we will allow data which is piecewise continuous with respect to an apriori defined partition

$$0 = c_1 < \cdots < c_\aleph \; = \; 1 \;, \tag{3.1.4}$$

with $\aleph \geq 1$. We denote the corresponding space of piecewise continuous functions by

$$\mathrm{C}_{pw} := \left\{ g \in L_\infty(0,1) \; | \; g|_{(c_j,c_{j+1})} \in \mathrm{C}(c_j,c_{j+1}), \; \text{for each } j = 1,\ldots,\aleph - 1 \right\} \;.$$

By the assumed uniform continuity, g then has well-defined left and right limits (which might be different) at each c_j. For definiteness we will assume that the value of $g(c_j)$ is taken to be the limit from the right for $j = 1, \cdots, \aleph - 1$ and the limit from the left for $j = \aleph$. The space C_{pw} is equipped with the usual uniform norm $\| \cdot \|_\infty$. Similarly, for any $\xi \in (0, 1]$, let

$$C_{pw}^\xi := \left\{ g \in C_{pw} \mid g|_{(c_j, c_{j+1})} \in C^\xi(c_j, c_{j+1}), \text{ for each } j = 1, \ldots, \aleph - 1 \right\},$$

with norm $\|g\|_{\xi, pw} := \max_{j=1}^{\aleph} \|g\|_{C^\xi(c_j, c_{j+1})}$.

We now make the following physically motivated assumptions on the input data.

Assumption 3.1.2 *(Input Data)*

1. *The cross-sections σ_S and σ_A are strictly positive and bounded above. We write*

$$\sigma_{\min} = \min_{x \in [0,1]} \sigma(x), \qquad \sigma_{\max} = \max_{x \in [0,1]} \sigma(x),$$

$$(\sigma_S)_{\min} = \min_{x \in [0,1]} \sigma_S(x) \quad and \quad (\sigma_S)_{\max} = \max_{x \in [0,1]} \sigma_S(x).$$

2. *There exists a partition* (3.1.4) *and $\eta \in (0, 1)$, such that $\sigma, \sigma_S, f \in C_{pw}^\eta$.*

3.1.1 Discretisation

To discretise (3.1.1) – (3.1.3) in angle, we use a $2N$-point quadrature rule

$$\int_{-1}^{1} g(\mu) \, d\mu \approx \sum_{|k|=1}^{N} w_k g(\mu_k), \tag{3.1.5}$$

with nodes $\mu_k \in [-1, 1] \backslash \{0\}$ and positive weights $w_k \in \mathbb{R}^+$. We assume the (anti-)symmetry properties $\mu_{-k} = -\mu_k$ and $w_{-k} = w_k$. To discretise in space, we introduce a mesh

$$0 = x_0 < x_1 < \ldots < x_M = 1, \tag{3.1.6}$$

which is assumed to resolve the break points $\{c_j\}$ introduced in (3.1.4) (which requires that $M \geq \aleph$). Further assumptions on the quadrature rule and mesh will be added in Section 3.3.

Our discrete scheme for (3.1.1) – (3.1.3) is then: Find the family of continuous piecewise-linear functions $\{\psi_k^{h,N}\}_{k=1}^{2N}$ (with nodal values $\{\psi_{k,j}^{h,N}\}$) such that

$$\mu_k \frac{\psi_{k,j}^{h,N} - \psi_{k,j-1}^{h,N}}{h_j} + \sigma_{j-1/2} \frac{\psi_{k,j}^{h,N} + \psi_{k,j-1}^{h,N}}{2} = \sigma_{S,j-1/2} \phi_{j-1/2}^{h,N} + f_{j-1/2}, \tag{3.1.7}$$

for $j = 1, ..., M$, $|k| = 1, \ldots, N$, where

$$\phi_{j-1/2}^{h,N} = \frac{1}{2} \sum_{|k|=1}^{N} w_k \frac{\psi_{k,j}^{h,N} + \psi_{k,j-1}^{h,N}}{2}, \quad j = 1, ..., M, \tag{3.1.8}$$

and with

$$\psi_{k,0}^{h,N} = 0, \quad \text{for} \quad k > 0 \quad \text{and} \quad \psi_{k,M}^{h,N} = 0, \quad \text{for} \quad k < 0. \tag{3.1.9}$$

Here $\sigma_{j-1/2}$ denotes the value of σ at the mid-point of the interval $I_j = (x_{j-1}, x_j)$, with the analogous meaning for $\sigma_{S,j-1/2}$ and $f_{j-1/2}$.

Solution Methods for Slab Geometry Transport

If the right-hand side of (3.1.7) were known, then (3.1.7) could be solved simply by sweeping from left to right (when $k > 0$) and from right to left (when $k < 0$). The appearance of $\phi_{j-1/2}^{h,N}$ on the right-hand side means that (3.1.7) and (3.1.8) constitute a coupled system, which can be written in matrix form as

$$\begin{pmatrix} T & -\Sigma_S \\ -W & I \end{pmatrix} \begin{pmatrix} \Psi \\ \Phi \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix} . \tag{3.1.10}$$

Here, the vector $\Phi \in \mathbb{R}^M$ contains the approximations of the scalar flux at the M midpoints of the spatial mesh. The matrix T is a block diagonal $2NM \times 2NM$ matrix, representing the left hand side of (3.1.7). The $2N$ diagonal blocks of T, one per angle, are themselves bi-diagonal. The $2NM \times M$ matrix Σ_S simply consists of $2N$ identical (since we assumed the scattering cross-section was isotropic in angle) diagonal blocks, one per angle, with $\Sigma_S \Phi$ representing the multiplication of Φ by σ_S at the midpoints of the mesh. The $M \times 2NM$ matrix W (applied to Ψ) represents the right hand side of (3.1.8), i.e. averaging at the midpoints and quadrature. The matrix I denotes the $M \times M$ identity matrix. The vector $\mathbf{f} \in \mathbb{R}^{2NM}$ contains $2N$ copies of the source term evaluated at the M midpoints of the spatial mesh.

3.1.2 Direct and Iterative Solvers

We now wish to find the (approximate) fluxes in the linear system (3.1.10). We note that the matrix T is invertible and has a useful sparsity structure that allows its inverse to be calculated in $\mathcal{O}(MN)$ operations. However, the bordered system (3.1.10) is not as easy to invert, due to the presence of Σ_S and W.

To exploit the sparsity of T, we do block elimination on (3.1.10) obtaining the Schur complement system for the scalar flux, i.e.,

$$\left(I - WT^{-1}\Sigma_S\right)\Phi = WT^{-1}\mathbf{f} , \tag{3.1.11}$$

which now requires the inversion of a smaller (dense) matrix. Note that (3.1.11) is a finite-dimensional version of the reduction of the integro-differential equation (3.1.1), (3.1.2) to the integral form of the RTE (see ahead to (3.1.22)). In this case, the two dominant computations with $\mathcal{O}(M^2N)$ and $\mathcal{O}(M^3)$ operations respectively, are the triple matrix product $WT^{-1}\Sigma_S$ in the construction of the Schur complement and the LU factorisation of the $M \times M$ matrix $\left(I - WT^{-1}\Sigma_S\right)$. This leads to a total

$$\text{theoretical cost of the direct solver} \ \sim \ \mathcal{O}(M^2(M + N)) . \tag{3.1.12}$$

We note that for stability reasons (see [93, §3], also [163] in a simpler context), the number of spatial and angular points should be related. A suitable choice could be $M \sim N$, leading to a cost of the direct solver of $\mathcal{O}(M^3)$ in general.

The second approach for solving (3.1.10) is an iterative solver commonly referred to as *source iteration*, cf. [32]. Re-writing (3.1.11) as

$$\Phi = WT^{-1}\left(WT^{-1}\Sigma_S\Phi + \mathbf{f}\right) ,$$

then naturally suggests the iteration

$$\Phi^{(k)} \;=\; WT^{-1}\left(\Sigma_S \Phi^{(k-1)} \;+\; \mathbf{f}\right) \;, \tag{3.1.13}$$

where $\Phi^{(k)}$ is the approximation at the kth iteration, and where we assume $\Phi^{(0)} = WT^{-1}\mathbf{f}$. This can be seen as a discrete version of an iterative method for the integral equation (3.1.22).

In practice, we truncate after K iterations. The dominant computations in the source iteration are the K multiplications with $WT^{-1}\Sigma_S$. Exploiting the sparsity of all the matrices involved, these multiplications cost $\mathcal{O}(MN)$ operations, leading to an overall

$$\text{theoretical cost of source iteration} \;\sim\; \mathcal{O}\left(M\,N\,K\right) \;. \tag{3.1.14}$$

A proof of a similar result, for a spatially two-dimensional problem, is given in Section 5.2.3.

The numerical experiments in [93] show that, for $2N = M$, the hidden constants in the two estimates (3.1.12) and (3.1.14) are approximately the same. Hence, whether the iterative solver is faster than the direct solver depends on whether the number of iterations K to obtain an accurate enough solution is smaller or larger than M. This will motivate the introduction of a third 'hybrid' solver, which we discuss later in Section 4.2.

There are sharp theoretical results on the convergence of (a non-discrete version of) source iteration for piecewise smooth cross-sections [32, Thm 2.20]. In particular, if $\phi^{(K)}$ denotes the approximation to ϕ after K iterations, then

$$\left\|\sigma^{1/2}\left(\phi - \phi^{(K)}\right)\right\|_2 \;\leq\; c'\left(c\left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^K \;, \tag{3.1.15}$$

for some constants $c' > 0$ and $c \leq 1$. That is, the error decays geometrically with rate no slower than the spatial maximum of σ_S/σ. For the case where the cross-sections are random (e.g. (2.2.2) with zero fission), then (3.1.15) will hold pathwise for each ω - for some realisations the spatial maximum of σ_S/σ will be close to 1. Using this result as a guide together with the assumption $0 < \sigma_A(x) < \infty$, for all $x \in [0,1]$, we assume that the convergence of the L_2-error with respect to K can be bounded by

$$\|\phi \,-\, \phi^{(K)}\|_2 \;\leq\; c\left\|\frac{\sigma_S}{\sigma}\right\|_\infty^K \;, \tag{3.1.16}$$

for some constant $c > 0$.

3.1.3 Abstract form of the problem

As preparation for analysing (3.1.1) – (3.1.3) and its discretisation, (3.1.7) – (3.1.9), consider the *pure transport problem* (previously discussed in Section 1.3): For fixed $\mu \in [-1,1]$, find $u = u(x)$, $x \in (0,1)$, such that

$$\mu\frac{du}{dx} + \sigma u = g, \quad \text{with } u(0) = 0, \text{ when } \mu > 0 \quad \text{and } u(1) = 0 \text{ when } \mu < 0, \tag{3.1.17}$$

with $g \in L_\infty$ a generic right-hand side. (Note that u depends on μ, but we suppress this in the notation. When $\mu = 0$ no boundary condition is needed.)

Following Appendix A.2, it is easy to show that a solution of this problem is $u := \mathcal{S}_\mu g$, where

we define

$$
\mathcal{S}_\mu g(x) \; := \; \left\{ \begin{array}{ll} \mu^{-1} \int_0^x \exp\left(\mu^{-1}\tau(x,y)\right) g(y) \; dy \; , & \mu > 0 \\ \sigma^{-1}(x) \; g(x) \; , & \mu = 0 \\ -\mu^{-1} \int_x^1 \exp\left(\mu^{-1}\tau(x,y)\right) g(y) \; dy \; , & \mu < 0 \end{array} \right. , \tag{3.1.18}
$$

and

$$
\tau(x,y) \; := \; \int_x^y \sigma(s) \; ds \; . \tag{3.1.19}
$$

The quantity $|\tau(x,y)|$ is often called the 'optical length' or 'optical path' [27]. To mimic the averaging process in (3.1.2) it is natural to also consider the integral operator:

$$
\mathcal{K}g(x) \; := \; \frac{1}{2} \int_{-1}^1 \mathcal{S}_\mu g(x) \; d\mu \; = \; \frac{1}{2} \int_0^1 E_1(|\tau(x,y)|)g(y) \; dy \; , \tag{3.1.20}
$$

where for $z > 0$, $E_1(z)$ is the exponential integral

$$
E_1(z) \; := \; \int_1^\infty \exp(-tz)\frac{dt}{t} \; . \tag{3.1.21}
$$

The operators $\mathcal{S}_\mu$ and $\mathcal{K}$ relate to (3.1.1) – (3.1.3) by the following proposition (and which is analogous to Theorem A.2.1).

Proposition 3.1.3 *Let ψ be a solution to (3.1.1) – (3.1.3). Then, ψ is uniquely determined by*

$$
\psi(x,\mu) \; = \; \mathcal{S}_\mu\left(\sigma_S\phi \; + \; f\right)(x) \; , \tag{3.1.22}
$$

and hence ϕ solves the integral equation

$$
\phi \; = \; \mathcal{K}\left(\sigma_S\phi \; + \; f\right) \; . \tag{3.1.23}
$$

We shall see later that (3.1.23) has a unique solution and this ensures that (3.1.1) – (3.1.3) has a unique solution. We also discussed the uniqueness of the solution in Remark 1.1.3.

Analogously we can consider the discrete system (3.1.7) – (3.1.9). Let V^h denote the space of continuous piecewise-linear functions with respect to the mesh $\{x_j\}_{j=0}^M$, and for any $v \in C$, let $\mathcal{P}^h v$ denote the piecewise constant function which interpolates v at the mid-points of the subintervals $I_j = (x_{j-1}, x_j)$. Then consider the discretisation of (3.1.17) defined by seeking $u^h \in V^h$ to satisfy

$$
\int_{I_j} \left(\mu \frac{du^h}{dx} + \mathcal{P}^h \sigma \; u^h \right) \; = \; \int_{I_j} g \; , \quad \text{for } j = 1, \ldots, M \; , \tag{3.1.24}
$$

with $u^h(0) = 0$ when $\mu > 0$ and $u^h(1) = 0$ when $\mu < 0$. This has a unique solution (the proof of which is given later during the proof of Theorem 3.3.5), which we write as $u^h = \mathcal{S}_\mu^h g$. Analogously to (3.1.20) we also define

$$
\mathcal{K}^{h,N} g \; = \; \frac{1}{2} \sum_{|k|=1}^N w_k \; \mathcal{S}_{\mu_k}^h g \; . \tag{3.1.25}
$$

Identifying any fully discrete solution $\psi_{k,j}^{h,N}$ of (3.1.7) – (3.1.9) with the function $\psi_k^{h,N} \in V^h$ by interpolation at the nodes $\{x_j\}$, we can see that (3.1.7) – (3.1.9) is equivalent to seeking $\psi_k^{h,N} \in V^h$,

$|k| = 1, ..., N$, that satisfy

$$\int_{I_j} \left(\mu_k \frac{\mathrm{d}\psi_k^{h,N}}{\mathrm{d}x} + \mathcal{P}^h \sigma \, \psi_k^{h,N} \right) \; = \; \int_{I_j} \mathcal{P}^h \left(\sigma_S \phi^{h,N} + f \right) \; , \quad j = 1, \ldots, M \; , \qquad (3.1.26)$$

where

$$\phi^{h,N} \; := \; \frac{1}{2} \sum_{|k|=1}^{N} w_k \psi_k^{h,N} \; , \qquad (3.1.27)$$

and

$$\psi_k^{h,N}(0) \; = \; 0 \; , \quad \text{when} \quad k > 0 \quad \text{and} \quad \psi_k^{h,N}(1) \; = \; 0 \; , \quad \text{when} \quad k < 0 \; . \qquad (3.1.28)$$

We then have the discrete analogue of Proposition 3.1.3:

Proposition 3.1.4 *The system* (3.1.26) *–* (3.1.28) *is equivalent to* (3.1.7) *–* (3.1.9), *and its solution can be written:*

$$\psi_k^{h,N} \; = \; \mathcal{S}_{\mu_k}^h \mathcal{P}^h (\sigma_S \phi^{h,N} + f) \; , \quad \text{for all } |k| = 1, \cdots, N \; . \qquad (3.1.29)$$

Moreover,

$$\phi^{h,N} \; = \; \mathcal{K}^{h,N} \mathcal{P}^h (\sigma_S \phi^{h,N} \; + \; f) \; . \qquad (3.1.30)$$

Now to estimate the error in our approximation to ϕ, we write

$$
\begin{aligned}
\phi - \phi^{h,N} \; &= \; \mathcal{K}(\sigma_S \phi + f) - \mathcal{K}^{h,N} \mathcal{P}^h (\sigma_S \phi^{h,N} + f) \\
&= \; \mathcal{K}(\sigma_S \phi + f) - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \phi + \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \phi - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \phi^{h,N} - \mathcal{K}^{h,N} \mathcal{P}^h f \\
&= \; \left(\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h \right) (\sigma_S \phi + f) + \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \left(\phi - \phi^{h,N} \right) \\
&= \; (I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S)^{-1} (\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h)(\sigma_S \phi + f) \; , \qquad (3.1.31)
\end{aligned}
$$

where we subtracted (3.1.30) from (3.1.23) to obtain the first equality and added $(\mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \phi - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \phi) = 0$ for the second equality. The fourth equality holds by subtracting $\mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \left(\phi - \phi^{h,N} \right)$ from both sides to acquire

$$\left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \right) \left(\phi - \phi^{h,N} \right) \; = \; \left(\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h \right) \left(\sigma_S \phi + f \right) \; ,$$

and then assuming that $(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S)^{-1}$ is a bounded map on C (which is proven later in (3.3.29)). Hence,

$$\| \phi - \phi^{h,N} \|_\infty \; \leq \; \| (I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S)^{-1} \|_{\mathrm{C} \mapsto \mathrm{C}} \| (\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h)(\sigma_S \phi + f) \|_\infty \; . \qquad (3.1.32)$$

3.2 Properties of the Operators

In this section, we will prove a number of technical results which will lead to bounds on both terms on the right hand side of (3.1.32) (stability and consistency) and will thus yield the deterministic error estimate in (3.0.1) – (3.0.2).

Notation 3.2.1 *To simplify presentation, for any $a \in \mathbb{R}$, we will use the notation $\bar{a} := \max\{1, a\}$. Also, from now on, we will use c to denote a constant that is positive, finite and independent of*

the cross-sections, mesh parameters and other relevant variables.

Throughout this section, we will make use of the following bounds, a consequence of Assumptions 3.1.2 and (3.1.19),

$$\sigma_{\min}|y - x| \leq \text{sgn}(y - x)\tau(x, y) \leq \sigma_{\max}|y - x| \ , \tag{3.2.1}$$

where $\text{sgn}(\cdot) = 1$, when its argument is positive, and (-1) when negative.

Lemma 3.2.2 *For $\mu \in [-1, 1]$, then $\|\mathcal{S}_\mu\|_{L_\infty \mapsto L_\infty} \ \leq \ \sigma_{\min}^{-1}$.*

Proof. Consider $\mu > 0$ and a function $g \in L_\infty$. By definition,

$$
\begin{aligned}
|\mathcal{S}_\mu g(x)| \ &= \ \left| \mu^{-1} \int_0^x \exp(\mu^{-1}\tau(x, y))g(y) \ dy \right| \\
&\leq \ \|g\|_\infty \mu^{-1} \int_0^x \exp(\mu^{-1}\tau(x, y)) \ dy \\
&\leq \ \|g\|_\infty \mu^{-1} \int_0^x \exp(\mu^{-1}\sigma_{\min}(y - x)) \ dy \\
&\leq \ \|g\|_\infty \sigma_{\min}^{-1} \left[\exp(\mu^{-1}\sigma_{\min}(y - x)) \right]_{y=0}^{y=x} \\
&\leq \ \|g\|_\infty \sigma_{\min}^{-1} \ ,
\end{aligned}
$$

where we have used that $y \in [0, x]$ implies $\tau(x, y) \leq \sigma_{\min}(y - x) \leq 0$ and that the exponential function is non-negative and monotonically increasing. The proof for $\mu < 0$ holds similarly. ∎

In the following two Lemmas we study the differentiability of $\mathcal{S}_\mu$ with respect to x and μ. Through (3.1.22), this relates directly to the differentiability of the angular flux ψ and hence will be fundamental to the convergence rate of the deterministic error estimates in Section 3.3.

Lemma 3.2.3

$$\left\| \frac{\partial}{\partial x} \mathcal{S}_\mu \right\|_{L_\infty \mapsto L_\infty} \ \leq \ 2 \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) |\mu|^{-1} \ , \quad \text{for all } \ \mu \in [-1, 1] \backslash \{0\} \ .$$

Moreover, $\mathcal{S}_\mu : L_\infty \mapsto \text{C}$ with $\|\mathcal{S}_\mu\|_{L_\infty \mapsto \text{C}} \leq \sigma_{\min}^{-1}$, for $\mu \in [-1, 1] \backslash \{0\}$.

Proof. Consider $\mu > 0$ and some function $g \in L_\infty$. By the definition of $\mathcal{S}_\mu$ and the Leibniz integral rule

$$
\begin{aligned}
\frac{\partial}{\partial x} \mathcal{S}_\mu g(x) &= \mu^{-1} \frac{\partial}{\partial x} \left\{ \int_0^x \exp(\mu^{-1}\tau(x, y))g(y) \ dy \right\} \\
&= \mu^{-1} \left(g(x) - \mu^{-1}\sigma(x) \int_0^x \exp(\mu^{-1}\tau(x, y))g(y) \ dy \right) \ .
\end{aligned}
$$

Applying the absolute value and bounding then gives

$$
\begin{aligned}
|(\mathcal{S}_\mu g)'(x)| &\leq \mu^{-1} \left(|g(x)| + \mu^{-1}\sigma_{\max} \int_0^x \exp(\mu^{-1}\tau(x, y))|g(y)| \ dy \right) \\
&\leq \mu^{-1}\|g\|_\infty \left(1 + \mu^{-1}\sigma_{\max} \int_0^x \exp(\mu^{-1}\tau(x, y)) \ dy \right) \\
&\leq \mu^{-1}\|g\|_\infty \left(1 + \sigma_{\max}\sigma_{\min}^{-1} \right) \ ,
\end{aligned}
$$

where the integral is bounded as in Lemma 3.2.2. The proof for $\mu < 0$ is similar.

The proof that $\mathcal{S}_\mu : L_\infty \mapsto \mathrm{C}$ is a simple consequence of the fact that, given any $g \in L_\infty$, the derivative of $\mathcal{S}_\mu g$ is bounded (except when $\mu = 0$). The corresponding bound is as proven in Lemma 3.2.2. ■

Lemma 3.2.4 *For $g \in L_\infty$ and any $\beta > 0$, then*

$$\sup_{x \in [0,1]} \int_{-1}^1 |\mu|^\beta \left| \frac{\partial}{\partial \mu}(\mathcal{S}_\mu g)(x) \right| \mathrm{d}\mu \; \le \; 2\beta^{-1}\sigma_{\min}^{-1}\|g\|_\infty \; .$$

Proof. Consider $\mu > 0$. By the product rule

$$\frac{\partial}{\partial \mu} \left(\mu^{-1}\exp(\mu^{-1}\tau(x,y)) \right) = -\mu^{-2}\exp(\mu^{-1}\tau(x,y))\left(1 + \mu^{-1}\tau(x,y) \right) \; . \tag{3.2.2}$$

Using (3.2.2), the definition of $\mathcal{S}_\mu g$ and the substitution $t = \mu^{-1}\tau(x,y) = \mu^{-1}\tau_x(y)$, then

$$\begin{aligned}
-\frac{\partial}{\partial \mu}(\mathcal{S}_\mu g)(x) &= -\int_0^x \frac{\partial}{\partial \mu}\left(\mu^{-1}\exp\left(\mu^{-1}\tau(x,y) \right) \right) g(y)\, dy \\
&= \mu^{-2} \int_0^x [1 + \mu^{-1}\tau(x,y)]\exp[\mu^{-1}\tau(x,y)]g(y)\, dy \\
&= \mu^{-1} \int_{\mu^{-1}\tau(x,0)}^0 (1+t)\,\exp(t)\sigma^{-1}(\tau_x^{-1}(\mu t))\, g(\tau_x^{-1}(\mu t))\, dt \; .
\end{aligned}$$

The function τ_x^{-1} exists because $y \in (0,x)$, for fixed x, and therefore $\tau_x(y) < 0$ - existence of τ_x^{-1} then follows by the inverse function theorem. Applying the absolute value and using the triangle inequality (i.e. $|1 + t| \le 1 + |t|$),

$$\left| \frac{\partial}{\partial \mu}(\mathcal{S}_\mu g)(x) \right| \le \mu^{-1}\|g\|_\infty \int_{\mu^{-1}\tau(x,0)}^0 (1+|t|)\exp(t)\sigma^{-1}(\tau_x^{-1}(\mu t))\, dt \; ,$$

which is an integral with a positive integrand. Hence, by bounding $\sigma^{-1}(\cdot) \le \sigma_{\min}^{-1}$ we can acquire an upper bound on the integral, by extending the domain of integration i.e.

$$\left| \frac{\partial}{\partial \mu}(\mathcal{S}_\mu g)(x) \right| \; \le \; \mu^{-1}\sigma_{\min}^{-1}\|g\|_\infty \int_{-\infty}^0 (1+|t|)\exp(t)\, dt \; \le \; 2\mu^{-1}\sigma_{\min}^{-1}\|g\|_\infty \; ,$$

where we note that $\int_{-\infty}^0 (1+|t|)\exp(t)\, dt = 2$. This implies that for any $\beta > 0$ and any $x \in [0,1]$

$$\int_0^1 |\mu|^\beta \left| \frac{\partial}{\partial \mu}(\mathcal{S}_\mu g) \right| \mathrm{d}\mu \; \le \; 2\int_0^1 \mu^{-1+\beta}\sigma_{\min}^{-1}\|g\|_\infty\, \mathrm{d}\mu \; = \; 2\beta^{-1}\sigma_{\min}^{-1}\|g\|_\infty \; .$$

The proof holds similarly for $\mu \in [-1,0)$, and $\mu = 0$ is trivial. ■

The next few results will allow us to prove that

$$\sigma_S, \sigma, f \in \mathrm{C}_{pw}^\eta, \text{ for any } \eta \in (0,1) \quad \Rightarrow \quad \phi \in \mathrm{C}^\xi, \text{ for all } \xi \in (0,1).$$

This result uses the smoothing property of the averaging operator $\mathcal{K}$.

Lemma 3.2.5 *On L_2, the operator $\mathcal{K}\sigma_S$ is bounded and $(I - \mathcal{K}\sigma_S)$ is invertible with the bound*

$$\| (I - \mathcal{K}\sigma_S)^{-1} \|_{L_2 \mapsto L_2} \; \le \; (\sigma_{\max}/\sigma_{\min})^{1/2}(1 - \|\sigma_S/\sigma\|_\infty)^{-1} \; . \tag{3.2.3}$$

Moreover, the scalar flux $\phi \in L_2$ and

$$\|\phi\|_2 \;\leq\; (\sigma_{\max}/\sigma_{\min})\,(\sigma_S)_{\min}^{-1}\,(1 - \|\sigma_S/\sigma\|_\infty)^{-1}\,\|f\|_2 \;. \tag{3.2.4}$$

Proof. Consider the weighted norm $\|g\|_{L_2^w} := \|\sigma^{1/2}g\|_2$, for any $g \in L_2^w$. This is equivalent to the usual L_2-norm in the sense that

$$\sigma_{\min}^{1/2}\|g\|_2 \leq \|g\|_{L_2^w} \leq \sigma_{\max}^{1/2}\|g\|_2 \;, \tag{3.2.5}$$

and hence any $g \in L_2^w$ also belongs to L_2. Using the result of [32, Thm 2.20], i.e.

$$\|\mathcal{K}\sigma_S\|_{L_2^w \mapsto L_2^w} \;\leq\; \|\sigma_S/\sigma\|_\infty \;<\; 1 \;, \tag{3.2.6}$$

then $\mathcal{K}\sigma_S$ is bounded on L_2^w, and hence by (3.2.5),

$$\|\mathcal{K}\sigma_S\|_{L_2 \mapsto L_2} \;\leq\; (\sigma_{\max}/\sigma_{\min})^{1/2}\,\|\sigma_S/\sigma\|_\infty \tag{3.2.7}$$

since, for $g \in L_2^w$

$$\sigma_{\min}^{1/2}\|\mathcal{K}\sigma_S g\|_2 \leq \|\mathcal{K}\sigma_S g\|_{L_2^w} \leq \|\mathcal{K}\sigma_S\|_{L_2^w \mapsto L_2^w}\|g\|_{L_2^w} \leq \sigma_{\max}^{1/2}\|\sigma_S/\sigma\|_\infty\|g\|_2 \;.$$

Moreover, by the Banach Lemma it follows that

$$\|\,(I - \mathcal{K}\sigma_S)^{-1}\,\|_{L_2^w \mapsto L_2^w} \leq \left(1 - \|\mathcal{K}\sigma_S\|_{L_2^w \mapsto L_2^w}\right)^{-1} \leq \left(1 - \|\sigma_S/\sigma\|_\infty\right)^{-1} \;. \tag{3.2.8}$$

The bound on $\|\,(I - \mathcal{K}\sigma_S)^{-1}\,\|_{L_2 \mapsto L_2}$ follows from (3.2.5) and (3.2.8).

Re-writing the integral equation (3.1.23) as $(I - \mathcal{K}\sigma_S)\,\phi = \mathcal{K}f$ and using (3.2.3), then

$$\begin{aligned}
\|\phi\|_2 &= \|\,(I - \mathcal{K}\sigma_S)^{-1}\,\mathcal{K}f\|_2 \;\leq\; \|\,(I - \mathcal{K}\sigma_S)^{-1}\,\|_{L_2 \mapsto L_2}\|\mathcal{K}f\|_2 \\
&\leq\; (\sigma_{\max}/\sigma_{\min})^{1/2}\,(1 - \|\sigma_S/\sigma\|_\infty)^{-1}\,\|\mathcal{K}f\|_2 \;.
\end{aligned}$$

Moreover, using the aforementioned bound on $\|\mathcal{K}\sigma_S\|_{L_2}$ i.e. (3.2.7),

$$\|\mathcal{K}f\|_2 \;\leq\; \|\mathcal{K}\sigma_S\|_{L_2 \mapsto L_2}\|\,(\sigma_S)^{-1}\,f\|_2 \;\leq\; (\sigma_{\max}/\sigma_{\min})^{1/2}\,\|\sigma_S/\sigma\|_\infty(\sigma_S)_{\min}^{-1}\|f\|_2 \;.$$

The bound (3.2.4) follows. ∎

Before we continue with our results on the operator $\mathcal{K}$, we must prove the following preliminary results relating to its integrand $E_1(\cdot)$.

Lemma 3.2.6 *For any $x \in (0,1)$ let $\delta > 0$ be such that $x + \delta \in (0,1]$. For $y \in [0,1]\backslash[x, x+\delta]$,*

$$\left| E_1(|\tau(x+\delta,y)|) \;-\; E_1(|\tau(x,y)|) \right| \;\leq\; \delta\,\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)\,\frac{1}{\min\{|y - (x+\delta)|,\ |y-x|\}} \;.$$

Proof. Using that $-E_1'(z) = E_0(z) := \exp(z)/z$ for $z > 0$ [1, eq.(5.1.26), pg.230] and the chain rule, then for $x \neq y$,

$$\frac{d}{dx}E_1(|\tau(x,y)|) \;=\; -\frac{\exp(-|\tau(x,y)|)}{|\tau(x,y)|}\,\frac{d}{dx}|\tau(x,y)| \;=\; \sigma(x)\mathrm{sgn}(y-x)\frac{\exp(-|\tau(x,y)|)}{|\tau(x,y)|} \;. \tag{3.2.9}$$

By the Mean Value Theorem and since $E_1(\cdot)$ is continuous on $(0,\infty)$, there exists $c \in (x, x+\delta)$

such that

$$E_1(|\tau(x+\delta,y)|) - E_1(|\tau(x,y)|) \;=\; \delta\mathrm{sgn}(y-c)\sigma(c)\frac{\exp(-|\tau(c,y)|)}{|\tau(c,y)|}\;.$$

Now when $y > x + \delta$, we have $\tau(c,y) \;=\; \int_c^y \sigma(z)\,dz \;\geq\; \sigma_{\min}(y-x-\delta) > 0$, and hence

$$\left| E_1|\tau(x+\delta,y)| \;-\; E_1|\tau(x,y)| \right| \;\leq\; \delta\frac{\sigma(c)}{\sigma_{\min}}\frac{\exp(-|\tau(c,y)|)}{(y-x-\delta)} \;\leq\; \delta\frac{\sigma_{\max}}{\sigma_{\min}}\frac{1}{(y-x-\delta)}\;.$$

The result for $y > x + \delta$ follows, and the result for $y < x$ holds similarly. ∎

In the next proof, we shall use the expansion (cf. [145, Eq. (1.8)], [1, Eq. (5.1.11)])

$$E_1(z) \;=\; \log(z) + \mathrm{Ein}(z) + \gamma^{\mathrm{E}}\;, \tag{3.2.10}$$

where $\gamma^{\mathrm{E}} \approx 0.5772$ is Euler's constant and

$$\mathrm{Ein}(z) \;:=\; \int_0^z \frac{1}{t}\left(1 - \exp(-t)\right)\,dt\;, \quad \text{for all } z > 0\,. \tag{3.2.11}$$

Elementary calculus shows that $\mathrm{Ein}(z) \leq z$, for all $z > 0$ - a proof is given in Section E.1.

Theorem 3.2.7 *The operator $\mathcal{K}$ maps L_2 to L_∞, and L_∞ to C^ξ, for all $0 < \xi < 1$. Moreover, the following bounds hold:*

(i) $\|\mathcal{K}\|_{L_2 \mapsto L_\infty} \;\leq\; \sigma_{\min}^{-1/2}$ *;*

(ii) $\|\mathcal{K}\|_{L_\infty \mapsto C^\xi} \;\leq\; c\,\dfrac{\overline{\sigma_{\max}}}{\sigma_{\min}}$ *.*

Proof. *(i)* Let $g \in L_2$ and $x \in [0,1]$. Using (3.1.20) and the Cauchy-Schwarz inequality, we have

$$2|\mathcal{K}g(x)| \;=\; \left| \int_0^1 E_1(|\tau(x,y)|)g(y)\,dy \right| \;\leq\; \|g\|_2 \left(\int_0^1 E_1^2(|\tau(x,y)|)\,dy \right)^{1/2}\;.$$

Since E_1^2 is strictly positive and monotonically decreasing on $\mathbb{R}^+$, we have (recalling (3.2.1))

$$\int_0^1 E_1^2(|\tau(x,y)|)\,dy \;\leq\; \int_0^1 E_1^2(\sigma_{\min}|x-y|)\,dy \;=\; \int_0^x E_1^2(\sigma_{\min}(x-y))\,dy + \int_x^1 E_1^2(\sigma_{\min}(y-x))\,dy\;. \tag{3.2.12}$$

Applying the substitution $r = x - y$ and using (3.1.21), the first integral on the right hand side of (3.2.12) becomes

$$\begin{aligned}
\int_0^x E_1^2(\sigma_{\min}r)\,dr &= \int_0^x \left(\int_1^\infty \exp\left(-s\sigma_{\min}r\right)s^{-1}\,ds \right)\left(\int_1^\infty \exp\left(-t\sigma_{\min}r\right)t^{-1}\,dt \right)\,dr \\
&= \int_1^\infty s^{-1} \int_1^\infty t^{-1} \left(\int_0^x \exp\left(-(t+s)\sigma_{\min}r\right)\,dr \right)\,dt\,ds \\
&= \int_1^\infty s^{-1} \int_1^\infty t^{-1} \left(\sigma_{\min}^{-1}\frac{1}{t+s}\left(1 - \exp\left(-(t+s)\sigma_{\min}x\right)\right) \right)\,dt\,ds \\
&\leq \sigma_{\min}^{-1} \int_1^\infty s^{-1} \int_1^\infty \frac{1}{t(t+s)}\,dt\,ds \;=\; 2\log(2)\sigma_{\min}^{-1}\;,
\end{aligned}$$

where the integrals can be interchanged by Fubini's theorem and the final equality holds by using partial fractions and integrating by parts. The result is also similar, up to a change in variables, to [1, eq.(5.1.33)].

The second integral on the right hand side of (3.2.12) can be bounded analogously and hence $\mathcal{K}: L_2 \mapsto L_\infty$, with $\|\mathcal{K}\|_{L_2 \mapsto L_\infty} \leq \log(2)^{1/2}\sigma_{\min}^{-1/2} \leq \sigma_{\min}^{-1/2}$.

(ii) Similarly, for any $g \in L_\infty$ and $x \in [0,1]$, then

$$2|\mathcal{K}g(x)| \leq \|g\|_\infty \int_0^1 E_1(|\tau(x,y)|) \, dy \leq \|g\|_\infty \int_0^1 E_1(\sigma_{\min}|y - x|) \, dy \,,$$

where we have used that the exponential integral is positive and monotonically decreasing. We will focus on the integral for $y \in [0, x]$ but the proof holds identically for the other case. Consider the substitution $r = x - y$, then by Fubini's theorem

$$\begin{aligned}
\int_0^x E_1(\sigma_{\min}|y - x|) \, dy &= \int_0^x E_1(\sigma_{\min} r) \, dr \\
&= \int_0^x \left(\int_1^\infty s^{-1} \exp(-s\sigma_{\min} r) \, ds \right) dr \\
&= \int_1^\infty s^{-1} \int_0^x \exp(-s\sigma_{\min} r) \, dr \, ds \\
&= \int_1^\infty s^{-1} \left[s^{-1}\sigma_{\min}^{-1} \exp(-s\sigma_{\min} r) \right]_{r=0}^{r=x} \, ds \\
&\leq \int_1^\infty s^{-2}\sigma_{\min}^{-1} \, ds = \sigma_{\min}^{-1} \,.
\end{aligned}$$

Hence,

$$\|\mathcal{K}g\|_\infty \leq \sigma_{\min}^{-1}\|g\|_\infty \,. \tag{3.2.13}$$

To bound the Hölder-seminorm, let $\xi \in (0,1)$ and consider

$$\sup_{x < z \leq 1} \frac{|\mathcal{K}g(z) - \mathcal{K}g(x)|}{|z - x|^\xi} = \max \left\{ \sup_{0 < \delta < \epsilon} \frac{|\mathcal{K}g(x + \delta) - \mathcal{K}g(x)|}{\delta^\xi} \,, \sup_{\epsilon \leq \delta \leq 1 - x} \frac{|\mathcal{K}g(x + \delta) - \mathcal{K}g(x)|}{\delta^\xi} \right\}, \tag{3.2.14}$$

where we have defined $\epsilon := (2\sigma_{\max})^{-1}$ and $\delta = \delta(x, z) := z - x$.

Let us first consider the case $\delta \in [\epsilon, 1 - x]$, i.e. the second term on the right hand side of (3.2.14). In that case, it follows by a simple application of the triangle inequality and the bound (3.2.13) that

$$\frac{|\mathcal{K}g(x + \delta) - \mathcal{K}g(x)|}{\delta^\xi} \leq 2\epsilon^{-\xi}\|\mathcal{K}g\|_\infty \leq 2(2\sigma_{\max})^\xi \sigma_{\min}^{-1}\|g\|_\infty \leq 4\frac{\sigma_{\max}}{\sigma_{\min}}\|g\|_\infty \,. \tag{3.2.15}$$

Now, let us consider the case $\delta \in (0, \epsilon)$ and let $B_\delta(x)$ be the closed ball of radius δ around x. To estimate the first term on the right hand side of (3.2.14), define $\mathcal{I}_\delta(x) := [0,1] \backslash B_\delta(x)$ and $\mathcal{J}_\delta(x) := B_\delta(x) \cap [0,1]$. Then,

$$\begin{aligned}
2\left| \mathcal{K}g(x + \delta) - \mathcal{K}g(x) \right| &= \left| \int_0^1 F(x, y, \delta)g(y) \, dy \right| \\
&\leq \|g\|_\infty \left(\int_{\mathcal{I}_\delta(x)} + \int_{\mathcal{J}_\delta(x)} \right) |F(x, y, \delta)| \, dy \,, \tag{3.2.16}
\end{aligned}$$

where $F(x, y, \delta) := E_1(|\tau(x + \delta, y)|) - E_1(|\tau(x, y)|)$. Consider the integral over $\mathcal{I}_\delta(x)$ in (3.2.16),

$$\int_{\mathcal{I}_\delta(x)} |F(x, y, \delta)| \, dy = \left(\int_0^{x-\delta} + \int_{x+\delta}^1 \right) |F(x, y, \delta)| \, dy \,, \tag{3.2.17}$$

and assume the first (second) integral on the right hand side is null when $x < \delta$ $(1 - x < \delta)$.

In the case where $x \geq \delta$, the first integral to the right of (3.2.17) is bounded by

$$
\begin{aligned}
\int_0^{x-\delta} |F(x,y,\delta)| \; dy \; &\leq \; \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) \delta \int_0^{x-\delta} (x-y)^{-1} \; dy \\
&= \; \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) \delta \left(\log \left(\frac{1}{\delta} \right) + \log x \right) \\
&\leq \; \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) \delta \log \left(\frac{1}{\delta} \right) \\
&\leq \; \frac{\sigma_{\max}}{\sigma_{\min}} \delta^\xi \; ,
\end{aligned}
$$

where we have used the result of Lemma 3.2.6 and the fact that $\delta \log(1/\delta) \leq \delta^\xi$, for all $\delta \leq 1$ and $\xi \in (0,1)$. The second integral on the right of (3.2.17) has the same estimate, when $\delta \leq 1 - x$. Hence,

$$
\int_{\mathcal{I}_\delta(x)} |F(x,y,\delta)| \; \mathrm{d}y \; \leq \; 2 \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) \delta^\xi \; . \tag{3.2.18}
$$

Now consider the integral over $\mathcal{J}_\delta(x)$ in (3.2.16), then

$$
\begin{aligned}
\int_{\mathcal{J}_\delta(x)} |F(x,y,\delta)| \; dy \; &\leq \; \int_{\mathcal{J}_\delta(x)} |E_1(|\tau(x+\delta,y)|)| \; dy + \int_{\mathcal{J}_\delta(x)} |E_1(|\tau(x,y)|)| \; dy \\
&\leq \; \int_{B_{2\delta}(x+\delta)} |E_1(|\tau(x+\delta,y)|)| \; dy + \int_{B_\delta(x)} |E_1(|\tau(x,y)|)| \; dy \; , \tag{3.2.19}
\end{aligned}
$$

where we note the subtle change in the limits of integration, that do not necessarily require the balls $B_{2\delta}(x+\delta)$ and $B_\delta(x)$ to be contained in $[0,1]$.

Using the expansion (3.2.10) for $E_1(z)$, the second integral on the right hand side of (3.2.19) can be re-written as

$$
\int_{B_\delta(x)} |E_1(|\tau(x,y)|)| \; dy \leq \int_{B_\delta(x)} |\log(|\tau(x,y)|)| \; dy + \int_{B_\delta(x)} |\mathrm{Ein}(|\tau(x,y)|)| \; dy + 2\gamma^{\mathrm{E}}\delta \; , \tag{3.2.20}
$$

Then, by (3.2.1) and using the fact that $\mathrm{Ein}(z) \in [0,z]$, for all $z > 0$, we can bound the second integral on the right hand side of (3.2.20)

$$
\int_{B_\delta(x)} |\mathrm{Ein}(|\tau(x,y)|)| \; dy \; \leq \; \int_{B_\delta(x)} |\tau(x,y)| \; \mathrm{d}y \; \leq \; 2\sigma_{\max}\delta^2 \; \leq \; \delta \; , \tag{3.2.21}
$$

where in the last step we used $\delta < \epsilon = (2\sigma_{\max})^{-1}$. For the first integral on the right hand side of (3.2.20), we use again (3.2.1) and the fact that $|\tau(x,y)| \leq \sigma_{\max}|x-y| < \epsilon\sigma_{\max} \leq 1/2$ to bound

$$
|\log|\tau(x,y)|| = -\log|\tau(x,y)| \leq -\log \left(\sigma_{\min}|y-x| \right) \; , \tag{3.2.22}
$$

since $-\log(x) = \log(x^{-1})$ and $|\tau(x,y)| \geq \sigma_{\min}|y-x|$. Thus, the first integral to the right of (3.2.20)

can be bounded by integrating the right hand side of (3.2.22), which gives

$$
\int_{B_\delta(x)} |\log|\tau(x,y)|| \, dy \;\leq\; -2 \int_0^\delta \log\left(\sigma_{\min} r\right) dr \tag{3.2.23}
$$

$$
\leq\; 2\delta\left(1 - \log\left(\sigma_{\min}\delta\right)\right)
$$

$$
\leq\; 2\delta^\xi \sigma_{\min}^{\xi-1} \frac{1}{\left(\frac{1}{\sigma_{\min}\delta}\right)^{1-\xi}} \left(1 + \log\left(\frac{1}{\sigma_{\min}\delta}\right)\right)
$$

$$
\leq\; c\, \delta^\xi \sigma_{\min}^{\xi-1} \;, \tag{3.2.24}
$$

where the third line holds by multiplying by $(\delta\sigma_{\min})^{1-\xi}/(\delta\sigma_{\min})^{1-\xi}$ and where the penultimate line used that $y^{-r}\left(1 + \log(y)\right)$ is bounded for any $y \geq 1$ and $r > 0$. Substituting the bounds in (3.2.21) and (3.2.24) into (3.2.20) we finally obtain

$$
\int_{B_\delta(x)} |E_1(|\tau(x,y)|)| \, dy \;\leq\; c\, \left(1 + \sigma_{\min}^{\xi-1}\right) \delta^\xi \;. \tag{3.2.25}
$$

The first integral on the right hand side of (3.2.19) can be bounded analogously. Thus, combining (3.2.16), (3.2.18), (3.2.19) and (3.2.25), we have shown that for $\delta \leq \epsilon$,

$$
\frac{|\mathcal{K}g(x+\delta) - \mathcal{K}g(x)|}{\delta^\xi} \;\leq\; c\, \frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \|g\|_\infty \;, \tag{3.2.26}
$$

uniformly in x, where we used that $\sigma_{\min}^{\xi-1} \leq (\overline{\sigma_{\max}}/\sigma_{\min})^{1-\xi} \leq \overline{\sigma_{\max}}/\sigma_{\min}$. Finally, combining (3.2.15) and (3.2.26) with the bound on $\|\mathcal{K}g\|_\infty$ we have

$$
\|\mathcal{K}g\|_\xi \;=\; \|\mathcal{K}g\|_\infty + \sup_{x,z\in[0,1]} \frac{|\mathcal{K}g(z) - \mathcal{K}g(x)|}{|z-x|^\xi} \;\leq\; c\, \frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \|g\|_\infty \;.
$$

$\blacksquare$

Lemma 3.2.8 *The operator $(I - \mathcal{K}\sigma_S)$ is invertible on C with the bound*

$$
\| (I - \mathcal{K}\sigma_S)^{-1} \|_{C\mapsto C} \;\leq\; 2\overline{\sigma_{\max}}^{1/2} \frac{\sigma_{\max}}{\sigma_{\min}} \left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1} \;=:\; \mathcal{R}_1(\sigma,\sigma_S) \;. \tag{3.2.27}
$$

Moreover, $(I - \mathcal{K}\sigma_S)$ is also invertible on C_{pw}, with the same bound as above.

Proof. Let $g \in C$ and suppose that

$$
(I - \mathcal{K}\sigma_S)\, v = g, \quad \text{or equivalently that} \quad v = \mathcal{K}\sigma_S v + g. \tag{3.2.28}
$$

This allows us to apply a bootstrapping argument. By Lemma 3.2.5, we have $v = (I - \mathcal{K}\sigma_S)^{-1} g \in L_2$ and using (3.2.3)

$$
\|v\|_2 \leq \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{1/2} \left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1} \|g\|_2 \leq \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{1/2} \left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1} \|g\|_\infty \;. \tag{3.2.29}
$$

Using (3.2.28) again, this time together with Theorem 3.2.7(i) we get $v \in L_\infty$, with the following bound

$$
\|v\|_\infty \leq \|\mathcal{K}\sigma_S v\|_\infty + \|g\|_\infty \leq \sigma_{\min}^{-1/2}(\sigma_S)_{\max}\|v\|_2 + \|g\|_\infty \leq \sigma_{\max}^{1/2}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{1/2}\|v\|_2 + \|g\|_\infty \;. \tag{3.2.30}
$$

Finally, using (3.2.28) with Theorem 3.2.7(ii) we conclude that $v \in C$ and therefore, since the supremum of a continuous function is the same as the essential supremum, the bound (3.2.30) holds on C. The bound in (3.2.27) follows by combining (3.2.29) and (3.2.30).

Now suppose $g \in C_{pw}$. Then $g \in L_2$ and the argument above holds verbatim to show that then $v \in C_{pw}$ and that the bounds in (3.2.29) and (3.2.30) hold again. ∎

The final result of this section follows from Theorem 3.2.7 and Lemma 3.2.8.

Corollary 3.2.9 *Let ϕ be the solution to (3.1.23), with $f \in C_{pw}^\eta$, for some $\eta \in (0,1)$ (Assumption 3.1.2). Then $\phi \in C^\xi$, for any $\xi \in (0,1)$, and the following two bounds hold:*

$$\|\phi\|_\infty \;\leq\; c\,\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\mathcal{R}_1(\sigma,\sigma_S)\|f\|_\infty \qquad and \qquad \|\phi\|_\xi \;\leq\; \mathcal{R}_2(\sigma,\sigma_S)\|f\|_\infty \;, \tag{3.2.31}$$

where $\mathcal{R}_1(\sigma,\sigma_S)$ is defined in (3.2.27) and

$$\mathcal{R}_2(\sigma,\sigma_S) \;:=\; c\,\frac{\overline{\sigma_{\max}}^3}{\sigma_{\min}^2}\,\mathcal{R}_1(\sigma,\sigma_S) \;. \tag{3.2.32}$$

Proof. Since, $C_{pw}^\eta \subset C_{pw}$ and $(I - \mathcal{K}\sigma_S)\phi = \mathcal{K}f$ the first bound in (3.2.31) follows directly from Lemma 3.2.8 and Theorem 3.2.7(ii), i.e.

$$\|\phi\|_\infty \leq \| (I - \mathcal{K}\sigma_S)^{-1} \|_{C \mapsto C}\|\mathcal{K}f\|_\infty \;\leq\; \mathcal{R}_1(\sigma,\sigma_S)\|\mathcal{K}\|_{L_\infty \mapsto C}\|f\|_\infty \;\leq\; \mathcal{R}_1(\sigma,\sigma_S)\left(c\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\right)\|f\|_\infty \;.$$

To obtain the second bound, we use Theorem 3.2.7(ii) again to obtain, for $\xi \leq \eta$,

$$\|\phi\|_\xi \;\leq\; \|\mathcal{K}\,(\sigma_S\phi + f)\|_\xi \;\leq\; \|\mathcal{K}\|_{L_\infty \mapsto C^\xi}\,\|\sigma_S\phi + f\|_\infty \;\leq\; c\,\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\left((\sigma_S)_{\max}\|\phi\|_\infty + \|f\|_\infty\right) \;,$$

and then combine this with the first bound in (3.2.31). We used again that $(\sigma_S)_{\max} \leq \sigma_{\max} \leq \overline{\sigma_{\max}}$. ∎

3.3 Deterministic Error Estimate

We now return to estimating the error $\phi - \phi^{h,N}$ using the formula (3.1.31). Introducing the operator

$$\mathcal{K}^N g(x) \;:=\; \frac{1}{2}\sum_{|k|=1}^{N} w_k\left(\mathcal{S}_{\mu_k}g\right)(x) \;=\; \frac{1}{2}\int_0^1 E_1^N(|\tau(x,y)|)g(y)\,dy \;, \tag{3.3.1}$$

with $E_1^N(z) := \sum_{k=1}^{N}\mu_k^{-1}w_k\exp(-\mu_k^{-1}z)$ denoting the N-point quadrature approximation of the exponential integral (3.1.21), we can write (3.1.31) as

$$\phi - \phi^{h,N} \;=\; \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}\left(e^N + e^{h,N}\right) \;, \tag{3.3.2}$$

where

$$e^N := \left(\mathcal{K} - \mathcal{K}^N\right)(\sigma_S\phi + f) \quad and \quad e^{h,N} = \left[\left(\mathcal{K}^N - \mathcal{K}^{h,N}\right) + \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\right)\right](\sigma_S\phi + f) \;. \tag{3.3.3}$$

We note that for $g \in L_\infty$,

$$\left(\mathcal{K} - \mathcal{K}^N\right)g(x) \;=\; \frac{1}{2}\int_{-1}^{1}\mathcal{S}_\mu g(x)\,d\mu - \frac{1}{2}\sum_{|k|=1}^{N} w_k\mathcal{S}_{\mu_k}g(x) \;. \tag{3.3.4}$$

Hence, by setting $g = \sigma_S \phi + f$, e^N can then be written as

$$e^N = \frac{1}{2} \int_{-1}^{1} \psi(x, \mu) \, d\mu - \frac{1}{2} \sum_{|k|=1}^{N} w_k \psi(x, \mu_k) \, ,$$

i.e. e^N is the error in approximating $\phi(x)$ by quadrature in angle. Likewise we note that by (3.1.25) and (3.3.1),

$$\left(\mathcal{K}^N - \mathcal{K}^{h,N} \right) g = \frac{1}{2} \sum_{|k|=1}^{N} w_k \left(\mathcal{S}_{\mu_k} - \mathcal{S}_{\mu_k}^h \right) g \, , \tag{3.3.5}$$

and

$$\mathcal{K}^{h,N} \left(I - \mathcal{P}^h \right) g = \frac{1}{2} \sum_{|k|=1}^{N} w_k \mathcal{S}_{\mu_k}^h \left(g - \mathcal{P}^h g \right) \, . \tag{3.3.6}$$

Finally, to obtain an error estimate we apply the supremum norm to (3.3.2), and by trivial manipulation write

$$\| \phi - \phi^{h,N} \|_\infty \leq \| \left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \right)^{-1} \|_{\mathrm{C} \mapsto \mathrm{C}} \left(\| e^N \|_\infty + \| e^{h,N} \|_\infty \right) \, . \tag{3.3.7}$$

Then the error estimate follows by showing that $\| e^N \|_\infty$ and $\| e^{h,N} \|_\infty$ both approach zero as $h \to 0$, $N \to \infty$ in an appropriate way and by finding a bound on $\| \left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S \right)^{-1} \|_{\mathrm{C} \mapsto \mathrm{C}}$. The first (the consistency) we do in Sections 3.3.1 and 3.3.2, while the second (the stability) is done in Section 3.3.3.

3.3.1 Consistency under Angular Discretisation

The convergence of the (quadrature) error e^N will rely on the regularity (in angle) of the solution (Lemma 3.2.4) and the following result from De Vore and Scott [55] (see also [163, Prop. 3.2]):

Proposition 3.3.1 *Consider the N-point Gauss-Legendre rule on $[0,1]$ and let m be a positive integer with $m \leq 2N - 1$. Then we have*

$$\left| \int_0^1 g(\mu) \, d\mu - \sum_{k=1}^{N} w_k g(\mu_k) \right| \leq c N^{-m} \int_0^1 [\mu (1 - \mu)]^{m/2} |g^{(m)}(\mu)| \, d\mu \, ,$$

whenever the integral on the right hand side exists.

Notation 3.3.2 *The* double Gauss rule *is the particular case of* (3.1.5) *where the N-point Gauss-Legendre rule is used on $[-1, 0]$ and on $[0, 1]$.*

We restrict the analysis to this rule from now on. We also note that the double Gauss rule satisfies all assumptions made on the quadrature rule in this text, see [163].

Theorem 3.3.3 *Let $\mathcal{K}^N$ be defined by* (3.3.1) *using the double Gauss rule. Then,*

$$\| \mathcal{K} - \mathcal{K}^N \|_{L_\infty \mapsto \mathrm{C}} \leq c \, \sigma_{\min}^{-1} N^{-1} \, . \tag{3.3.8}$$

Proof. Using (3.3.4), the (anti-)symmetry properties of the double Gauss rule, Proposition 3.3.1

(with $m = 1$) and Lemma 3.2.4 (with $\beta = 1/2$), we obtain for any $g \in L_\infty$,

$$
\begin{aligned}
\left| \left(\mathcal{K} - \mathcal{K}^N \right) g(x) \right| &= \frac{1}{2} \left| \int_0^1 \left(\mathcal{S}_\mu + \mathcal{S}_{-\mu} \right) g(x) \, d\mu - \sum_{k=1}^N \left(w_k \mathcal{S}_{\mu_k} + w_{-k} \mathcal{S}_{\mu_{-k}} \right) g(x) \right| \\
&\leq \frac{1}{2} \left| \int_0^1 \mathcal{S}_\mu g(x) \, d\mu - \sum_{k=1}^N w_k \mathcal{S}_{\mu_k} g(x) \right| + \frac{1}{2} \left| \int_0^1 \mathcal{S}_{-\mu} g(x) \, d\mu - \sum_{k=1}^N w_k \mathcal{S}_{-\mu_k} g(x) \right| \\
&\leq c \, N^{-1} \left[\int_0^1 \mu^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x) \right) \right| \, d\mu + \int_{-1}^0 (-\mu)^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x) \right) \right| \, d\mu \right] \\
&\leq c \, N^{-1} \int_{-1}^1 |\mu|^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x) \right) \right| \, d\mu \\
&\leq c \, N^{-1} \sigma_{\min}^{-1} \| g \|_\infty \ .
\end{aligned}
$$

Hence, $\left(\mathcal{K} - \mathcal{K}^N \right) : L_\infty \mapsto L_\infty$ satisfies the bound in (3.3.8). The extension to C holds because $\mathcal{S}_\mu$ maps from L_∞ to C, see Lemma 3.2.3. ∎

Corollary 3.3.4 *Under the conditions of Theorem 3.3.3, $e^N \in C$ with the bound*

$$
\| e^N \|_\infty \ \leq \ c \ \left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \right)^2 \mathcal{R}_1(\sigma, \sigma_S) N^{-1} \| f \|_\infty \ .
$$

Proof. By Theorem 3.3.3, (3.3.3) and Corollary 3.2.9, we obtain

$$
\begin{aligned}
\| e^N \|_\infty &\leq c N^{-1} \sigma_{\min}^{-1} \left(\sigma_{\max} \| \phi \|_\infty + \| f \|_\infty \right) \\
&\leq c N^{-1} \sigma_{\min}^{-1} \left(\sigma_{\max} \frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \mathcal{R}_1(\sigma, \sigma_S) + 1 \right) \| f \|_\infty
\end{aligned}
$$

from which the estimate follows. ∎

3.3.2 Consistency under Spatial Discretisation

We recall the operator $\mathcal{P}^h$ (defined in Section 3.1.3) and note that it maps C_{pw} to C_{pw}, and for any $g \in C_{pw}$, $\| \mathcal{P}^h g \|_\infty \leq \| g \|_\infty$. Moreover, for any $g \in C_{pw}^\xi$ with $0 < \xi \leq 1$,

$$
\| (I - \mathcal{P}^h) g \|_\infty \ \leq \ h^\xi \| g \|_{\xi, pw} \ . \tag{3.3.9}
$$

The simple proof of (3.3.9) is given in Section E.3.

From now on we assume that our mesh $\{ x_j \}_{j=0}^M$ is quasi-uniform, i.e., for some constant $\rho \geq 1$, the subinterval lengths $h_j := x_j - x_{j-1}$ satisfy

$$
\max_{j=1,\ldots,M} h_j \ =: \ h \ \leq \ \rho \min_{j=1,\ldots,M} h_j \ . \tag{3.3.10}
$$

Lemma 3.3.5 *Let $\mu \in [-1,1] \backslash \{0\}$. For $\mathcal{S}_\mu^h$ defined by (3.1.24), $\mathcal{S}_\mu^h : L_\infty \mapsto V^h \subset C$, and*

$$
\| \mathcal{S}_\mu^h \|_{L_\infty \mapsto V^h} \ \leq \ 2\rho \sigma_{\min}^{-1} \left(1 + \sigma_{\max} \frac{h}{|\mu|} \right) \ .
$$

Proof. Without loss of generality, assume $\mu > 0$ and let $g \in L_\infty$. Using the notation $\alpha_j = h_j \sigma_{j-1/2} / (2\mu)$, we can re-write (3.1.24) as

$$
\mu \left(1 + \alpha_j \right) U_j = \mu \left(1 - \alpha_j \right) U_{j-1} + \int_{I_j} g \ , \qquad j = 1, \cdots, M \ , \tag{3.3.11}
$$

because the solution to (3.1.24) has nodal values $\{U_j\}$, which can found using

$$\mu \frac{U_j - U_{j-1}}{h_j} + \sigma_{j-1/2} \frac{U_j + U_{j-1}}{2} \;=\; \frac{1}{h_j} \int_{I_j} g \;,$$

for all $j = 1, \cdots, M$, with analogous boundary conditions.

Then, using the notation $p_j = 1 - \alpha_j$, $q_j = 1 + \alpha_j$ and $r_j = p_j/q_j$, (3.3.11) becomes

$$U_j \;=\; r_j U_{j-1} + \frac{1}{\mu q_j} \int_{I_j} g \;.$$

We now iterate this formula to prove existence and uniqueness of the solution. Suppose there exist $r, q \geq 0$ such that

$$\left.\begin{array}{rcl} |r_j| \leq & r & < \; 1 \;, \\ q_j \geq & q & \geq \; 1 \;, \end{array}\right\} \quad \text{for all } \; 1 \leq j \leq M \;. \tag{3.3.12}$$

Then,

$$
\begin{aligned}
|U_j| \;\leq\;& r|U_{j-1}| + \frac{h}{\mu q}\|g\|_\infty \\
\leq\;& r^2|U_{j-2}| + (r+1)\frac{h}{\mu q}\|g\|_\infty \\
\leq\;& \left(\sum_{i=1}^{j} r^{j-i}\right) \frac{h}{\mu q}\|g\|_\infty \\
\leq\;& \frac{h}{\mu q(1-r)}\|g\|_\infty \;,
\end{aligned} \tag{3.3.13}
$$

where we have used the boundary condition $U_0 = 0$ (for $\mu > 0$). Since $\mathcal{S}_\mu^h g \in V^h$, we have

$$\|\mathcal{S}_\mu^h g\|_\infty \leq \max_j |U_j| \;, \tag{3.3.14}$$

and therefore in order to find an explicit bound of $\|\mathcal{S}_\mu^h\|_{L_\infty \mapsto V^h}$ we must first find expressions for the bounds r and q in (3.3.12). The second inequality is fairly simple

$$q_j \;\geq\; q \;:=\; \max\left\{1, \; \sigma_{\min}\frac{h}{2\rho\mu}\right\} \;\geq\; 1 \;. \tag{3.3.15}$$

For the first inequality in (3.3.12), we consider three separate cases: $h \leq 2\mu/\sigma_{\max}$ (or equivalently $\sigma_{\max}h/2\mu \leq 1$); $h \geq 2\rho\mu/\sigma_{\min}$; and $2\mu/\sigma_{\max} \leq h \leq 2\rho\mu/\sigma_{\min}$. Considering each of these cases, it is possible to to find an r that satisfies the condition in (3.3.12) and one example is defined as follows:

$$r \;=\; \begin{cases} \left(1 + \frac{\sigma_{\min}h}{2\rho\mu}\right)^{-1} & \text{if } h \leq 2\mu/\sigma_{\max} \;, \\[2mm] \left(1 + \frac{2\mu}{\sigma_{\max}h}\right)^{-1} & \text{if } h \geq 2\rho\mu/\sigma_{\min} \;, \\[2mm] \left(1 - \frac{\sigma_{\min}}{\sigma_{\max}\rho}\right) \Big/ \left(1 + \frac{\sigma_{\min}}{\sigma_{\max}\rho}\right) & \text{otherwise} \;. \end{cases} \tag{3.3.16}$$

To see this for the first two cases is fairly simple, by bounding α_j and manipulating. For the third case, note that $h/2\mu \geq 1/\sigma_{\max}$ and hence $1 + \sigma_{\max}h/2\mu \geq 1 + 1 > 1 + (\sigma_{\min}/\sigma_{\max}\rho)$, and then the bound follows.

To finish the proof we show that

$$\frac{h}{\mu q(1-r)} \;\leq\; \frac{2\rho}{\sigma_{\min}}\left(1 + \sigma_{\max}\frac{h}{\mu}\right) \;. \tag{3.3.17}$$

Considering, for example, the case when $2\mu/\sigma_{\max} \le h \le 2\rho\mu/\sigma_{\min}$, then r is given by the third equation in (3.3.16) and

$$\frac{1}{q(1-r)} \le \frac{1}{2}\left(1 + \rho\frac{\sigma_{\max}}{\sigma_{\min}}\right) = \frac{\rho}{\sigma_{\min}}\left(\frac{\sigma_{\min}}{2\rho} + \frac{\sigma_{\max}}{2}\right) < \frac{\rho}{\sigma_{\min}}\left(\frac{\mu}{h} + \frac{\sigma_{\max}}{2}\right) .$$

So

$$\frac{h}{\mu q(1-r)} \le \frac{\rho}{\sigma_{\min}}\left(1 + \frac{\sigma_{\max}}{2}\frac{h}{\mu}\right) ,$$

which yields (3.3.17). The proof of (3.3.17) for the other values of r is similar. ∎

Lemma 3.3.6 *Let $u = \mathcal{S}_\mu g$ and $u^h = \mathcal{S}_\mu^h g$, for some $g \in L_\infty$. Recall that $\sigma \in \mathrm{C}_{pw}^\eta$, for some $\eta \in (0,1)$ (Assumption 3.1.2). Then, the piecewise linear interpolant $\widehat{u}$ of u onto the mesh (3.1.6) satisfies*

$$\|u^h - \widehat{u}\|_\infty \le \|\mathcal{S}_\mu^h\|_{L_\infty \mapsto V^h}\left(\sigma_{\max}\|u - \widehat{u}\|_\infty + h^\eta\|\sigma\|_{\eta,pw}\|u\|_\infty\right) .$$

Proof. Using (3.1.24) and (3.1.17),

$$\int_{I_j} \mu\frac{d}{dx}u^h + \mathcal{P}^h\sigma\, u^h\, \mathrm{d}x = \int_{I_j} \mu\frac{d}{dx}u + \sigma u\, \mathrm{d}x ,$$

and hence subtracting $\int_{I_j} \mu\frac{d}{dx}\widehat{u} + (\mathcal{P}^h\sigma)\widehat{u}$ from both sides gives

$$\int_{I_j} \mu\frac{d}{dx}(u^h - \widehat{u}) + \mathcal{P}^h\sigma\, (u^h - \widehat{u}) = \int_{I_j} \mu\frac{d}{dx}(u - \widehat{u}) + \sigma u - \mathcal{P}^h\sigma\, \widehat{u}$$

$$= \int_{I_j} \left(\mathcal{P}^h\sigma\, (u - \widehat{u}) + (\sigma - \mathcal{P}^h\sigma)u\right) , \tag{3.3.18}$$

where the derivative term in (3.3.18) vanishes because u and $\widehat{u}$ coincide at the mesh nodes. Hence, by the definition of $\mathcal{S}_\mu^h$ in (3.1.24),

$$u^h - \widehat{u} = \mathcal{S}_\mu^h\left[\mathcal{P}^h\sigma\, (u - \widehat{u}) + (\sigma - \mathcal{P}^h\sigma)u\right] ,$$

and therefore

$$\|u^h - \widehat{u}\|_\infty \le \|\mathcal{S}_\mu^h\|_{L_\infty \mapsto V^h}\left[\|\mathcal{P}^h\sigma\|_\infty\|u - \widehat{u}\|_\infty + \|\sigma - \mathcal{P}^h\sigma\|_\infty\|u\|_\infty\right] ,$$

from which the result follows by (3.3.9) since $\sigma \in \mathrm{C}_{pw}^\eta$, for $\eta \in (0,1)$. ∎

As mentioned in the introduction to this chapter, the main deterministic error estimate (Theorem 3.3.11) will contain an $h\log N$ term. The next result is the first indication of this, showing that $\|\mathcal{S}_\mu - \mathcal{S}_\mu^h\|_{L_\infty \mapsto \mathrm{C}}$ can blow up as $|\mu| \to 0$ for fixed h.

Lemma 3.3.7 *There is a constant $c > 0$, independent of all parameters, such that*

$$\|\mathcal{S}_\mu - \mathcal{S}_\mu^h\|_{L_\infty \mapsto \mathrm{C}} \le c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2\frac{h}{|\mu|} + \|\sigma\|_{\eta,pw}h^\eta\right) . \tag{3.3.19}$$

Proof. Recall that, for any $g \in L_\infty$, $u = \mathcal{S}_\mu g \in \mathrm{C}$ is the solution of (3.1.17) and $u^h = \mathcal{S}_\mu^h g \in V^h$ is its continuous piecewise linear approximation, defined by (3.1.24). Then, with $\widehat{u}$ as in Lemma 3.3.6,

$$\|\left(\mathcal{S}_\mu - \mathcal{S}_\mu^h\right)g\|_\infty = \|u - u^h\|_\infty \le \|u - \widehat{u}\|_\infty + \|\widehat{u} - u^h\|_\infty . \tag{3.3.20}$$

Hence, by Lemmas 3.3.5 and 3.3.6 and noting that $\sigma_{\max}\|\mathcal{S}_\mu^h\|_{L_\infty \mapsto V^h} \geq 1$, we have

$$\|u - u^h\|_\infty \; \leq \; 2\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right)\left(\frac{3}{2}\sigma_{\max}\|u - \widehat{u}\|_\infty + \|\sigma\|_{\eta,pw}h^\eta\|u\|_\infty\right) . \tag{3.3.21}$$

Now since (by Lemma 3.2.3) we have $u' \in L_\infty$, it follows that $\|u - \widehat{u}\|_\infty \leq ch\|u'\|_\infty$. Combining the result (3.3.21) with Lemmas 3.2.2 and 3.2.3, then

$$\begin{aligned}
\|\left(\mathcal{S}_\mu - \mathcal{S}_\mu^h\right)g\|_\infty \; &\leq \; c\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right)(\sigma_{\max}h\|u'\|_\infty + \|\sigma\|_{\eta,pw}h^\eta\|u\|_\infty) \\
&\leq \; c\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right)\left(\frac{\sigma_{\max}^2}{\sigma_{\min}}\frac{h}{|\mu|} + \frac{\|\sigma\|_{\eta,pw}}{\sigma_{\min}}h^\eta\right)\|g\|_\infty .
\end{aligned}$$

Assuming that $\sigma_{\max}h/|\mu| \leq 1$, then we get the desired result. Otherwise $\sigma_{\max}h/|\mu| > 1$ and then by the triangle inequality and Lemmas 3.2.2 and 3.3.5 we have

$$\begin{aligned}
\|\left(\mathcal{S}_\mu - \mathcal{S}_\mu^h\right)g\|_\infty \; &\leq \; \|\mathcal{S}_\mu g\|_\infty + \|\mathcal{S}_\mu^h g\|_\infty \\
&\leq \; \sigma_{\min}^{-1}\|g\|_\infty + 2\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right)\|g\|_\infty \\
&\leq \; c\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right)\|g\|_\infty \\
&\leq \; c\rho\sigma_{\min}^{-1}\frac{\sigma_{\max}}{\sigma_{\min}}\sigma_{\max}\frac{h}{|\mu|}\|g\|_\infty \\
&\leq \; c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2\frac{h}{|\mu|} + \|\sigma\|_{\eta,pw}h^\eta\right)\|g\|_\infty ,
\end{aligned}$$

where we used $\rho \geq 1$, $\sigma_{\max}/\sigma_{\min} \geq 1$, $\sigma_{\max}h/|\mu| > 1$ and $\|\sigma\|_{\eta,pw}h^\eta \geq 0$. $\blacksquare$

The next result is analogous to Theorem 3.3.3.

Lemma 3.3.8 *Let $\mathcal{K}^N$ and $\mathcal{K}^{h,N}$ be defined by (3.3.1) and (3.1.25) respectively, with $\{\mu_k\}$ and $\{w_k\}$ given by the double Gauss rule. Then, for $N \geq 2$,*

$$\|\mathcal{K}^N - \mathcal{K}^{h,N}\|_{L_\infty \mapsto C} \; \leq \; c\rho\sigma_{\min}^{-2}\left[\sigma_{\max}^2 h \log N + \|\sigma\|_{\eta,pw}h^\eta\right] .$$

Proof. Using (3.3.5) and theorem 3.3.7 for any $g \in L_\infty$, then we have,

$$\begin{aligned}
\|\left(\mathcal{K}^N - \mathcal{K}^{h,N}\right)g\|_\infty \; &\leq \; \sum_{|k|=1}^N w_k\|\mathcal{S}_{\mu_k} - \mathcal{S}_{\mu_k}^h\|_{L_\infty \mapsto L_\infty}\|g\|_\infty \\
&\leq \; c\rho\sigma_{\min}^{-2}\|g\|_\infty\sum_{|k|=1}^N w_k\left(\sigma_{\max}^2\frac{h}{|\mu_k|} + \|\sigma\|_{\eta,pw}h^\eta\right) . \tag{3.3.22}
\end{aligned}$$

Now, since the Gauss rule is exact for constant functions, we have

$$\sum_{|k|=1}^N w_k \; = \; 2 . \tag{3.3.23}$$

Moreover, from [163, Lemma 3.1] we know that for any quadrature rule satisfying the assumptions of Section 3.1.1 and the additional assumption:

$$\sum_{k=1}^n w_k \; \leq \; c\mu_n , \tag{3.3.24}$$

where the (first N) angles are ordered such that $0 < \mu_1 < \cdots < \mu_N \leq 1$ (and recall that $-\mu_k = \mu_{-k}$) - which holds true for the double Gauss rule - then

$$\sum_{|k|=1}^{N} w_k |\mu_k|^{-1} \;\leq\; c\,(1 + |\log \mu_1|) \;. \tag{3.3.25}$$

For the case of a double Gauss rule, points near the origin are spaced $\mathcal{O}(N^{-2})$ and hence $|\log \mu_1| \sim 2 \log N \geq 1$, for $N \geq 2$.

Using (3.3.22), (3.3.23) and (3.3.25) gives the desired bound for $(\mathcal{K}^N - \mathcal{K}^{h,N}) : L_\infty \mapsto L_\infty$. Similarly to Theorem 3.3.3, the extension to C then follows because $\mathcal{S}_\mu$, $\mathcal{S}_\mu^h : L_\infty \mapsto$ C, by Lemma 3.2.3 and Lemma 3.3.5, respectively. ∎

Theorem 3.3.9 *Suppose the assumptions of Lemma 3.3.8 hold and assume that $h \in (0,1)$. Then $e^{h,N} \in$ C with the bound*

$$\|e^{h,N}\|_\infty \;\leq\; c\rho\sigma_{\min}^{-2}\overline{\|\sigma_S\|_{\eta,pw}}\mathcal{R}_2(\sigma,\sigma_S)\left(\sigma_{\max}^2 h \log N + \|\sigma\|_{\eta,pw} h^\eta\right)\|f\|_{\eta,pw} \;,$$

where $e^{h,N}$ is defined in (3.3.3) and $\mathcal{R}_2(\sigma,\sigma_S)$ is defined in Corollary 3.2.9.

Proof. Using (3.3.9) and Lemma 3.3.8 we have

$$\begin{aligned}
\|e^{h,N}\|_\infty &\leq \Big(\|\left(\mathcal{K}^N - \mathcal{K}^{h,N}\right)\|_{C_{pw}^\eta \mapsto C} + \|\mathcal{K}^{h,N}\|_{C_{pw}\mapsto C}\|\left(I - \mathcal{P}^h\right)\|_{C_{pw}^\eta \mapsto C_{pw}}\Big)\|\sigma_S\phi + f\|_{\eta,pw} \\
&\leq c\rho\sigma_{\min}^{-2}\left[\left(\sigma_{\max}^2 h \log N + \|\sigma\|_{\eta,pw} h^\eta\right) + \sigma_{\min}\left(1 + \sigma_{\max} h \log N\right) h^\eta\right]\|\sigma_S\phi + f\|_{\eta,pw} \\
&\leq c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2 h \log N(1 + h^\eta) + \|\sigma\|_{\eta,pw} h^\eta\right)\|\sigma_S\phi + f\|_{\eta,pw} \\
&\leq c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2 h \log N + \|\sigma\|_{\eta,pw} h^\eta\right)\|\sigma_S\phi + f\|_{\eta,pw} \;,
\end{aligned}$$

where $\sigma_S\phi + f \in C_{pw}^\eta$, and we have used that

$$\|\mathcal{K}^{h,N}\|_{L_\infty \mapsto C} \;\leq\; c\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max} h \log N\right) \;, \tag{3.3.26}$$

which can be shown by a similar sequence of steps to the proof of Lemma 3.3.8. The result then follows by writing

$$\begin{aligned}
\|\sigma_S\phi + f\|_{\eta,pw} &\leq \|\sigma_S\|_{\eta,pw}\|\phi\|_{\eta,pw} + \|f\|_{\eta,pw} \;\leq\; \left(\overline{\|\sigma_S\|_{\eta,pw}}\mathcal{R}_2(\sigma,\sigma_S) + 1\right)\|f\|_{\eta,pw} \\
&\leq \overline{\|\sigma_S\|_{\eta,pw}}\mathcal{R}_2(\sigma,\sigma_S)\|f\|_{\eta,pw} \;,
\end{aligned}$$

where we use Corollary 3.2.9 and that $\mathcal{R}_2(\sigma,\sigma_S) \geq 1$.

∎

3.3.3 Stability

Up to now we have shown $\|e^N\|_\infty$ and $\|e^{h,N}\|_\infty$ approach zero as $h \log N \to 0$ and $N \to \infty$, see Corollary 3.3.4 and Theorem 3.3.9 respectively. To prove a final bound on (3.3.7) we need to show stability, i.e. show $\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}$ exists and can be bounded in the $\|\cdot\|_{C\mapsto C}$ norm, independently of h and N. To do this it is useful to use the identity

$$\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1} \;=\; I + \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S \;. \tag{3.3.27}$$

A proof of the identity is given in Lemma E.2.1. We have already bounded $\|\mathcal{P}^h \sigma_S\|_{C \mapsto C_{pw}}$ in (3.3.9) and $\|\mathcal{K}^{h,N}\|_{C_{pw} \mapsto C}$ in (3.3.26), (although the bounds are not independent of h or N). Therefore,

$$\left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S\right)^{-1} \text{ is bounded on C} \iff \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)^{-1} \text{ is bounded on C}_{pw} .$$

The identity (3.3.27) allows us to prove simpler results on $C_{pw} \supset C$, which allow us to bound $\| \left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S\right)^{-1} \|_{C \mapsto C}$.

Theorem 3.3.10 *If h and N^{-1} are sufficiently small, such that $h \log N \leq 1$ and*

$$\frac{1}{h^\eta + h \log N + N^{-1}} \geq c \left(\frac{\|\sigma_S\|_{\eta,pw}}{(\sigma_S)_{\min}}\right)^2 \sigma_{\min}^{-1} \max\{ \overline{\sigma_{\max}}^2, \|\sigma\|_{\eta,pw}\} \mathcal{R}_1(\sigma, \sigma_S) =: \mathcal{R}_3(\sigma, \sigma_S) , \tag{3.3.28}$$

then $\left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S\right)^{-1}$ maps on C with the bound

$$\| \left(I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S\right)^{-1} \|_{C \mapsto C} \leq c\rho \left(\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\right)^2 \overline{\sigma_{\max}} \mathcal{R}_1(\sigma, \sigma_S) =: \mathcal{R}_4(\sigma, \sigma_S) , \tag{3.3.29}$$

where $\mathcal{R}_1(\sigma, \sigma_S)$ is defined in Lemma 3.2.8.

Proof. Fix $0 < \epsilon < 1$ and set $\mathcal{A}(h, N) = I - (I - \sigma_S \mathcal{K})^{-1} \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)$. Suppose that there exists a h and N^{-1}, sufficiently small, such that

$$\|\mathcal{A}(h, N)\|_{C_{pw} \mapsto C_{pw}} \leq \epsilon . \tag{3.3.30}$$

Then it would follow from the Banach Lemma that $\| \left(I - \mathcal{A}(h, N)\right)^{-1} \|_{C_{pw} \mapsto C_{pw}} \leq 1/(1 - \epsilon)$ and so

$$\| \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)^{-1} (I - \sigma_S \mathcal{K}) \|_{C_{pw} \mapsto C_{pw}} = \| \left(I - \mathcal{A}(h, N)\right)^{-1} \|_{C_{pw} \mapsto C_{pw}} \leq \frac{1}{1 - \epsilon} . \tag{3.3.31}$$

This would then imply,

$$\| \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)^{-1} \|_{C_{pw} \mapsto C_{pw}} = \| \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)^{-1} (I - \sigma_S \mathcal{K}) (I - \sigma_S \mathcal{K})^{-1} \|_{C_{pw} \mapsto C_{pw}}$$

$$\leq \frac{1}{1 - \epsilon} \| (I - \sigma_S \mathcal{K})^{-1} \|_{C_{pw} \mapsto C_{pw}} \leq \frac{1}{1 - \epsilon} \frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}} \mathcal{R}_1(\sigma, \sigma_S) ,$$

where we used the identity $(I - \sigma_S \mathcal{K})^{-1} = \sigma_S (I - \mathcal{K} \sigma_S)^{-1} \sigma_S^{-1}$ and Lemma 3.2.8. Thus the desired result follows, *on the assumption* that (3.3.30) holds.

To verify (3.3.30), we write

$$\|\mathcal{A}(h, N)\|_{C_{pw} \mapsto C_{pw}} = \| (I - \sigma_S \mathcal{K})^{-1} \left[(I - \sigma_S \mathcal{K}) - \left(I - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\right)\right] \|_{C_{pw} \mapsto C_{pw}}$$

$$\leq \| (I - \sigma_S \mathcal{K})^{-1} \|_{C_{pw} \mapsto C_{pw}} \|\mathcal{P}^h \sigma_S \mathcal{K}^{h,N} - \sigma_S \mathcal{K}\|_{C_{pw} \mapsto C_{pw}}$$

$$\leq \frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}} \| (I - \mathcal{K} \sigma_S)^{-1} \|_{C_{pw} \mapsto C_{pw}} \|\sigma_S \mathcal{K} - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\|_{C_{pw} \mapsto C_{pw}}$$

$$\leq \frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}} \mathcal{R}_1(\sigma, \sigma_S) \|\sigma_S \mathcal{K} - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\|_{C_{pw} \mapsto C_{pw}} , \tag{3.3.32}$$

where we again used Lemma 3.2.8. To estimate the right hand side of (3.3.32) we write

$$\|\sigma_S \mathcal{K} - \mathcal{P}^h \sigma_S \mathcal{K}^{h,N}\|_{C_{pw} \mapsto C_{pw}} \leq \| \left(I - \mathcal{P}^h\right) \sigma_S \mathcal{K}\|_{C_{pw} \mapsto C_{pw}} \tag{3.3.33}$$

$$+ \|\mathcal{P}^h \sigma_S\|_{C \mapsto C_{pw}} \left(\|\mathcal{K} - \mathcal{K}^N\|_{C_{pw} \mapsto C} + \|\mathcal{K}^N - \mathcal{K}^{h,N}\|_{C_{pw} \mapsto C}\right) .$$

Then we can bound the second term on the right hand side of (3.3.33) by using the trivial bound on $\|\mathcal{P}^h \sigma_S\|_\infty$, Theorem 3.3.3 and Theorem 3.3.8. Moreover, by defining

$$\xi \; := \; \begin{cases} \eta \,, & \text{if } \eta < 1 \\ 1 - \delta \,, & \text{otherwise} \end{cases} \,,$$

for $0 < \delta \ll 1$, and using (3.3.9) and Theorem 3.2.7(ii), then we can show that for the first term on the right hand side of (3.3.33)

$$\|(I - \mathcal{P}^h)\sigma_S \mathcal{K}\|_{C_{pw} \mapsto C_{pw}} \;\leq\; \|\sigma_S\|_{\xi,pw}\|\mathcal{K}\|_{C_{pw} \mapsto C^\xi} h^\xi \;\leq\; c\left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\right)\|\sigma_S\|_{\eta,pw}\left(h^\eta + h^{1-\delta}\right) \,,$$

where we have used that $\|\cdot\|_{\xi,pw} \leq \|\cdot\|_{\eta,pw}$, for all $\xi \leq \eta$.

Bringing the results together gives

$$\|\sigma_S\mathcal{K} - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\|_{C_{pw} \mapsto C_{pw}} \tag{3.3.34}$$
$$\leq\; c\rho \frac{\|\sigma_S\|_{\eta,pw}}{\sigma_{\min}} \max\{\ \overline{\sigma_{\max}},\ \sigma_{\min}^{-1}\sigma_{\max}^2,\ \sigma_{\min}^{-1}\|\sigma\|_{\eta,pw}\}\left(h^\eta + h\log N + N^{-1}\right)$$
$$\leq\; c\rho \frac{\|\sigma_S\|_{\eta,pw}}{\sigma_{\min}^2} \max\{\ \overline{\sigma_{\max}}^2,\ \|\sigma\|_{\eta,pw}\}\left(h^\eta + h\log N + N^{-1}\right) \,.$$

Using (3.3.34) and (3.3.32) allows us to state the condition (3.3.28) - which ensures that $\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}$ maps on C_{pw} and

$$\left\|\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\right\|_{C_{pw} \mapsto C_{pw}} \;\leq\; c\,\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\mathcal{R}_1(\sigma,\sigma_S) \,.$$

It is then simple to prove (3.3.29) by (3.3.27). ■

3.3.4 The Error Estimate

We now have all the ingredients to prove the following error estimate, and one of the main results of this thesis.

Theorem 3.3.11 *Let $\phi^{h,N}$ denote the approximation to ϕ, using a double Gauss quadrature rule and a Crank-Nicolson scheme. Then, for h and N^{-1} sufficiently small according to (3.3.28) and $h\log N \leq 1$,*
$$\|\phi - \phi^{h,N}\|_\infty \;\leq\; \mathcal{R}(\sigma,\sigma_S)\left(N^{-1} + h\log N + h^\eta\right)\|f\|_{\eta,pw} \,,$$
where $\mathcal{R}(\sigma,\sigma_S) = c\rho\mathcal{R}_4(\sigma,\sigma_S)\mathcal{R}_2(\sigma,\sigma_S)\sigma_{\min}^{-2}\overline{\|\sigma\|_{\eta,pw}}^2\overline{\|\sigma_S\|_{\eta,pw}}$.

Proof. The proof follows by considering (3.3.7) and applying the bounds of the stability estimate (in (3.3.29)), and the angular (Corollary 3.3.4) and spatial (Theorem 3.3.9) consistency conditions, i.e.

$$\|\phi - \phi^{h,N}\|_\infty$$
$$\leq\; c\rho\mathcal{R}_4(\sigma,\sigma_S)\left[\left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\right)^2\mathcal{R}_1(\sigma,\sigma_S)N^{-1} + \frac{\overline{\|\sigma_S\|_{\eta,pw}}}{\sigma_{\min}^2}\mathcal{R}_2(\sigma,\sigma_S)\left(\sigma_{\max}^2 h\log N + \|\sigma\|_{\eta,pw}h^\eta\right)\right]\|f\|_{\eta,pw}$$
$$\leq\; c\rho\mathcal{R}_4(\sigma,\sigma_S)\mathcal{R}_2(\sigma,\sigma_S)\left(\frac{\overline{\|\sigma\|_{\eta,pw}}}{\sigma_{\min}}\right)^2\overline{\|\sigma_S\|_{\eta,pw}}\left(N^{-1} + h\log N + h^\eta\right)\|f\|_{\eta,pw} \,,$$
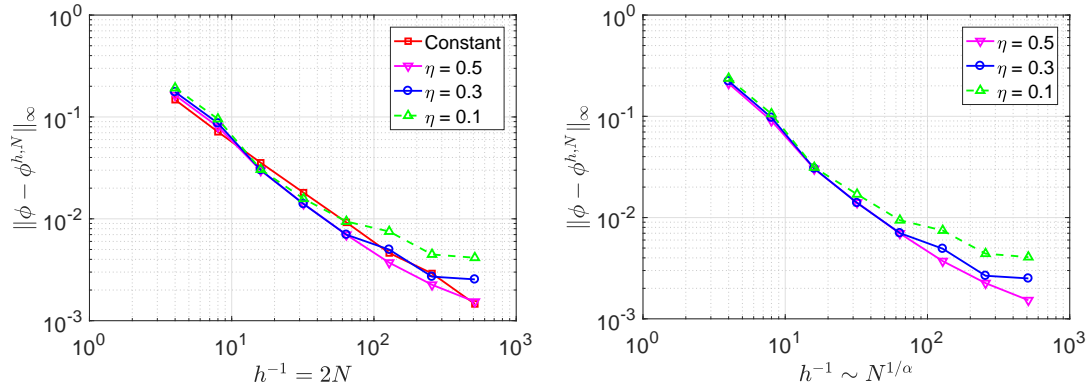
Figure 3-1: The error of the approximation to the scalar flux. (Left) $N = (2h)^{-1}$; (Right) $N = 2\lceil \alpha^{-1} h^{-\alpha} \rceil$, where $\alpha := \min\{1, \eta\}$.

where we used $\mathcal{R}_1(\sigma, \sigma_S) \leq \mathcal{R}_2(\sigma, \sigma_S)$. ∎

3.4 Numerical Results

We now present numerical results to support the error estimate in Theorem 3.3.11.

For the input data, we assume the absorption cross-section and the fixed source term are constant, i.e. $\sigma_A \equiv \exp(0.25)$ and $f \equiv \exp(1)$. Moreover, we consider four deterministic fields for σ_S. The first is a constant coefficient case, with $\sigma_S \equiv \exp(0.75)$. The remaining fields are (pointwise) η-Hölder continuous fields [183] (with parameters $\eta = 0.1, 0.3, 0.5$) defined by

$$\sigma_S(x) \;=\; |x - 0.4999|^\eta \; \sin\left(\frac{1}{|x - 0.4999|^2}\right) \; . \tag{3.4.1}$$

The total cross-section is then defined by $\sigma = \sigma_A + \sigma_S$.

For the discretisation, we choose a uniform spatial mesh with mesh width h and the double Gauss quadrature rule with $2N$ points. We consider two spatial-angular relationships; $N = (2h)^{-1}$ and $N = 2\lceil \frac{1}{\min\{1,\eta\}} h^{-\min\{1,\eta\}} \rceil$. The approximation to the scalar flux is computed at different resolutions, with $h^{-1} = 4, 8, 16, ..., 512$, and compared to a reference solution calculated when $h^{-1} = 1024$ and $N = 512$. The error is estimated in the supremum norm.

The numerical results are presented in Figure 3-1. For $h^{-1} \leq 64$ and for all choices of σ_S and N, we observe $\mathcal{O}(h)$ convergence. This is in line with the constant coefficient result in [163]. For larger $h^{-1} > 64$ the convergence begins to slow, for the low regularity fields. We conjecture that the convergence is $\mathcal{O}(h^{\min\{1,\eta\}})$, i.e. essentially the $\mathcal{O}(h^\eta)$ convergence (when $\eta < 1$, or the $\mathcal{O}(h \log N)$ convergence when $\eta \geq 1$) in Theorem 3.3.11 - implying the error estimate is sharp.

78

Chapter 4

Multilevel Monte Carlo Theory for Heterogeneous Transport

Contents

In this chapter, we introduce *uncertainty into the input data* of the problem considered in Chapter 3, i.e. the radiative transport equation in one spatial and one angular dimension discretised using the classical discrete ordinates method and a classical diamond differencing (or Crank-Nicolson) scheme. We will extend the error estimate of Chapter 3 to a probabilistic error estimate (see ahead to (4.0.1)). The probabilistic error estimate then facilitates a rigorous proof of the parameters α, β and γ, that arise in the computational ϵ-cost of a (multilevel) Monte Carlo estimator, i.e. (2.4.13) and (2.4.34) (see ahead to (4.0.2) and (4.0.3)). Numerical results are presented, where we use standard Monte Carlo, quasi-Monte Carlo and their multilevel variants as estimators. As far as we know, these methods have not been applied to radiative transport until now. In the numerical section, the cross-sections are assumed to be log-normal random fields equipped with the Matérn class of covariances, and represented by a Karhunen-Loève expansion (recall Section 2.3.2 and Section 2.3.1) - we note that the theoretical results are more general than this case.

We now give an overview of the results of this chapter, without the detailed analysis. For this introduction (to ensure that the spatial and angular errors are equal order) we set $N = N(h) = \max\{\lceil ch^{-\eta} \rceil, 4\}$, for some constant $c > 0$ independent of h.

The first result in this Chapter is a probabilistic counterpart of (3.0.2). Here we have to deal with the fact that the deterministic estimate (3.0.2) is subject to the "mesh resolution condition" or "stability condition" (3.0.1), which in turn arises from the non-self-adjointness of the RTE. In the case of coercive self-adjoint PDEs with random data and Galerkin discretisation (e.g. [48, 199]) one obtains a probabilistic error estimate by interpreting the deterministic error estimate pathwise and then taking expectation. This does not work here because of the pathwise stability estimate (3.0.1). To get around this problem, given a path independent mesh width $h < 1$, for each realisation $\sigma = \sigma(\cdot, \omega), \sigma_S = \sigma_S(\cdot, \omega)$, we let $\overline{h}_\omega$ denote (the largest) mesh diameter which satisfies

the path-dependent criterion (3.0.1) and finally set $h_\omega = \min\{h, \overline{h}_\omega\}$. Then the approximation to $\phi = \phi(\cdot, \omega)$ is taken to be $\Phi^h = \phi^{h_\omega, N(h_\omega)}$. We prove in Theorem 4.1.3 that

$$\|\phi - \Phi^h\|_{L_p(\Omega; L_\infty)} \leq c(p, r) \, h^\eta \, \|f\|_{L_r(\Omega; C_{pw}^\eta)} \, . \tag{4.0.1}$$

for any $1 \leq p \leq r$, provided the norm on the right-hand side is finite and the cross sections σ, σ_S have bounded moments of any finite order. Here, $c(p, r)$ denotes an absolute constant depending only on (p, r), and the norms are the usual Bochner norms with respect to the probability space Ω (defined in Section 2.1). This result shows that the error in the Bochner norm on the left-hand side decreases with deterministic rate h^η, provided we are willing to use a finer mesh for any particular sample where the resolution criterion (3.0.1) demands it. If we assume furthermore that the cost $\mathcal{C}(\cdot)$ to compute a single sample of $\Phi^h = \phi^{h_\omega, N(h_\omega)}$ (e.g. measured in floating point operations) satisfies

$$\mathcal{C}(\phi^{h_\omega, N(h_\omega)}) \leq c'(\omega) h_\omega^{-\gamma} \, ,$$

for some $\gamma > 0$ and that the sample-dependent constant $c' \in L_p(\Omega)$, for some $p > 1$, then the third main result of this paper in Theorem 4.1.5 is that

$$\mathbb{E}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-\gamma}\right) \, , \tag{4.0.2}$$

where the hidden constant is independent of h. The important observation is that, on average, the cost to compute a sample from Φ^h has the same cost growth rate (with respect to h) as the sample-wise cost (with respect to h_ω), despite some samples $\Phi^h(\omega, x)$ being computed on a mesh with $h_\omega \ll h$ in order to satisfy the stability criterion.

Estimates, such as (4.0.1) and (4.0.2), play a crucial role in the complexity analysis of (multilevel) Monte Carlo methods for computing the expectation of (functionals of) the solution ϕ of a randomised version of (3.1.1). Suppose $\mathfrak{L}(\phi)$ is such a functional - the quantity of interest - and to simplify notation we write this as $Q = \mathfrak{L}(\phi)$ (a random variable). We approximate Q by $Q_h := Q(\Phi^h)$ with Φ^h described above and then approximate $\mathbb{E}[Q]$, by applying a sampling method of choice to $\mathbb{E}[Q_h]$ – we denote the result as $\widehat{Q}_h$.

As we have already mentioned, finding an accurate and efficient estimator $\widehat{Q}_h$ of $\mathbb{E}[Q]$ is at the heart of the forward problem of Uncertainty Quantification (UQ), and to compare such methods, we consider the *computational ϵ-cost* $\mathcal{C}_\epsilon(\widehat{Q}_h)$ of an estimator $\widehat{Q}_h$. That is, if ϵ denotes a desired accuracy (in the sense of root mean-squared error, see (2.4.2)), then $\mathcal{C}_\epsilon(\widehat{Q}_h)$ is defined to be the total cost for $\widehat{Q}_h$ to achieve an accuracy of ϵ.

By the general theory in [48], i.e. Theorem 2.4.4, the ϵ-cost of standard and multilevel Monte Carlo methods can be computed in terms of the parameter η in (3.0.2) (related to the regularity of the data), the parameter γ in (4.0.2) (related to the cost per sample), as well as another parameter β that quantifies the speed of variance reduction (recall (2.4.33)) between levels of the multilevel scheme and can also be derived from (4.0.1). In the fourth main result of this paper in Theorem 4.1.8, we prove rigorously that

$$\beta \geq 2\eta \, . \tag{4.0.3}$$

To provide a bound on γ that only depends on the regularity of the data it is necessary to fix the solution method. Two particular examples that were used in our numerical results are given in Example 4.1.6 and have already been discussed in Section 3.1.2. In particular, for the asymptotically cheaper one of the two methods, the so-called source iteration, we have $\gamma \leq 1 + \eta$ (if

certain assumptions on the relationship between h and $N = N(h)$ are made). The general theory in [48] then leads to the following respective upper bounds on the ϵ-costs of the standard and the multilevel Monte Carlo estimators $\widehat{Q}_h^{MC}$ and $\widehat{Q}_h^{MLMC}$:

$$\mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\Big] = \mathcal{O}\big(\epsilon^{-(4+\frac{1-\eta}{\eta})}\big) \quad \text{and} \quad \mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\Big] = \mathcal{O}\big(\epsilon^{-(2+\frac{1-\eta}{\eta})}\big),$$

i.e. a theoretical gain of up to two orders of magnitude in ϵ-cost. However, we will see in the numerical section that this is overly optimistic, since the bound on the ϵ-cost of the standard Monte Carlo estimator is not sharp. Nevertheless, we do observe gains of (at least) one order of magnitude.

The structure of the remaining chapter is as follows. In Section 4.1.1, we assume certain statistical properties on our nuclear input data, which allows us to prove that the explicit coefficients we found in the results of Chapter 3 (e.g. the $\mathcal{R}(\sigma, \sigma_S)$ in (3.0.2)) belong to certain probabilistic Lebesgue space(s). Using this result allows a simple proof of the probabilistic counterpart to (3.0.2) to be made - with one caveat - we have to deal with the fact that the deterministic estimate is subject to the stability condition (3.0.1) and hence we define $\overline{h}_\omega$, the largest mesh diameter which satisfies the stability condition. Then, we prove (4.0.2) under certain assumptions, which are satisfied for given examples. In Section 4.1.2, we use the results of Section 4.1.1 to rigorously estimate the α and β parameters from (2.4.3) and (2.4.33), respectively (i.e. (4.0.3)) and then make rigorous statements about the standard and multilevel Monte Carlo methods. Finally, in Section 4.2 we present a hybrid direct-iterative solver which allows us to more efficiently estimate the scalar flux on the hierarchy of discretisations that multilevel Monte Carlo introduces. Numerical results are presented in support of the chapter.

4.1 Application in Uncertainty Quantification

In Theorem 3.3.11, in the previous chapter, we proved a deterministic error estimate that is *explicit in its dependence on the input data*, through the appearance of the terms $\mathcal{R}(\sigma, \sigma_S)$ (for the cross-sections) and $\|f\|_{\eta,pw}$ (for the forcing). The explicit dependence will be very important as we turn our attention to the original objective - an error estimate when the input data is random.

4.1.1 Random Input Data and Probabilistic Error Estimates

To formally describe the random model, recall that we use $\omega \in \Omega$ to denote a random event from the sample space Ω and $\mathbb{P} : \Omega \mapsto [0,1]$ to denote an associated probability measure. Moreover, we defined the Bochner space $L_p(\Omega; X) := \{g : \Omega \mapsto X \mid \int_\Omega \|g\|_X^p \mathrm{d}\mathbb{P}(\omega) < \infty\}$ with associated norm $\|g\|_{L_p(\Omega;X)} = \big(\mathbb{E}\left[\|g\|_X^p\right]\big)^{1/p}$, for some normed space $(X, \|\cdot\|_X)$ - see Section 2.2. Now we have the framework to state the appropriate counterpart to Assumption 3.1.2, for random input data.

Assumption 4.1.1 *(Random Input Data) Assume $\sigma_S = \sigma_S(\omega, \cdot)$, $\sigma = \sigma(\omega, \cdot)$ and $f = f(\omega, \cdot)$ are (possibly dependent or correlated) random C^η fields, for $\eta \in (0,1)$, on each sub-interval of $\{c_j\}_{j=1}^{\aleph}$ (i.e. C_{pw}^η random fields), such that*

(a) σ, $\sigma_S \in L_p(\Omega; \mathrm{C}_{pw}^\eta)$ *and* $(\sigma_S)_{\min}^{-1}, \sigma_{\min}^{-1} \in L_p(\Omega)$, *for all $p \in [1, \infty)$;*

(b) $f \in L_{p_*}(\Omega; \mathrm{C}_{pw}^\eta)$, *for some $p_* \in (1, \infty]$.*

The set $\{c_j\}_{j=1}^{\aleph}$ is defined as in Assumption 3.1.2. Moreover, we will again assume that the value of each function at c_j is taken to be the right limit for $j = 1, \cdots, \aleph - 1$, and the left limit for $j = \aleph$.

A class of suitable random fields that can be shown to satisfy Assumption 4.1.1, for some $\eta >$ 0 and for all $p_* \in [1, \infty)$, are log-normal random fields with an underlying Matérn covariance (see Section 2.3.1), provided the means of $\log \sigma_S$, $\log \sigma$ and $\log f$ belong to C_{pw}^η and the Matérn smoothness parameter ν is sufficiently big for each of the fields [199]. This will be discussed further in Section 4.3.

Finally, for simplicity, we restrict ourselves to deterministic $\sigma_A = \sigma_A(x) \in C_{pw}^\eta$ with

$$0 \; < \; (\sigma_A)_{\min} \; \leq \; \sigma_A(x) \; \leq \; (\sigma_A)_{\max} \; < \; \infty \;, \quad \text{for all} \quad x \in [0,1] \;, \tag{4.1.1}$$

i.e., the absorption does not vanish anywhere and it is known. This implies that the distributions of $\sigma = \sigma_S + \sigma_A$ and σ_S differ only in their mean. This is by no-means a necessary assumption, and the results generalise also to $\sigma_A(\omega, \cdot) \in L_p(\Omega; C_{pw}^\eta)$, for all $p \in [1, \infty)$.

To achieve a probabilistic equivalent to Theorem 3.3.11, the main result in the earlier part of this paper, we only need to show that $\mathcal{R}(\sigma, \sigma_S) \in L_p(\Omega)$. Recall that we write $\overline{a}$ to denote $\max\{1, a\}$, for any scalar value a, and we note that in the case of a scalar random variable $a \in L_p(\Omega)$ we also have $\overline{a} \in L_p(\Omega)$.

Lemma 4.1.2 *Consider the auxiliary functions $\mathcal{R}_i(\sigma, \sigma_S)$, $i = 1, \ldots, 4$, defined in (3.2.27), (3.2.32), (3.3.28) and (3.3.29), respectively, where σ_S and σ are assumed to satisfy Assumption 4.1.1. Then, for all $p \in [1, \infty)$,*

$$\mathcal{R}_i(\sigma, \sigma_S) \in \; L_p(\Omega), \quad \text{for all} \;\; i = 1, \cdots, 4, \tag{4.1.2}$$

and hence, the constant in Theorem 3.3.11 also satisfies $\mathcal{R}(\sigma, \sigma_S) \in L_p(\Omega)$, for all $p \in [1, \infty)$.

Proof. Consider $\mathcal{R}_i\big(\sigma(\omega, \cdot), \sigma_S(\omega, \cdot)\big)$, for any $i = 1, \ldots, 4$ and for any $\omega \in \Omega$. To simplify the presentation, throughout this proof we will suppress the ω notation, i.e. we write $\sigma_S = \sigma_S(\omega, \cdot)$, $\sigma = \sigma(\omega, \cdot)$. Note first that since $\sigma = \sigma_S + \sigma_A$ we have $\sigma_{\min}^{-1} \leq (\sigma_A)_{\min}^{-1}$, which we assumed to be a constant independent of ω. Moreover, any of the terms

$$\sigma_{\max}, \; (\sigma_S)_{\max}, \; \|\sigma\|_{\eta, pw}, \; \|\sigma_S\|_{\eta, pw}, \; \overline{\sigma_{\max}}, \; \overline{(\sigma_S)_{\max}} \; \text{and} \; \overline{\|\sigma\|_{\eta, pw}}$$

can be bounded by a constant c times $\overline{\|\sigma_S\|_{\eta, pw}}$, where c may depend on $(\sigma_A)_{\max}$ or $\|\sigma_A\|_{\eta, pw}$, but is again independent of ω. From here it is easy to check that the functions $\mathcal{R}_i(\sigma, \sigma_S)$, $i = 1, \cdots, 4$, can be bounded by an expression of the form

$$c \, (\sigma_S)_{\min}^{-r_1} \, \overline{\|\sigma_S\|_{\eta, pw}}^{r_2/2} \, \left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty \right)^{-1} \;, \tag{4.1.3}$$

for some $r_1 \in \{0, 1, 2\}$ and $r_2 \in \mathbb{N}$, where $c > 0$ is a constant independent of ω (but possibly dependent on $(\sigma_A)_{\min}^{-1}$ and $\|\sigma_A\|_{\eta, pw}$).

Let $p \in [1, \infty)$. To see that each of the functions $\mathcal{R}_i(\sigma, \sigma_S)$ is in $L_p(\Omega)$ let us first show that $(1 - \|\sigma_S/\sigma\|_\infty)^{-1} \in L_p(\Omega)$. Re-writing

$$0 \; < \; \frac{\sigma_S}{\sigma} \; = \; \frac{\sigma_S}{\sigma_S + \sigma_A} \; = \; 1 - \frac{\sigma_A}{\sigma_S + \sigma_A} \; \leq \; 1 - \frac{(\sigma_A)_{\min}}{(\sigma_S)_{\max} + (\sigma_A)_{\max}} \; < \; 1 \;,$$

it follows that

$$\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty \right)^{-1} \leq \left(1 - \left(1 - \frac{(\sigma_A)_{\min}}{(\sigma_S)_{\max} + (\sigma_A)_{\max}}\right)\right)^{-1} = \frac{(\sigma_S)_{\max} + (\sigma_A)_{\max}}{(\sigma_A)_{\min}} \; \leq \; c \, \overline{(\sigma_S)_{\max}} \;,$$

which is in $L_p(\Omega)$ due to Assumption 4.1.1.

Substituting this bound into (4.1.3), we deduce that each of the functions $\mathcal{R}_i(\sigma, \sigma_S)$ is bounded by a product of random variables that are in $L_p(\Omega)$, for all $1 \le p < \infty$. Therefore, (4.1.2) is a consequence of the Hölder and Minkowski inequalities and it follows that $\mathcal{R}(\sigma, \sigma_S) \in L_p(\Omega)$, for all $1 \le p < \infty$. $\blacksquare$

We now establish a probabilistic counterpart of Theorem 3.3.11. To simplify the presentation, let us suppose in the following that

$$N = N(h) = \max \left\{ \lceil ch^{-\eta} \rceil, 4 \right\}, \qquad (4.1.4)$$

for some constant $c > 0$ independent of h and ω and where $\eta \in (0,1)$ is given in Assumption 4.1.1. The assumption $N \ge 4$ guarantees that $h \log N > h$ and so

$$h \le N^{-1} + h \log N + h^\eta \le c' h^\eta, \qquad (4.1.5)$$

for some different constant $c' > 1$ independent of h and ω.

From the definition of $\mathcal{R}_3(\sigma, \sigma_S)$ in (3.3.28) and Assumption 4.1.1, $\mathcal{R}_3(\sigma, \sigma_S)$ does not in general belong to $L_\infty(\Omega)$ and hence it is impossible to satisfy the stability constraint (3.3.28) uniformly across all samples $\omega \in \Omega$, for any fixed mesh size $h > 0$. Instead, we will consider a *sample-dependent mesh size*.

Let $h > 0$ be chosen arbitrarily, but independently of ω such that $h \log N(h) \le 1$. For each $\omega \in \Omega$, let $\overline{h}_\omega$ be the largest possible value for the mesh width such that the stability constraint (3.3.28) is satisfied with $\sigma_S = \sigma_S(\omega, x)$, $\sigma = \sigma(\omega, x)$ and $N = N(\overline{h}_\omega)$. The stable numerical approximation of the random field $\phi(\omega, x)$ with maximum mesh width h is then defined to be the field

$$\Phi^h(\omega, x) := \phi^{h_\omega, N(h_\omega)}(\omega, x),$$

where for each ω, the approximation $\phi^{h_\omega, N(h_\omega)}(\omega, x)$ is computed as described in Section 3.1.1 above, with spatial mesh size

$$h_\omega := \min\{h, \overline{h}_\omega\} \qquad (4.1.6)$$

and with $2N(h_\omega)$ quadrature points.

This choice of h_ω guarantees that (3.3.28) is satisfied for each realisation of $\sigma_S(\omega, x)$ and $\sigma(\omega, x)$. However, since the random fields σ_S and σ are in general not uniformly bounded away from 0 or infinity in Assumption 4.1.1, even if h is small there will be a set $\Omega_{\text{bad}} \subset \Omega$ (of non-zero measure) such that $h_\omega < h$, for $\omega \in \Omega_{\text{bad}}$, i.e. a set of realisations where the numerical approximation with mesh width h is not guaranteed to be stable. Due to Assumption 4.1.1(a) the measure of the set Ω_{bad} converges to 0 in the limit, as $h \to 0$ - we prove this in Section E.4.

Theorem 4.1.3 *Let $h \in (0,1)$ be such that $h \log N(h) \le 1$ and suppose that Assumption 4.1.1 holds for some $\eta \in (0,1)$ and for some $p_* \in (1, \infty]$. Then, $\Phi^h(\omega, \cdot)$ exists for all $\omega \in \Omega$ and for any $1 \le p < r \le p_*$, there exists a positive constant $c(p, r) > 0$ such that*

$$\|\phi - \Phi^h\|_{L_p(\Omega; L_\infty)} \le c(p, r) \|f\|_{L_r(\Omega; \mathrm{C}_{pw}^\eta)} h^\eta, \qquad (4.1.7)$$

Proof.

Using the definition of Φ^h, Theorem 3.3.11, (4.1.4) and (4.1.5) together with Hölder's inequality

we have

$$
\begin{aligned}
\|\phi - \Phi^h\|_{L_p(\Omega;L_\infty)}^p &= \mathbb{E}\left[\|\phi - \phi^{h_\omega, N(h_\omega)}\|_\infty^p\right] \\
&\leq \mathbb{E}\left[\left(\mathcal{R}(\sigma,\sigma_S)\,c'\,h_\omega^\eta\,\|f(\omega,\cdot)\|_{\eta,pw}\right)^p\right] \\
&\leq (c'\,h^\eta)^p\,\left(\mathbb{E}\left[\mathcal{R}(\sigma,\sigma_S)^q\right]\right)^{p/q}\left(\mathbb{E}\left[\|f(\omega,\cdot)\|_{\eta,pw}^r\right]\right)^{p/r} \\
&= \left(c(p,r)\,\|f\|_{L_r(\Omega;\mathrm{C}_{pw}^\eta)}\,h^\eta\right)^p\;,
\end{aligned}
$$

where

$$
q^{-1} + r^{-1} \;=\; p^{-1}\;,\quad\text{and}\quad c(p,r) \;:=\; c'\,\|\mathcal{R}(\sigma,\sigma_S)\|_{L_q(\Omega)} < \infty
$$

due to Lemma 4.1.2. ∎

Remark 4.1.4 (Uniform Random Input Data) The cross sections σ_S and σ are not assumed to be bounded away from 0 or infinity uniformly in Assumption 4.1.1(a). If we strengthen the assumption to hold for $p = \infty$ (i.e. uniformly bounded random fields) then we can choose $p = r = p_*$ in (4.1.7). In particular, if $p_* = \infty$ then $\|\phi - \Phi^h\|_{L_\infty(\Omega;L_\infty)}$ converges with order h^η. Moreover, since $p = \infty$ in Assumption 4.1.1 ensures $\mathcal{R}_3(\sigma,\sigma_S) \in L_\infty(\Omega)$, there exists a $h_{\max} > 0$, such that $\overline{h}_\omega \geq h_{\max}$ for all $\omega \in \Omega$. As a consequence, for all $h \leq h_{\max}$, $\Omega_{\mathrm{bad}} = \emptyset$ and the stability constraint (3.3.28) is satisfied with uniform mesh size $h_\omega \equiv h$. As already outlined, this is not the case under Assumption 4.1.1, in general.

For the general case of sample-dependent discretisations it is important to discuss the *expected computational cost* per sample. Recall that for $\omega \in \Omega$, we choose h_ω according to (4.1.6) and $N(h_\omega)$ according to (4.1.4). Let us assume that with those choices of discretisation parameters, the cost to compute one sample is bounded by

$$
\mathcal{C}(\phi^{h_\omega, N(h_\omega)}) \;\leq\; c'(\omega)h_\omega^{-\gamma}\;, \tag{4.1.8}
$$

for some $\gamma > 0$. In Theorem 4.1.6 below, we give examples of solvers for (3.1.7)-(3.1.8) where (4.1.8) holds with $\gamma \in [1,3]$ and $c' \in L_p(\Omega)$, for some $p > 1$. In the following lemma, we will show that the expected cost to compute a sample of Φ^h will be of order $h^{-\gamma}$, even though some samples $\Phi^h(\omega,x)$ may need to be computed with a spatial mesh size $h_\omega \ll h$. This result exploits the fact that the measure of Ω_{bad} goes to 0 as $h \to 0$.

Lemma 4.1.5 *Let $h \in (0,1)$ such that $h \log N(h) \leq 1$ and suppose that Assumption 4.1.1 holds for some $\eta \in (0,1)$. Furthermore, suppose that the computational cost $\mathcal{C}(\Phi^h(\omega,x))$ for solving (3.1.7)-(3.1.8), for a given $\omega \in \Omega$, satisfies (4.1.8) with $c' \in L_p(\Omega)$ for some $p > 1$. Then,*

$$
\mathbb{E}[\mathcal{C}(\Phi^h)] \;=\; \mathcal{O}\left(h^{-\gamma}\right)\;,
$$

and the hidden constant is independent of h.

Proof. Using (4.1.8), the first inequality in (4.1.5) and the definition of $\overline{h}_\omega$, we get

$$
\begin{aligned}
\mathcal{C}(\Phi^h(\omega,x)) &\leq c'(\omega)h_\omega^{-\gamma} \\
&= c'(\omega)\max\{h^{-\gamma},\,(\overline{h}_\omega)^{-\gamma}\} \\
&\leq c'(\omega)\left(h^{-\gamma} + \mathcal{R}_3(\sigma(\omega,\cdot),\sigma_S(\omega,\cdot))^\gamma\right)\;. \tag{4.1.9}
\end{aligned}
$$

Now, taking the expectation of (4.1.9) and applying Hölders inequality

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{C}(\Phi^h(\omega, x))\right] &\leq \mathbb{E}\left[c'(\omega)\right] h^{-\gamma} + \mathbb{E}\left[c'(\omega)\mathcal{R}_3(\sigma(\omega, \cdot), \sigma_S(\omega, \cdot))^\gamma\right] \\
&\leq \mathbb{E}\left[c'(\omega)\right] h^{-\gamma} + \left(\mathbb{E}\left[c'(\omega)^p\right]\right)^{1/p}\left(\mathbb{E}\left[\mathcal{R}_3(\sigma(\omega, \cdot), \sigma_S(\omega, \cdot))^{\gamma q}\right]\right)^{1/q},
\end{aligned}
$$

where $1 \leq q \leq \infty$ is such that $p^{-1} + q^{-1} = 1$. By Theorem 4.1.2 and Assumption 4.1.1, $\mathcal{R}_3(\sigma, \sigma_S)^\gamma \in L_q(\Omega)$ and so the result follows, since h was assumed to be less than 1. ■

Example 4.1.6 We will now give examples of Lemma 4.1.5 in practice. In Section 3.1.2, two methods for computing the solution to (3.1.7) – (3.1.8) are presented. The first method is a direct solver where first ψ is eliminated from the coupled system (3.1.7) – (3.1.8) and then LU factorisation is applied to the resulting Schur complement system (i.e. the direct solver (3.1.11)). The cost for this method is of order $h_\omega^{-2}\left(N(h_\omega) + h_\omega^{-1}\right)$ (see (3.1.12)) with a constant independent of ω [93]. Using (4.1.4) this implies that (4.1.8) holds with c' independent of ω and $\gamma = 3$.

The second method is a type of Richardson iteration known as *source iteration* (i.e. the iterative solver (3.1.13)). In that case, it can be shown that the cost is of order $h_\omega^{-1}N(h_\omega)$ (see (3.1.14)) with a constant proportional to $-\log\left(\|\sigma_S(\omega, \cdot)/\sigma(\omega, \cdot)\|_\infty\right)^{-1}$ [32, 93]. Using again (4.1.4), this implies that (4.1.8) holds with $\gamma = 1 + \eta$.

Corollary 4.1.7 *Suppose the assumptions of Lemma 4.1.5 hold and system* (3.1.7) – (3.1.8) *is solved with either of the Methods 1 or 2 in Example 4.1.6. Then, condition* (4.1.8) *holds with* $c'(\omega) \in L_\infty(\Omega)$ *(Method 1,* (3.1.11)*) and* $c'(\omega) \in L_p(\Omega)$*, for all* $1 \leq p < \infty$ *(Method 2,* (3.1.13)*). Hence,*

$$
\mathbb{E}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-3}\right) \quad and \quad \mathbb{E}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-1-\gamma}\right), \quad respectively,
$$

and the hidden constants are independent of h*.*

Proof. For the direct solver, c' is independent of ω, and hence trivially $c' \in L_\infty(\Omega)$.

In the case of the iterative solver, $c'(\omega)$ is proportional to $-\log\left(\|\sigma_S(\omega, \cdot)/\sigma(\omega, \cdot)\|_\infty\right)^{-1}$ [32]. Since $\|\sigma_S(\omega, \cdot)/\sigma(\omega, \cdot)\|_\infty \in (0, 1)$, for almost all $\omega \in \Omega$, and since $-\log(1 - y) = \sum_{k=1}^\infty y^k/k > y$, for all $y \in (0, 1)$, we have

$$
c'(\omega) \leq -c\log\left(\left\|\frac{\sigma_S(\omega, \cdot)}{\sigma(\omega, \cdot)}\right\|_\infty\right)^{-1} \leq c\left(1 - \left\|\frac{\sigma_S(\omega, \cdot)}{\sigma(\omega, \cdot)}\right\|_\infty\right)^{-1},
$$

where $c > 0$ is a constant independent of ω. As we have seen in Lemma 4.1.2, this implies that $c' \in L_p(\Omega)$, for all $1 \leq p < \infty$. The expected costs follow from Lemma 4.1.5. ■

4.1.2 Multilevel Monte Carlo Acceleration

From a practical point of view, we are interested in methods that allow us to *accurately and efficiently* estimate the statistics of functionals of ϕ (or ψ). We focus on estimating $\mathbb{E}[Q]$, the expected value of Q, where $Q(\omega) \in \mathbb{R}$ denotes a functional $\mathfrak{L}$ of ϕ representing some quantity we are interested in. Here, we will consider the non-linear functional

$$
Q(\omega) = \|\phi\|_1^q := \left(\int_0^1 |\phi(\omega, x)|\, dx\right)^q, \quad \text{for some integer } 1 \leq q < \infty, \tag{4.1.10}
$$

i.e. the qth moment of the spatial average of $|\phi|$ (over $[0, 1]$), but we note the methodology also applies to many other functionals. For each $\omega \in \Omega$, we will approximate the random variable $Q(\omega)$

by

$$Q_h(\omega) \;=\; \left(h \sum_{j=1}^{M} |\Phi^h(\omega, x_j)| \right)^q , \tag{4.1.11}$$

where $\Phi^h(\omega, \cdot)$ is the approximation of $\phi(\omega, \cdot)$ using a spatial discretisation with mesh size h_ω given by (4.1.6) and a double Gauss rule with $2N(h_\omega)$ angular points defined by (4.1.4). Note that the cost of computing Q_h is negligible compared to the cost of finding Φ^h itself.

From Section 2.4, we recall the (standard) Monte Carlo (MC) estimator

$$\widehat{Q}_h^{MC} \;:=\; \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} Q_h(\mathbf{Z}^n) , \tag{4.1.12}$$

where N_{MC} is the number of Monte Carlo points/samples and $\mathbf{Z}^n := (\sigma^n, \sigma_S^n, f^n)^T$, where (for each $n \in \mathbb{N}$) we assume $\sigma^n, \sigma_S^n, f^n$ denote, respectively, realisations of the random fields σ, σ_S and f - they are assumed to satisfy Assumption 4.1.1.

Moreover, recall the assumptions (2.4.3), (2.4.4), i.e. there exist two constants $\alpha, \gamma > 0$ such that

$$\left| \mathbb{E}\left[Q - Q_h \right] \right| \;=\; \mathcal{O}\left(h^\alpha \right) , \tag{4.1.13}$$

$$\mathbb{E}\left[\mathcal{C}(Q_h) \right] \;=\; \mathcal{O}\left(h^{-\gamma} \right) , \tag{4.1.14}$$

which allowed us to prove (in (2.4.13)) that the mean ϵ-cost of the standard Monte Carlo estimator is

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC}) \right] \;=\; \mathbb{E}\left[\sum_{n=1}^{N_{MC}} \mathcal{C}(Q_h(\mathbf{Z}^n)) \right] \;=\; N_{MC}\, \mathbb{E}\left[\mathcal{C}(Q_h) \right] \;=\; \mathcal{O}\left(\epsilon^{-2 - \frac{\gamma}{\alpha}} \right) . \tag{4.1.15}$$

Likewise, recall the MLMC estimator

$$\widehat{Q}_h^{MLMC} \;:=\; \sum_{\ell=0}^{L} \widehat{Y}_\ell^{MC} \;=\; \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} Y_\ell(\mathbf{Z}^{\ell,n}) , \tag{4.1.16}$$

where $\{\mathbf{Z}^{\ell,n}\}_{n=1}^{N_\ell}$ denotes the set of i.i.d. samples on level ℓ, of the random fields σ, σ_S, f, chosen independently from the samples on the other levels. The optimal choice of N_ℓ is given in (2.4.32). Using the assumption (2.4.33), i.e. that there exists $\beta > 0$ such that

$$\mathbb{V}[Y_\ell] \;=\; \mathcal{O}\left(h_\ell^\beta \right) , \tag{4.1.17}$$

then it can be proven (see Theorem 2.4.4) that the mean ϵ-cost of the multilevel Monte Carlo estimator is

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC}) \right] \;=\; \mathcal{O}\left(\epsilon^{-2 - \max\{0,\, (\gamma - \beta)/\alpha\}} \right) , \quad \text{for} \;\; \beta \neq \gamma , \tag{4.1.18}$$

with a similar result when $\beta = \gamma$.

Given the clear importance of the parameters α, β, γ, we would now like to estimate them theoretically. We have already estimated the parameter γ for two different solvers in Corollary 4.1.7 using the sample dependent mesh size. Let us now estimate α and β for the functional Q of ϕ in (4.1.10). These estimates can be deduced directly from the general estimate for $\|\phi - \Phi^h\|_{L_p(\Omega; L_\infty)}$ in Theorem 3.3.11, following the proof of [47, Prop. 4.2].

Theorem 4.1.8 *Consider the quantity of interest Q with $q \in [1, \infty)$ and its approximation Q_h defined in (4.1.10) and in (4.1.11), respectively. Let $h \in (0, 1)$ such that $h \log N(h) \leq 1$ and*

suppose that Assumption 4.1.1 holds for some $\eta \in (0,1)$ and for some $p_* > 2q$. Then, the bounds in (4.1.13) and (4.1.17) hold with

$$\alpha = \eta \quad and \quad \beta = 2\alpha ,$$

Proof. Let $\delta \in (0,1)$ and $\alpha = \min\{\eta, 1 - \delta\}$. Note first that it follows from (4.1.4), (4.1.5) and the assumptions made on h that there exists a constant $c > 0$ such that, for any $\omega \in \Omega$,

$$N(h_\omega)^{-1} + h_\omega \log N(h_\omega) + h_\omega^\eta \ \leq \ \min\{c\, h_\omega^\alpha, 4\} . \tag{4.1.19}$$

Also note that $\big|\mathbb{E}[Q - Q_h]\big| \leq \mathbb{E}\big[|Q - Q_h|\big]$ and that

$$\mathbb{V}[Y_\ell] \ \leq \ \mathbb{E}\big[Y_\ell^2\big] \ \leq \ 2 \left(\mathbb{E}\left[|Q - Q_{h_\ell}|^2\right] + \mathbb{E}\big[|Q - Q_{h_{\ell-1}}|^2\big] \right) .$$

Thus, to establish bounds of the form (4.1.13) and (4.1.17) it suffices to bound the expected value of $|Q - Q_h|^k$, for $k = 1$ and 2, in terms of h.

For each $\omega \in \Omega$, let the random variable $Q_h = Q_{h_\omega}(\omega)$ denote the quantity of interest in (4.1.11) computed using the sample dependent mesh size in (4.1.6) with $\Phi^h(\omega, \cdot) = \phi^{h_\omega, N(h_\omega)}(\omega, \cdot)$. By using the expansion $a^q - b^q = (a - b) \sum_{j=0}^{q-1} a^{q-1-j} b^j$, we have

$$Q - Q_h \ = \ \left(\int_0^1 |\phi| dx \right)^q - \left(\int_0^1 |\Phi^h| dx \right)^q \ = \ \left(\int_0^1 |\phi| - |\Phi^h| dx \right) \sum_{j=0}^{q-1} \|\phi\|_1^{q-1-j} \|\Phi^h\|_1^j .$$

Applying the absolute value and using the reverse and the standard triangle inequality, as well as the fact that $\|\phi\|_1 \leq \|\phi\|_\infty$, we can deduce that

$$\begin{aligned}
|Q - Q_h| \ &\leq \ \|\phi - \Phi^h\|_1 \sum_{j=0}^{q-1} \|\phi\|_1^{q-1-j} \|\Phi^h\|_1^j \\
&\leq \ \|\phi - \Phi^h\|_\infty \sum_{j=0}^{q-1} \|\phi\|_\infty^{q-1-j} \left(\|\phi\|_\infty + \|\phi - \Phi^h\|_\infty \right)^j . \tag{4.1.20}
\end{aligned}$$

Now, using Corollary 3.2.9 and Theorem 3.3.11 together with (4.1.19) to bound $\|\phi\|_\infty$ and $\|\phi - \Phi^h\|_\infty$, respectively, it follows that

$$\begin{aligned}
|Q - Q_h| \ &\leq \ c\|\phi - \Phi^h\|_\infty \sum_{j=0}^{q-1} \left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \mathcal{R}_1(\sigma, \sigma_S)\|f\|_\infty \right)^{q-1-j} \\
&\qquad \left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \mathcal{R}_1(\sigma, \sigma_S)\|f\|_\infty + 4\mathcal{R}(\sigma, \sigma_S)\|f\|_{\eta, pw} \right)^j ,
\end{aligned}$$

for some constant $c > 0$. Hence, using again Theorem 3.3.11, (4.1.19), as well as the inequalities $\|f\|_\infty \leq \|f\|_{\eta, pw}$ and $\mathcal{R}_1(\sigma, \sigma_S) \leq \mathcal{R}(\sigma, \sigma_S)$, we get

$$\begin{aligned}
|Q - Q_h| \ &\leq \ c\|\phi - \Phi^h\|_\infty \left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \mathcal{R}(\sigma, \sigma_S) \|f\|_{\eta, pw} \right)^{q-1} \sum_{j=0}^{q-1} 5^j \\
&\leq \ \frac{c\, 5^q}{4} \left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \right)^q \mathcal{R}(\sigma, \sigma_S)^q \|f\|_{\eta, pw}^q \, h_\omega^\eta . \tag{4.1.21}
\end{aligned}$$

For simplicity in the remaining presentation, let us define

$$\mathcal{R}_5(\sigma, \sigma_S) \;:=\; \overline{\frac{\sigma_{\max}}{\sigma_{\min}}} \, \mathcal{R}(\sigma, \sigma_S) \;,$$

from which it is easy to show $\mathcal{R}_5(\sigma, \sigma_S) \in L_p(\Omega)$, for all $p \in [1, \infty)$, following Theorem 4.1.2.

Finally, since $f \in L_{p_*}(\Omega; \mathrm{C}^\eta_{pw})$ and $\mathcal{R}_5(\sigma, \sigma_S) \in L_p(\Omega)$, for all $p \in [1, \infty)$, it follows from (4.1.21) together with Hölder's inequality that, for $k = 1, 2$,

$$\mathbb{E}\left[|Q - Q_h|^k\right] \;\le\; \left(\frac{c\,5^q}{4}\right)^k \|\mathcal{R}_5(\sigma, \sigma_S)\|^{kq}_{L_r(\Omega)} \, \|f\|^{kq}_{L_{p_*}(\Omega; \mathrm{C}^\eta_{pw})} \, h^{k\eta} \;\le\; c'\, h^{k\eta} \;,$$

where $r \in [1, \infty)$ is defined by $r^{-1} + p_*^{-1} = kq^{-1}$ and $c' > 0$ is a constant independent of h. This completes the proof. ∎

Corollary 4.1.9 *Suppose the assumptions of Theorem 4.1.8 hold and system (3.1.7)-(3.1.8) is solved with Method 2 in Example 4.1.6, and let $\delta \in (0, 1)$ be arbitrary. Then, the ϵ-costs of the Monte Carlo method and of the multilevel Monte Carlo method satisfy, respectively,*

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}^{MC}_h)\right] \;=\; \mathcal{O}\big(\epsilon^{-4-\chi}\big) \quad and \quad \mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}^{MLMC}_h)\right] \;=\; \mathcal{O}\big(\epsilon^{-2-\chi}\big) \;,$$

where $\chi := \max\left\{\dfrac{1 - \eta}{\eta}, \delta\right\} > 0$.

Corollary 4.1.9 implies that, when $\chi = \delta$, the multilevel Monte Carlo method is *optimal* i.e. the computational cost is $\mathcal{O}(\epsilon^{-2})$ (recall from Section 2.4.3 that this is proportional to the cost of a *single* standard Monte Carlo sample). This occurs when $\eta \ge 1$ (which our analysis does not cover, although we will numerically consider this case in the next section).

We will now introduce a more efficient way of computing an estimate of the scalar flux ϕ, when using a sequence of discretisation parameters e.g. MLMC. We note that it is trivial to show that the result of Corollary 4.1.9 extends to this new solver.

4.2 A Hybrid Direct-Iterative Solver

To compute samples of an estimate of the scalar flux, and thus of a quantity of interest, we propose a hybrid version of the direct and the iterative solver for the Schur complement system (3.1.11) described in Section 3.1.2.

The cost of the iterative solver depends on the number K of iterations that we take. For each ω, we aim to choose K such that the L_2-error $\|\phi(\omega, \cdot) - \phi^{(K)}(\omega, \cdot)\|_2 < \epsilon$. To estimate K we fix $h = 1/1024$, $N = (2h)^{-1}$ and $d = 3600$ and use the direct solver to compute $\phi^h = \phi^{h, N(h)}$ for each sample ω. Let $\varrho(\omega) := \|\sigma_S(\cdot, \omega)/\sigma(\cdot, \omega)\|_\infty$. For a sufficiently large number of samples, we then evaluate

$$\frac{\log\left(\left\|\phi^h(\omega, \cdot) \;-\; \phi^{h,(K)}(\omega, \cdot)\right\|_2\right)}{K \log\left(\varrho(\omega)\right)}$$

and find that this quotient is less than $\log(0.5)$ in more than 99% of the cases (at least in the specific case considered in the numerical results in Section 4.3), for $K = 1, \ldots, 150$, so that we can choose $c = 0.5$ in (3.1.16). We repeat the experiment also for larger values of h and smaller values of d to verify that this bound holds in at least 99% of the cases independently of the discretisation parameter h and of the truncation dimensions d.

Hence, a sufficient a-priori condition to achieve $\|\phi^h(\omega) - \phi^{h,(K)}(\omega)\|_2 < \epsilon$ in at least 99% of the cases is

$$K = K(\epsilon, \omega) = \max\left\{ 1, \left\lceil \frac{\log(2\epsilon)}{\log(\varrho(\omega))} \right\rceil \right\}, \qquad (4.2.1)$$

where $\lceil \cdot \rceil$ denotes the ceiling function. It is important to note that K is no longer a deterministic parameter for the solver (like M or N). Instead, K is a random variable that depends on the particular realisation of σ_S. It follows from (4.2.1), using the results in [47, §2], [89] as in Section 3.1, that $\mathbb{E}[K(\epsilon, \cdot)] = \mathcal{O}(\log(\epsilon))$ and $\mathbb{V}[K(\epsilon, \cdot)] = \mathcal{O}\left(\log(\epsilon)^2\right)$, with more variability in the case of the exponential field.

Recall from (3.1.12) and (3.1.14) that, in the case $N = (2h)^{-1} = M/2$, the costs for the direct and iterative solvers are $c_1 M^3$ and $c_2 K M^2$, respectively, for two constants $c_1, c_2 > 0$. In our numerical experiments, we found that in fact $c_1 \approx c_2$, for this particular relationship between M and N. This motivates a third "hybrid" solver, which we present in Algorithm 1, where the iterative solver is chosen when $K(\omega) < M$ and the direct solver when $K(\omega) \geq M$. This allows us to use the optimal solver for each particular sample - in the case of the hierarchies that appear in MLMC, we adaptively adjust the solver dependent on the level and the sample.

We finish this section with a study of timings in seconds (here referred to as the cost) of the three solvers (i.e. (3.1.11), (3.1.13) and Algorithm 1). In Fig. 4-1, we plot the average cost (over $16,384$ MC samples) divided by M_ℓ^3, against the level parameter ℓ. We observe that, as expected, the (scaled) expected cost of the direct solver is almost constant and the iterative solver is more efficient for larger values of M_ℓ. Over the range of values of M_ℓ considered in our experiments, a best fit for the rate of growth of the cost with respect to the discretisation parameter h_ℓ in (2.4.4) is $\gamma \approx 2.1$, for both fields (asymptotically this is the same as the source iteration method).

Algorithm 1: Hybrid direct-iterative solver of (3.1.10), for one realisation of the input data

Data: Realisation of input data, σ_S and σ;

Data: A desired accuracy ϵ;

Result: An estimate of the scalar flux

1 Compute $\varrho := \|\sigma_S/\sigma\|_\infty$ and

$$K = \left\lceil \frac{\log(2\epsilon)}{\log(\varrho)} \right\rceil$$

   if $K < M$ **then**

2   Compute $\Phi = \Phi^{(K)}$ in (3.1.10), using K (source) iterations of the method (3.1.13);

3 **else**

4   Compute Φ in (3.1.10), using the direct method (3.1.11);

5 **end**

4.3 Numerical Results

Until now, the random structure assumed in Assumption 4.1.1 has been unspecified. For our numerical results, we assume $\log \sigma_S$ is a correlated zero mean Gaussian random field, with covariance function defined by the Matérn class (recall Section 2.3.1)

$$C_\nu(x, y) = \sigma_{var}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(2\sqrt{\nu} \frac{|x - y|}{\lambda_C} \right)^\nu K_\nu \left(2\sqrt{\nu} \frac{|x - y|}{\lambda_C} \right). \qquad (4.3.1)$$
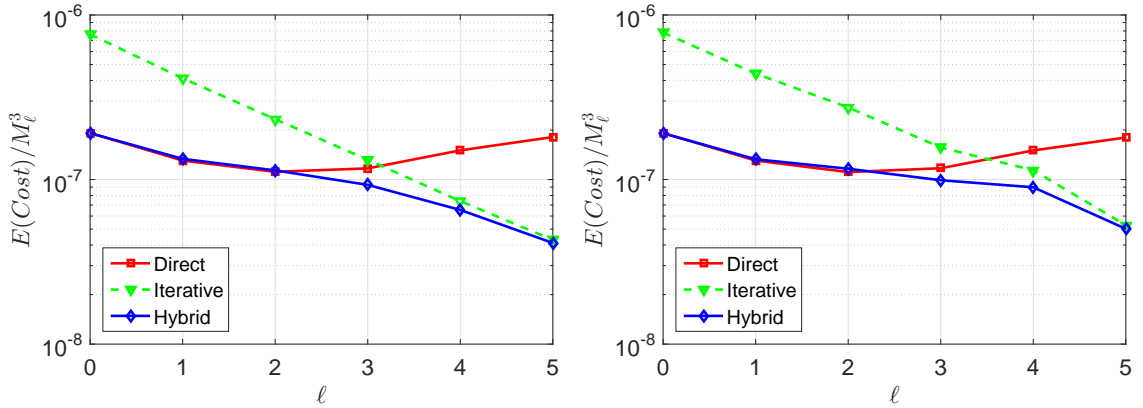
Figure 4-1: Comparison of the average costs of the solvers (actual timings in seconds divided by M_ℓ^3) for the (Left) Matérn field and (Right) exponential field.

It is parametrised by the smoothness parameter $\nu \geq 0.5$; λ_C is the correlation length, σ_{var}^2 is the variance, Γ is the gamma function and K_ν is the modified Bessel function of the second kind.

To sample from σ_S we use the (d-truncated) Karhunen-Loève (KL) expansion of $\log \sigma_S$, i.e.,

$$\log \sigma_S(x, \omega) \;=\; \sum_{i=1}^{d} \sqrt{\xi_i} \; \eta_i(x) \; y_i(\omega) \;, \tag{4.3.2}$$

where $y_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and the ξ_i and η_i are the eigenvalues and the $L_2(0,1)$-orthogonal eigenfunctions of the covariance integral operator associated with kernel given by the covariance function. This was discussed in more detail in Section 2.3.2.

For our specific examples we consider σ_S as two Matérn fields that belong to C^η, for some $\eta > 0$. For the first case, we choose $\nu = \eta = 0.5$[1]. This corresponds to the exponential covariance (i.e. (2.3.6)) and in the following will be referred to as the 'exponential field'. For the second case, denoted the 'Matérn field', we choose $\eta = 1.5$. The correlation length and variance for both cases are $\lambda_C = 1$ and $\sigma_{var}^2 = 1$, respectively. As in Section 3.4, we assume $\sigma_A \equiv \exp(0.25)$ and $f \equiv \exp(1)$.

For the discretisation, we choose the same uniform spatial mesh and quadrature rule as in Section 3.4 and fix the number of angles by $N = (2h)^{-1}$ for the Matérn field, and $N = 2\lceil 2h^{-1/2} \rceil$ for the exponential field. We empirically choose the (finite) number of KL terms as $8h^{-1}$ for the Matérn field and $225h^{-1/2}$ for the exponential field. This is to ensure that the error due to the truncation is negligible compared with other errors. We note that, even for such large number of KL modes, the sampling cost of the KL expansion does not dominate because the randomness only exists in the (one) spatial dimension.

We will consider two measurements of error in the mean, $\mathbb{E}[\|\phi - \phi^{h,N(h)}\|_\infty]$ and $\mathbb{E}[\|Q - Q_h\|_\infty]$ (where Q is defined in (4.1.10) and we take $q = 1$). We estimate ϕ by a reference solution with $h^{-1} = 512$, $N = 256$, and we choose 2048 / 3600 KL modes for the Matérn and exponential fields, respectively. The expectation is estimated using a standard Monte Carlo estimator (cf. (2.4.8) or [93]) with $32,768$ samples.

In Figure 4-2, we present the numerical results which support the error estimate in Theorem 4.1.3. For this error estimate, we observe $\mathcal{O}(h)$ convergence, for both random fields, despite

---

[1]we recall from Chapter 2, that a Gaussian random field equipped with a Matérn covariance function (with smoothing parameter ν) is $\lceil \nu \rceil - 1$ times differentiable
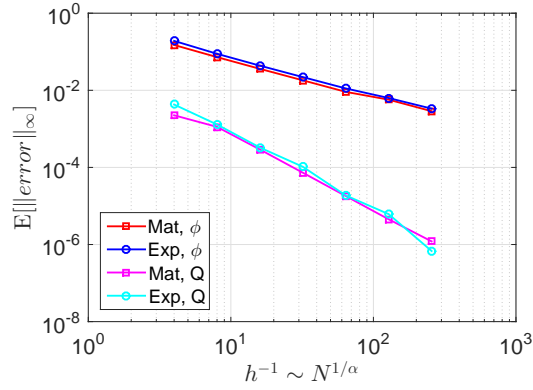
Figure 4-2: Convergence of the mean error(s) $\mathbb{E}[\|\phi - \phi^{h,N(h)}\|_\infty]$ and $\mathbb{E}[\|Q - Q_h\|_\infty]$, with $N = (2h)^{-1}$ (Matérn) and $N = 2\lceil 2h^{-1/2}\rceil$ (Exponential). In the image, 'Mat' and 'Exp' refer to the Matérn and exponential random fields respectively, and ϕ and Q denote the mean error in ϕ and Q respectively.
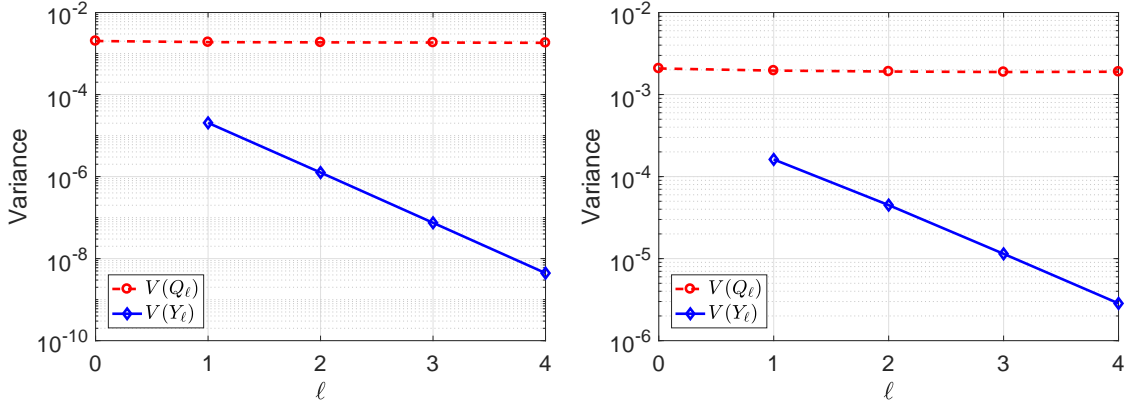
Figure 4-3: Comparison of the variance of the quantity of interest on each level, Q_ℓ, and the variance of the difference process $Y_\ell = Q_\ell - Q_{\ell-1}$, for two random fields. (Left) Matérn and (Right) exponential field.

$\eta = 0.5$ for the exponential random field. For deterministic or random fields, there exists a $\mathcal{O}(h^\eta)$ convergence part which would ensure our estimates are sharp. However, in practice we do not always observe it. We also note the accelerated $\mathcal{O}(h^2)$ convergence for the error $\mathbb{E}[\|Q - Q_h\|_\infty]$, which arises due to the additional smoothness Q has over ϕ.

We now wish to study the efficiency of the Monte Carlo method(s) and justify the underlying assumptions (2.4.3), (2.4.4) and (2.4.33) (and (2.4.24) for QMC), we numerically estimate the parameters α, γ and β (and λ) in those assumptions. In Figure 4-2, we already observed that $\mathbb{E}[\|Q - Q_h\|_\infty] = \mathcal{O}(h^2)$ for both random fields, i.e. $\alpha = 2$. Likewise, in Figure 4-1 we observed that $\gamma \approx 2.1$ (for $N = N(h) = (2h)^{-1}$, for the Matérn field) - we also observed $\gamma \approx 1.5$ for the exponential field (where $N = 2\lceil 2h^{-1/2}\rceil$). Moreover, from Figure 4-3 we observe $\beta \approx 4.1$ and $\beta \approx 2.0$, for the Matérn and exponential fields, respectively. These rates are summarised in Table 4.1.

To estimate the parameter λ in (2.4.38), we need to study the convergence rate of the QMC method with respect to the number of samples N_{QMC}. This study is illustrated in Fig. 4-4. As expected, the sampling error of the standard MC estimator converges with $\mathcal{O}(N_{MC}^{-1})$. On the other hand, we observe that the sampling error of the QMC estimator converges approximately with
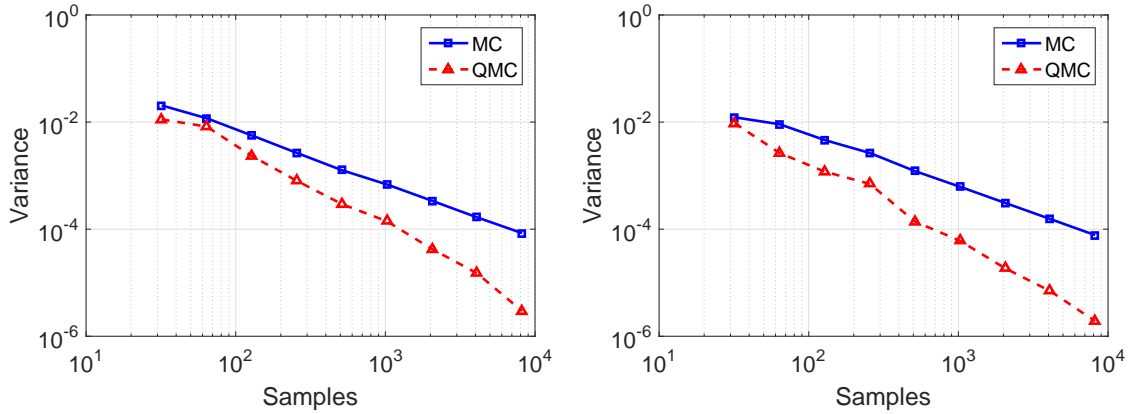
Figure 4-4: Sampling error plotted against the number of samples, for the standard Monte Carlo and quasi-Monte Carlo estimators: (Left) Matérn and (Right) exponential field.

$\mathcal{O}(N_{QMC}^{-1.6})$ and $\mathcal{O}(N_{QMC}^{-1.4})$ (or $\lambda = 0.62$ and $\lambda = 0.71$) for the Matérn field and for the exponential field, respectively. This estimate of λ is also summarised in Table 4.1.

Furthermore, in Figure 4-5 we present numerical results comparing the computational ϵ-cost of the standard, quasi, multilevel and multilevel quasi-Monte Carlo methods, when estimating $\mathbb{E}[Q]$. We observe r, where the computational cost is ϵ^{-r}, to be: 3.3, 2.2, 2.1 and 1.3, for the Matérn field; and 3.4, 2.8, 2.4, 2.3, for the exponential field; for the standard, quasi, multilevel and multilevel quasi-Monte Carlo methods respectively. To help the unfamiliar reader with the computational ϵ-cost plots presented in this thesis (e.g. Figure 4-5), we outline some important details of these plots below in Remark 4.3.1.

Remark 4.3.1 *Computational ϵ-cost plots, such as Figure 4-5, are taken on a log-log scale. They measures the change in (log) cost (along the y-axis) compared with the change in (log) root-MSE accuracy (see (2.4.2), along the x-axis).*

The plots should be read right to left (for increasing accuracy) and from the bottom to the top (for increasing cost). Initially it may seem unintuitive to read right to left, however the plot is structured in this way as the $\epsilon > 0$ is larger to the right and smaller to the left.

To estimate the rate of growth of the computational ϵ-cost of a particular estimator, we compute the gradient of the curve. That is, if $(\epsilon_1, \mathcal{C}_1)$ and $(\epsilon_2, \mathcal{C}_2)$ each denote an accuracy and cost pair (that lies on the line), such that $\epsilon_1 > \epsilon_2$ and $\mathcal{C}_1 \leq \mathcal{C}_2$, then the rate r (where the computational cost is ϵ^{-r}) can be estimated by

$$r = \frac{\log\left(\frac{\mathcal{C}_2}{\mathcal{C}_1}\right)}{\log\left(\frac{\epsilon_1}{\epsilon_2}\right)}.$$

For example, for the MLMC estimator in the left plot in Figure 4-5, we see that (approximately) $\mathcal{C}_1 = 10^{-2}$ cost corresponds to $\epsilon_1 = 10^{-3}$ root-MSE accuracy and $\mathcal{C}_2 = 10^{+2}$ cost corresponds to $\epsilon_2 = 10^{-5}$ root-MSE accuracy. Hence,

$$r = \frac{\log\left(\frac{10^{+2}}{10^{-2}}\right)}{\log\left(\frac{10^{-3}}{10^{-5}}\right)} = \frac{\log(10,000)}{\log(100)} = 2.$$

In Table 4.1 we give compare the rates of: (i) the theoretically estimated ϵ-cost, using (2.4.13), (2.4.27), Theorem 2.4.4 (also Corollary 4.1.9) and (2.4.39), along with the theoretically derived

	η	α_{theo}	β_{theo}	γ_{theo}	Theoretical	α_{obs}	β_{obs}	γ_{obs}	λ_{obs}	Estimated	Observed
MC					4.0					3.1	3.3
MLMC					2.0				-	2.0	2.1
QMC	1.5	≈ 1.0	≈ 2.0	2.0		2.0	4.1	2.1		2.3	2.2
MLQMC					-				0.62	1.2	1.3
MC					5.0					2.8	3.4
MLMC					3.0				-	2.0	2.4
QMC	0.5	0.5	1.0	1.5		2.0	2.0	1.5		2.2	2.8
MLQMC					-				0.71	1.4	2.3

Table 4.1: Summary of computational ϵ-cost rates r, where for an estimator $\widehat{Q}$, $\mathbb{E}[\mathcal{C}(\widehat{Q})] = \mathcal{O}(\epsilon^{-r})$. We estimate r in the following ways: 'Theoretical' uses the parameters in (4.3.3); 'Estimated' uses the numerically observed α_{obs}, β_{obs}, γ_{obs} and λ_{obs}; 'Observed' uses the observed rates from Figure 4-5.

rates (4.3.3), i.e.[2]

$$\alpha_{\text{theo}} \;=\; \min\{\eta, 1 - \delta\} \;, \quad \beta_{\text{theo}} \;=\; 2\alpha_{\text{theo}} \;, \quad \gamma_{\text{theo}} \;=\; \min\{2, 1 + \eta\} \;; \qquad (4.3.3)$$

(ii) the numerically estimated ϵ-cost, using (2.4.13), (2.4.27), Theorem 2.4.4 and (2.4.39), along with the numerically observed parameters α_{obs}, β_{obs}, γ_{obs} presented in the table; (iii) the numerically observed ϵ-cost, estimated from Figure 4-5 and already mentioned in the previous paragraph.

The discrepancy between α_{obs} and β_{obs} compared with the theoretical parameters in (4.3.3) leads to the theoretically proved ϵ-cost rate(s) in Table 4.1 to be pessimistic. As noted above, this is due to not always observing the $\mathcal{O}(h^\eta)$ convergence part for certain problems, and because Q is smoother than ϕ. However, we observe good agreement between the mean ϵ-cost of the Monte Carlo estimator (4.1.15) and the multilevel Monte Carlo estimator (4.1.18), for the numerically observed parameters.

We conclude that the multilevel Monte Carlo method gives us excellent gains over the Monte Carlo method, with up to two orders of magnitude gain theoretically and over one order of magnitude gain numerically, for both random fields. The discrepancy between the theory and numerics here arises because, for the two specific cases we consider, our error estimate is not sharp.

We note that we also observe substantial gains from the QMC methods. In particular, we observe almost a full order of magnitude gain numerically for the quasi-Monte Carlo estimator over the standard Monte Carlo estimator, and a similar gain for the multilevel quasi-Monte Carlo estimator over the standard multilevel Monte Carlo estimator (for Matérn field). This is despite using an "off the shelf" generating vector. It is left as future work to theoretically justify the improvement that the QMC rules bring, and to find an optimal generating vector for this problem.

---

[2]for problems where we consider random fields with smoothness parameter $\eta \geq 1$, we note that $\mathrm{C}_{pw}^{1-\delta} \subset C_{pw}^\eta$ (and similar). Hence, results such as Theorem 4.1.8 hold, when η is replaced by $1 - \delta$.
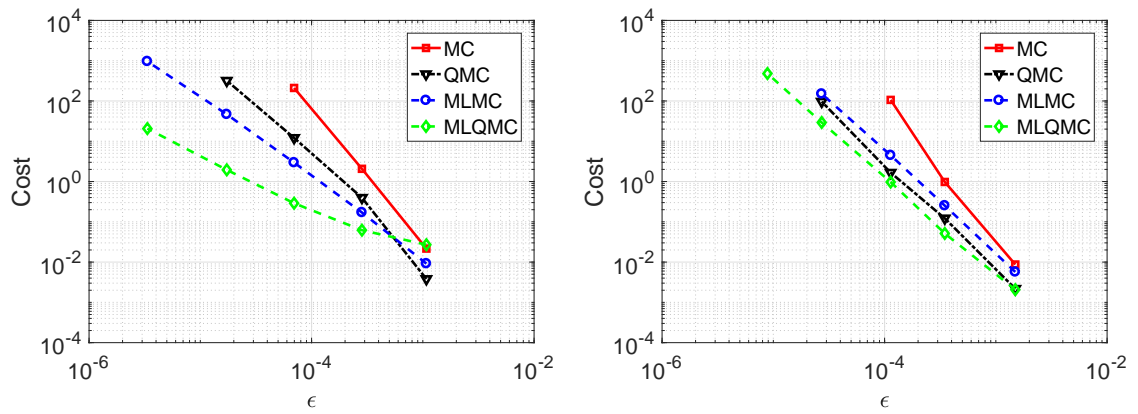
Figure 4-5: Cost (in seconds) plotted against ϵ (accuracy) on level L for standard, quasi, multilevel and multilevel quasi-Monte Carlo. (Left) Matérn field and (Right) exponential field. Details on reading this plots is given in Remark 4.3.1.

Chapter 5

Practical Numerical Tests

Contents

In this chapter we will apply the Monte Carlo techniques for Uncertainty Quantification (previously discussed in Chapter 2) to more physically relevant problems in radiative transport (than the mono-energetic 1D slab geometry fixed source problem considered in Chapter 3 – Chapter 4). In particular, we will consider the criticality problem (Section 1.1.3) in the 1D slab geometry, and we will also extend our previous 1D fixed source calculations to include problems in two spatial dimensions.

For the criticality problem, the random model will be based on the (deterministic and homogeneous) Los Alamos benchmark (two) presented in [190]. We introduce uncertainty by adding uniformly distributed random variable(s), representing measurement error for example, to the deterministic nuclear input data. To compute the eigenpair of interest, we propose an iterative eigensolver which uses a single source iteration within each iteration of a shifted inverse iteration.

We also consider two fixed source problems in a 2D spatial domain. The first problem is a simplified version of the C5-MOX problem presented in [43, 212]. The second problem mimics a concrete shielding problem, with a source of particles to the left side of the domain (e.g. particles leaving a reactor) and a detector on the right side of the domain. As part of the shielding problem, we have designed a novel model for nuclear cross-sections within heterogeneous concrete - which uses a realisation of a Gaussian random field, to ensure spatial correlation, combined with a sequence of maps which enforce distinct interfaces between different material types (and hence cross-sections) within the concrete.

For each of the aforementioned problems, we will begin by discussing the relevant radiative transport problem. We then define our chosen discretisation and present the methodology used to estimate a solution. Only then will we discuss the specific model of uncertainty considered. We finish by presenting numerical results which allow us to compare the efficiency of the variants of Monte Carlo sampling.

5.1 Criticality Problem in One Spatial Dimension

5.1.1 Model Problem

The criticality problem considered in this section comes from the *mono-energetic 1D slab geometry problem*, originally discussed in (1.3.6). We assume the cross-sections are isotropic in angle and the (spatial) domain contains a single spatially homogeneous material. Then, given some random input data

$$\mathbf{Z}(\omega, \cdot) = \mathbf{Z}(\omega) = [\sigma_S(\omega), \sigma_A(\omega), \sigma_F(\omega), \nu(\omega)] \ .$$

For each realisation of the input data, we are interested in the eigenvalue problem: Find the smallest eigenvalue $\lambda_{\mathrm{crit}}(\omega) \in \mathbb{R}^+$ and the angular flux $\psi(\omega, x, \mu) \not\equiv 0$ such that

$$\mu \frac{\partial \psi}{\partial x}(\omega, x, \mu) + \sigma(\omega)\psi(\omega, x, \mu) = [\sigma_S(\omega) + \lambda_{\mathrm{crit}}(\omega)\nu(\omega)\sigma_F(\omega)] \phi(\omega, x) \ , \tag{5.1.1}$$

$$\text{where} \quad \phi(\omega, x) = \frac{1}{2} \int_{-1}^1 \psi(\omega, x, \mu') \ d\mu' \ , \tag{5.1.2}$$

for $x \in [0, L_{\mathrm{(LA)}}]$, $\mu \in [-1, 1]$ and for almost all $\omega \in \Omega$, where ϕ denotes the scalar flux and $L_{\mathrm{(LA)}} \in \mathbb{R}^+$ denotes the width of the interval. The existence and uniqueness of λ_{crit} was previously discussed in Remark 1.1.2. We select the zero incoming flux boundary condition:

$$\psi(\omega, 0, \mu) = 0, \ \text{for} \ \mu > 0 \quad \text{and} \quad \psi(\omega, L_{\mathrm{(LA)}}, \mu) = 0, \ \text{for} \ \mu < 0 \ , \tag{5.1.3}$$

which hold for almost all $\omega \in \Omega$. We note that for the Los Alamos benchmark (two) [190] considered in this section, $L_{\mathrm{(LA)}} := 3.707444$ centimetres.

5.1.2 Discretisation

For each realisation $\omega \in \Omega$, (5.1.1) – (5.1.3) represents an eigenvalue problem in two independent variables, space and angle. For ease of presentation, let us suppress the dependence on ω for the moment.

We consider the same discretisation of (5.1.1) – (5.1.3) that was previously used for the spatially one-dimensional fixed source problem (see Chapter 3). That is, we discretise in angle using a double Gauss quadrature rule, with $2N$-points $\mu_k \in [-1, 1] \backslash \{0\}$ and the corresponding weights $w_k \in \mathbb{R}^+$. We discretise in space by using a Crank Nicolson scheme on a uniform mesh $0 = x_0 < x_1 < \ldots < x_M = L_{\mathrm{(LA)}}$. Subsequently, the discrete scheme for (5.1.1) – (5.1.3) is: Find the smallest $\lambda^{h,N} > 0$ and the family of (non-zero) continuous piecewise-linear functions $\{\psi_k^{h,N}\}_{k=1}^{2N}$ (with nodal values $\{\psi_{k,j}^{h,N}\}$) such that

$$\mu_k \frac{\psi_{k,j}^{h,N} - \psi_{k,j-1}^{h,N}}{h_j} + \sigma \frac{\psi_{k,j}^{h,N} + \psi_{k,j-1}^{h,N}}{2} = [\sigma_S + \lambda^{h,N}\nu\sigma_F] \phi_{j-1/2}^{h,N} \ , \tag{5.1.4}$$

for $j = 1, \cdots, M$, $|k| = 1, \ldots, N$, where

$$\phi^{h,N}_{j-1/2} \;=\; \frac{1}{2} \sum_{|k|=1}^{N} w_k \frac{\psi^{h,N}_{k,j} + \psi^{h,N}_{k,j-1}}{2} \;, \qquad \text{for } j = 1, ..., M \;, \qquad (5.1.5)$$

and with the no-inflow boundary conditions

$$\psi^{h,N}_{k,0} = 0, \quad \text{for} \quad k > 0 \quad \text{and} \quad \psi^{h,N}_{k,M} = 0, \quad \text{for} \quad k < 0 \;. \qquad (5.1.6)$$

5.1.3 Solution Methods for the Criticality Problem

The criticality problem, previously discussed in (1.1.16), is an example of a *generalised eigenvalue problem* [174, 178], so-called as it is a generalisation of a typical eigenvalue problem to the problem: Find (λ, x) satisfying $\mathfrak{A}x = \lambda\mathfrak{B}x$, for given operators $\mathfrak{A}$ and $\mathfrak{B}$.

In the context of (general) radiative transport before discretisation, the criticality problem seeks the smallest (real and positive) eigenvalue λ_{crit} (or the largest $k_{\text{eff}} := (1/\lambda_{\text{crit}}) \in \mathbb{R}^+$) and its corresponding eigenfunction $\psi \not\equiv 0$, such that

$$(\mathcal{T} - \mathcal{S})\,\psi \;=\; \lambda_{\text{crit}} \mathcal{F}\psi \;, \qquad (5.1.7)$$

where $\mathcal{T}$, $\mathcal{S}$ and $\mathcal{F}$ are defined for the full radiative transport equation in (1.1.8) – (1.1.10) respectively. The existence and uniqueness of the eigenpair was already discussed in Remark 1.1.2, for problems with homogeneous cross-sections and no-inflow boundary conditions.

There are a vast array of techniques available to find approximations of the eigenpair $(\lambda_{\text{crit}}, \psi)$ satisfying (5.1.7), we refer the reader to the books [15, 174, 201] for introductory material on some of these methodologies. Broadly speaking the techniques can be separated into two categories; direct eigensolvers, and iterative eigensolvers.

Direct eigensolvers, such as the the QR algorithm - which involves computing a full QR factorisation of an appropriate matrix - have the advantage of being able to compute the entire spectrum at once, but they can be prohibitively expensive and in our problem we are only interested in the eigenpair corresponding to the *smallest* positive real eigenvalue.

In comparison, iterative eigensolvers typically find only a subset of the spectrum. They consist of an *(outer) iteration*, where for each iteration we must solve a fixed source problem. If we choose an iterative solver, to estimate the solution of each fixed source problem, then we also have an *inner iteration*. Such outer-inner iteration combinations for eigenvalue problems are known as *inexact iterative methods* (assuming that the iterative solver does not compute the solution to the inner source problem exactly). There is an extensive literature on the theory of inexact iterative methods outside of radiative transport, see for example [71, 88, 126] and related references, but only recent literature within radiative transport, e.g. [178].

Below we will discuss an iterative eigensolver for computing the eigenpair in (5.1.7) - the power iteration and its variants. Whilst this is a popular deterministic method for finding eigenpairs in radiative transport, there are also many other possibilities to choose from. For example, [206] detail the application of a type of Krylov subspace method, and stochastic methods e.g. MONK® (developed by the ANSWERS® software service) are commonplace in industry.

Remark 5.1.1 *The statements made in the proceeding discussion will be general statements about the power iteration method (and its variants). However, we illustrate the details of the power iteration for the (specific) transport criticality problem (5.1.7) and in this case, simplifications to*

the aforementioned statements can be made. For example, the statement 'the smallest eigenvalue in absolute value' - regarding (inverse) power iteration - is equivalent to the statement 'the smallest real and positive eigenvalue' for the transport problem (5.1.7), because (5.1.7) only has positive real eigenvalues (see Remark 1.1.2).

5.1.4   (Inverse) Power Iteration

One of the simplest and most well known methods for computing eigenpairs is the so-called *power iteration*, which aims to find the *largest eigenvalue* (in absolute value) and its corresponding eigenfunction satisfying a given eigenproblem. However, we are interested in the *smallest eigenvalue* (in absolute value). This leads us to one popular method known as the *Inverse Power iteration* [15, 178, 206]. As the name suggests it is closely related to the power iteration - in fact, it is a power iteration but the eigenproblem is re-written. To see why this gives us the smallest eigenvalue (in absolute value), re-write (5.1.7) as

$$\frac{1}{\lambda_{\text{crit}}} \psi \;=\; (\mathcal{T} - \mathcal{S})^{-1} \mathcal{F} \psi \;, \tag{5.1.8}$$

and then apply the standard power iteration to this eigenproblem, i.e. iteratively apply $(\mathcal{T} - \mathcal{S})^{-1} \mathcal{F}$ to the previous estimate (and normalise):

$$\psi^{(n+1)} \;=\; c_n \, (\mathcal{T} - \mathcal{S})^{-1} \mathcal{F} \psi^{(n)} \;, \quad \text{for} \;\; n = 0, 1, \cdots \;, \tag{5.1.9}$$

where c_n denotes a normalisation constant, e.g. ensuring $\|\psi^{(n+1)}\| = 1$ in a chosen norm. Under appropriate conditions, this gives us the largest $k_{\text{eff}} = (1/\lambda_{\text{crit}})$ (in absolute value), and the corresponding eigenfunction, satisfying (5.1.8). This is equivalent to acquiring the smallest (in absolute value) eigenvalue λ_{crit}, and the corresponding eigenfunction, satisfying (5.1.7). Note that for each n, finding the next estimate $\psi^{(n+1)}$ in (5.1.9) is equivalent to solving a fixed source problem, with fixed source $\mathcal{F} \psi^{(n)}$ (and then normalising).

After each iteration of (5.1.9), λ_{crit} can be estimated by the (generalised) Rayleigh quotient on $L_2(\mathcal{D} \times \mathcal{A})$, defined by

$$\lambda^{(n)} \;:=\; \frac{\langle \psi^{(n)}, (\mathcal{T} - \mathcal{S}) \, \psi^{(n)} \rangle_{L_2(\mathcal{D} \times \mathcal{A})}}{\langle \psi^{(n)}, \mathcal{F} \psi^{(n)} \rangle_{L_2(\mathcal{D} \times \mathcal{A})}} \;, \tag{5.1.10}$$

where $\langle \cdot, \cdot \rangle_{L_2(\mathcal{D} \times \mathcal{A})}$ denotes the standard L_2-inner product over space ($\mathcal{D}$) and angle ($\mathcal{A}$). To get some intuition on where (5.1.10) comes from, consider (5.1.7), multiply it by ψ and integrate over the domain of $\mathcal{D} \times \mathcal{A}$ and then re-arrange:

$$\int_{\mathcal{D} \times \mathcal{A}} \psi \, (\mathcal{T} - \mathcal{S}) \, \psi \;=\; \int_{\mathcal{D} \times \mathcal{A}} \lambda \psi \mathcal{F} \psi \quad \Longleftrightarrow \quad \lambda \;=\; \frac{\langle \psi, (\mathcal{T} - \mathcal{S}) \, \psi \rangle_{L_2(\mathcal{D} \times \mathcal{A})}}{\langle \psi, \mathcal{F} \psi \rangle_{L_2(\mathcal{D} \times \mathcal{A})}} \;,$$

assuming the denominator does not vanish, and where we note that for two functions $f, g \in L_2(X)$, for some space X, then $\langle f, g \rangle_X := \int_X f \, g$.

The (inverse) power iteration (5.1.9) combined with (5.1.10) is sometimes called the *Rayleigh quotient iteration*.

Now consider the discretised version of (5.1.7), i.e. when the operators $\mathcal{T}$, $\mathcal{S}$ and $\mathcal{F}$ are approximated by matrices (T, S and F respectively) and the eigenfunction ψ is approximated by a vector of coefficients $\boldsymbol{\Psi}$ (applied to some set of basis functions). Then provided that: (i) the smallest eigenvalue (in absolute value) is simple; (ii) the initial guess $\boldsymbol{\Psi}^{(0)}$ (again applied to a set of basis functions) is not orthogonal to the direction of the eigenfunction; it can be shown that the

inverse power iteration converges to the smallest eigenvalue (in absolute value) [174]. Moreover, it is well known (see [174, Theorem 4.1]) that the convergence of the inverse power iteration depends on the so-called *dominance ratio*, $|\lambda_1/\lambda_2|$, where λ_1, λ_2 denote the smallest and second smallest eigenvalues (in absolute value) respectively.

The problem, in the context of radiative transport, is that for many real-life reactors the dominance ratio is close to 1 [158, 206], i.e. the inverse power iteration exhibits poor convergence. In such cases, the paper [206] observed fewer iterations for its Krylov subspace method. Below we will discuss a standard method for accelerating the convergence of the inverse power iteration, known as *shifted inverse iteration*.

Shifted Inverse Iteration

Consider a scalar-valued shift $\rho \in \mathbb{R}$, chosen such that

$$|\lambda_1 - \rho| < |\lambda_2 - \rho| \leq |\lambda_3 - \rho| \leq \cdots , \qquad (5.1.11)$$

where we assume $\lambda_1 \neq \lambda_2$ denote the first and second smallest eigenvalues (in absolute value) for the non-discrete problem (5.1.7). In the context of the transport problem (5.1.7), $\lambda_1 = \lambda_{\mathrm{crit}}$. Then, the *shifted inverse iteration* is simply the inverse power iteration applied to the following eigenproblem

$$(\mathcal{T} - \mathcal{S} - \rho\mathcal{F})\,\psi \;=\; (\lambda_{\mathrm{crit}} - \rho)\,\mathcal{F}\psi , \qquad (5.1.12)$$

i.e. (5.1.7) with $\rho\mathcal{F}\psi$ subtracted from both sides (inverse power iteration and shifted inverse iteration are equivalent when $\rho \equiv 0$). Subsequently, we know the convergence of the shifted inverse iteration will depend on the (shifted) dominance ratio, i.e.

$$\frac{|\lambda_1 - \rho|}{|\lambda_2 - \rho|} \ .$$

Therefore, if we choose $\rho \approx \lambda_1$ (and closer to λ_1 than λ_2) and the assumption (5.1.11) holds, then the convergence of the shifted inverse iteration can be faster than standard inverse iteration.

For transport problems within reactor physics, acquiring a good choice for ρ is often simple. This is because most reactors are designed to ensure that $\lambda_{\mathrm{crit}} = 1$ (or at least close to 1, with some safety margin) and hence choosing a fixed shift $\rho = 1$ is often a good starting guess. The shift ρ does not have to be fixed for all (outer) iterations, and can be updated at each iteration (e.g. see ahead to the Rayleigh quotient shift (5.1.17)).

The shifted inverse iteration will be an important method for our computations and hence we present an algorithmic representation of it in Algorithm 2. In this representation, we have assumed the use of the source iteration method (see (3.1.13)) for the inner source problem, but the method can easily be adapted to use another iterative, or a direct, solver. There are five main steps to this algorithm for shifted inverse iteration: (i) a stopping criterion for the outer iteration; (ii) a stopping criterion for the inner iteration; (iii) a choice of normalisation; (iv) an estimate of the eigenvalue; (v) an update of the shift. There are many possibilities one may choose here, however in Algorithm 2 we present the following suggestions (and in our later numerical computations we will use the discretised version of Algorithm 2).

We define the outer stopping criterion as: For any $n \in \mathbb{N}$, find $\lambda^{(n)}$ and $\psi^{(n)}$ such that

$$\mathbf{res}_{\mathrm{out}}^{(n)} \;=\; \left\| \left(\mathcal{T} - \mathcal{S} - \lambda^{(n)}\mathcal{F}\right)\psi^{(n)} \right\|_I \;\leq\; \epsilon_{\mathrm{out}} , \qquad (5.1.13)$$

where we introduce the norm[1] $\| \cdot \|_I := \| \mathcal{W} \mathcal{T}^{-1} \cdot \|_{L_2(\mathcal{D})}$ and note that the operator $\mathcal{W}$ is defined in (1.1.4). At the $(n-1)$th outer iteration, with $\rho^{(n-1)}$ and $\mathcal{F}\psi^{(n-1)}$ given and fixed, we can define the inner stopping criterion as (however, in our computations we will follow a simpler approach, as outlined later in Section 5.1.5): For any $k \in \mathbb{N}$, find $\tilde{\psi}^{(n-1,k)}$ such that

$$\mathbf{res}_{\text{inn}}^{(n-1,k)} := \left\| \left(\mathcal{T} - \mathcal{S} - \rho^{(n-1)} \mathcal{F} \right) \tilde{\psi}^{(n-1,k)} - \mathcal{F}\psi^{(n-1)} \right\|_I \leq \epsilon_{\text{in}} , \qquad (5.1.14)$$

where $\tilde{\psi}^{(\cdot,k)}$ denotes the approximation to the angular flux ψ (given a fixed source $\mathcal{F}\psi^{\cdot}$) at the kth iteration of the inner solver (i.e. source iteration). Once k is sufficiently large such that (5.1.14) is satisfied, then we define $\tilde{\psi}^{(n)} := \tilde{\psi}^{(n-1,k)}$. Subsequently, we acquire the new eigenfunction estimate $\psi^{(n)}$, by normalising $\tilde{\psi}^{(n)}$:

$$\psi^{(n)} := \frac{\tilde{\psi}^{(n)}}{\| \mathcal{W} \tilde{\psi}^{(n)} \|_{L_2(\mathcal{D})}} , \qquad (5.1.15)$$

where we assume that the denominator is non-zero (we have observed this in our computations). For the eigenvalue update, we use the (generalised) Rayleigh quotient, i.e.

$$\lambda^{(n)} := \frac{\langle \psi^{(n)}, (\mathcal{T} - \mathcal{S}) \psi^{(n)} \rangle_{L_2(\mathcal{D} \times \mathcal{A})}}{\langle \psi^{(n)}, \mathcal{F}\psi^{(n)} \rangle_{L_2(\mathcal{D} \times \mathcal{A})}} , \qquad (5.1.16)$$

where $\langle \cdot, \cdot \rangle_{L_2(\mathcal{D} \times \mathcal{A})}$ denotes the L_2-inner product over space ($\mathcal{D}$) and angle ($\mathcal{A}$). Finally, we choose the Rayleigh quotient shift, i.e.

$$\rho^{(n)} := \lambda^{(n)} . \qquad (5.1.17)$$

Another option for the shift would be the 'non-standard Rayleigh quotient shift' given in [178], which is proven to have better convergence properties than the standard Rayleigh quotient shift (5.1.17). We do not consider this here.

---

[1] the $\| \cdot \|_I$ norm links the non self-adjoint eigenvalue problem (5.1.7), with an underlying self-adjoint eigenvalue problem (related to the integral equation form of the RTE), see [178] for details

---

Algorithm 2: Shifted Inverse Iteration with inner source iteration solve

---

    Data: Initial estimate of the eigenfunction $\psi^{(0)}$ and an initial shift $\rho^{(0)}$;

    Data: Desired accuracies for the outer iteration, ϵ_{out}, and the inner iteration, ϵ_{in}.

    Result: Approximation of the eigenpair: $\lambda_{\text{crit}} \approx \lambda^{(n+1)}$ and $\psi \approx \psi^{(n+1)}$;

1 Initialise $n = 0$;

2 **while** $\boldsymbol{res}_{out}^{(n)} > \epsilon_{out}$ **do**

3   Compute $\mathcal{F}\psi^{(n)}$ (in some sense, the new fixed source);

4   Initialise $k = 0$ and $\tilde{\psi}^{(n,-1)} \equiv 0$;

5   **while** $\boldsymbol{res}_{inn}^{(n,k)} > \epsilon_{in}$ **do**

6   Compute $\tilde{\psi}^{(n,k)}$ such that

$$\tilde{\psi}^{(n,k)} \;=\; \mathcal{T}^{-1}\left(\sigma_S \mathcal{W}\tilde{\psi}^{(n,k-1)} \;+\; \mathcal{F}\psi^{(n)}\right)$$

        Update the residual $\mathbf{res}_{\text{inn}}^{(n,k+1)}$ according to (5.1.14);

7 $k = k + 1$;

8   **end**

9   Set $\tilde{\psi}^{(n+1)} := \tilde{\psi}^{(n,k-1)}$;

10   Acquire $\psi^{(n+1)}$ by normalising $\tilde{\psi}^{(n+1)}$ according to (5.1.15);

11   Update the eigenvalue estimate $\lambda^{(n+1)}$ according to (5.1.16);

12   Update the shift $\rho^{(n+1)}$ according to (5.1.17);

13   Update the residual $\mathbf{res}_{\text{out}}^{(n+1)}$ according to (5.1.13);

14 $n = n + 1$;

15 **end**

5.1.5 One Inner Iteration

The shifted inverse iteration can lead to substantial gains over the standard inverse iteration [158, 178, 206] - but it is not without its issues. The solution ($\psi^{(n)}$ at the nth outer iteration) to the inner fixed source problem

$$\left(\mathcal{T} - \mathcal{S} - \rho^{(n)}\mathcal{F}\right)\psi^{(n)} \;=\; \left(\lambda_{\text{crit}} - \rho^{(n)}\right)\mathcal{F}\psi^{(n-1)}\,, \tag{5.1.18}$$

can be approximated by solving the linear system associated with the matrix discretisation of (5.1.18). However, particularly when choosing the Rayleigh quotient shift $\rho^{(n)}$, defined in (5.1.17), the matrix discretisation of $\left(\mathcal{T} - \mathcal{S} - \rho^{(n)}\mathcal{F}\right)$ becomes nearly singular (for large n), and iterative methods (such as source iteration) may struggle to converge (leading to a large, or potentially infinite, number of required inner iterations to accurately estimate the solution to the inner fixed source problem)[2].

We now suggest an alternative that seeks to avoid this issue. Motivated by the analysis in [178, §3.2], we see that finding even a very rough solution of the inner problem can lead the inexact inverse iteration to converge - where it is shown that quadratic convergence of Algorithm 2 can be achieved (under a fairly weak condition on ϵ_{in}). We omit any further details as it is beyond the scope of thesis and refer the interested reader to [178]. Moreover, we find by experiment that solving the (approximate) inner fixed source problem by *one source iteration* (in line 6 of Algorithm

---

[2]provided the matrix isn't singular, this actually aids the outer iteration. This is because, and quoting [15, pg.629], "any resulting large perturbations in the solution will be rich in the eigenvector [for the matrix discretisation of (5.1.18)]"

2) gives a reliable overall convergent approximate inverse iteration method for the problems we have considered. For the remainder of this thesis this method is referred to as the *one inner iteration* method.

Remark 5.1.2 *The (quadratic) convergence result mentioned above is relevant when the shift in Algorithm 2 is taken to be the non-standard Rayleigh quotient shift defined in [178, eq.(3.23)]. However, a similar result was also proven ([178, Corollary 3.6]) which shows linear convergence of Algorithm 2 when a fixed shift is taken (but with a stronger condition on ϵ_{in}). Moreover, the "convergence analysis ... is given only for the continuous problem", i.e. (5.1.7). Nonetheless "it provides a guide to how iterations behave in discrete cases" [178].*

The use of 'single inner solves' appears in a number of other contexts. For example, the `CACTUS` module within the `WIMS` code (developed by the ANSWERS® team) and the specific method of perturbation technique discussed in [178], employ a similar strategy.

It is left for future work to analyse and compare the *one inner iteration* to the standard shifted inverse iteration, and other eigensolvers. From our (limited) numerical results, the one inner iteration method exhibits the same linear[3] convergence (of a sequence, see Remark 5.1.3 for details) of $\lambda^{(n)}$ (with respect to iterations, see ahead to Figure 5-2) that is observed in [178] for the standard inverse iteration. Moreover, in our numerical experiments the cost of the one inner iteration is smaller than that of the shifted inverse iteration because, for the shifted inverse iteration, the number of source iterations used to estimate the solution of the *final* (inner) source problem reaches the maximum number of iterations that we allow in our code.

Remark 5.1.3 *A sequence $\lambda^{(1)}$, $\lambda^{(2)}$, $\cdots$ is said to converge with order r if*

$$\lim_{n \to \infty} \frac{|\lambda_{crit} - \lambda^{(n+1)}|}{|\lambda_{crit} - \lambda^{(n)}|^r} \leq c , \tag{5.1.19}$$

for some constant $c > 0$. Linear convergence is a special case when $r = 1$, $c \in (0,1)$ and the in-equality is replaced by equality. Practically, we can estimate r by using the following approximation

$$r = \frac{\log\left(\left|\frac{\lambda^{(n+1)} - \lambda^{(n)}}{\lambda^{(n)} - \lambda^{(n-1)}}\right|\right)}{\log\left(\left|\frac{\lambda^{(n)} - \lambda^{(n-1)}}{\lambda^{(n-1)} - \lambda^{(n-2)}}\right|\right)} . \tag{5.1.20}$$

In the specific case below, i.e. Figure 5-2, the approximation (5.1.20) is not necessary as it is easy to see that we have linear convergence. This is because for (almost) all iterations considered the line is straight and therefore the ratio between $|\lambda^{(n+1)} - \lambda^{(n)}|$ and $|\lambda^{(n)} - \lambda^{(n-1)}|$ must remain fixed for (almost) all n (the c in (5.1.19) is also approximated by the gradient of this line).

We will now discuss the random model, which will be used for our numerical results below.

5.1.6 Random Model: Uniform Los Alamos

We now consider a random variant of one of the benchmarks given in [190], and references therein. Before we introduce the uncertainty, let us first give details of the deterministic problem.

The deterministic problem is given by the mono-energetic 1D slab geometry problem equipped with no-inflow boundary conditions and defined on the 1D spatial interval $[0, L_{(LA)}]$, with $L_{(LA)} :=$ 3.70744 (in centimetres). The problem assumes that the cross-sections are isotropic in angle and

---

[3]theoretically we expect quadratic convergence. However because Crank Nicolson is not a symmetry preserving scheme, the convergence is reduced. See [178] for further details

$L_{\text{(LA)}}$	σ_A	σ_S	σ_F	ν	σ
3.707444	0.019584	0.225216	0.0816	3.24	0.3264

Table 5.1: The input data for problem 2 in the Los Alamos benchmark set [190].

(a) σ_S (b) σ_F
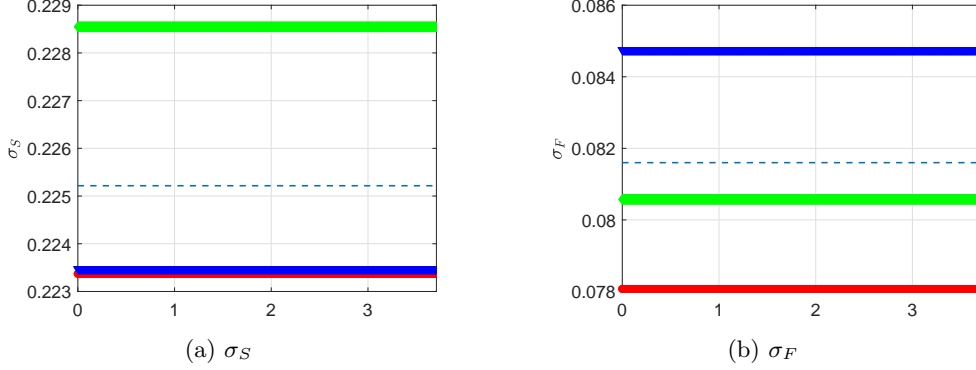
Figure 5-1: Three realisations of the scattering (left) and fission (right) cross-sections, under the random variation in (5.1.21). The colours green, blue and red denote realisations 1, 2 and 3 respectively. The corresponding cross-sections for the deterministic problem are given by the dashed blue lines.

that the spatial domain contains a single homogeneous material. The nuclear input data for the homogeneous material is given in Table 5.1. For this particular benchmark problem, it is known (analytically) that the smallest positive real eigenvalue is $\lambda_{\text{crit}} = 1$.

We will now introduce uncertainty into the nuclear input data of the aforementioned problem, with details given in Problem 5.1.4. Subsequently, we present realisations of the scattering and fission cross-section(s) in Figure 5-1.

Problem 5.1.4 *Let $z_1, \cdots, z_4 \sim \mathcal{U}(1 - \epsilon, 1 + \epsilon)$ denote uniform random variables, with $\epsilon := 0.05$. Then, the (random) uniform Los Alamos problem is defined by (5.1.1) – (5.1.3) with (random) nuclear input data given by*

$$\sigma_A = z_1 \sigma_A^{(det)}, \quad \sigma_S = z_2 \sigma_S^{(det)}, \quad \sigma_F = z_3 \sigma_F^{(det)}, \quad \nu = z_4 \nu^{(det)}. \tag{5.1.21}$$

Here $\sigma_A^{(det)}, \sigma_S^{(det)}, \sigma_F^{(det)}, \nu^{(det)}$ denote the deterministic values (given in Table 5.1) for σ_A, σ_S, σ_F and ν respectively.

We will be interested in estimating $\mathbb{E}[Q]$, where our quantity of interest is

$$Q(\omega) := \lambda_{\text{crit}}(\omega), \tag{5.1.22}$$

for $\omega \in \Omega$, i.e. the smallest (positive real) eigenvalue for Problem 5.1.4.

Numerical Results

We will now present a number of numerical results relating to Problem 5.1.4. We consider the discretisation outlined in Section 5.1.2, with a uniform spatial mesh of width h and the double Gauss quadrature rule with $2N$ angles, chosen such that $N = N(h) := (2h)^{-1}$. The eigenpair is estimated using the one inner iteration method, introduced in Section 5.1.5, where we select the (outer) stopping criterion, flux normalisation, eigenvalue estimate and shift according to discrete approximations of (5.1.13), (5.1.15), (5.1.16) and (5.1.17) respectively. To estimate the solution

(a) $|\lambda_{\text{crit}} - \lambda^{(n)}|$

(b) $\left\| \left(T - S - \lambda^{(n)} F \right) \psi^{(n)} \right\|_I$

Figure 5-2: Two error metrics showing the convergence for each outer iteration, when using the one inner iteration method (see Section 5.1.5) for the inner source problem. The $\lambda^{(n)}$ denotes the approximation to λ_{crit} at the nth outer iteration. We show the convergence at three different resolutions $M = h^{-1} = 32, 64, 128$ (green, red and blue respectively).
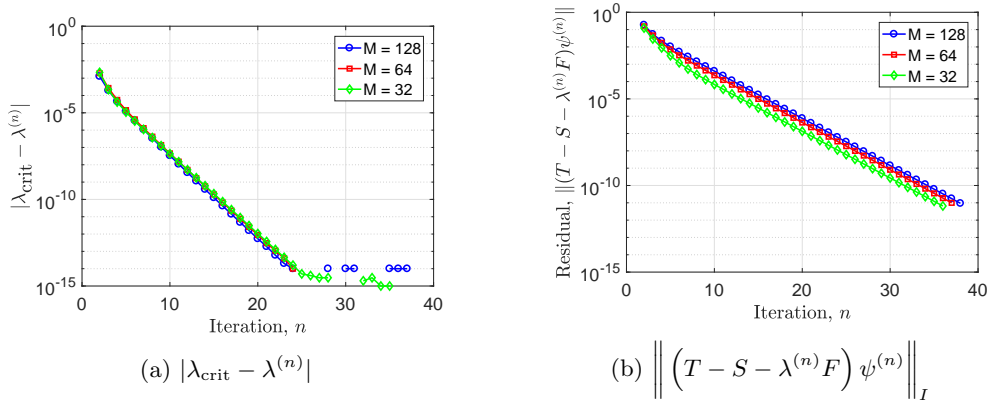
	α_{obs}	β_{obs}	γ_{obs}	Estimated	Observed
Monte Carlo				2.9	2.9
Multilevel	2.0	3.7	1.8	2.0	2.0

Table 5.2: Summary of computational ϵ-cost rates r, where for an estimator $\widehat{Q}$, $\mathbb{E}[\mathcal{C}(\widehat{Q})] = \mathcal{O}(\epsilon^{-r})$. We estimate r in the following ways: 'Estimated' uses the numerically observed α_{obs}, β_{obs}, γ_{obs}; 'Observed' uses the observed rates from Figure 5-4.

to the inner fixed source problem, i.e. the discrete version of (5.1.18), we use the source iteration method previously discussed in (3.1.13) and outlined in Algorithm 2.

We introduce a hierarchy of levels $\ell = 0, \cdots, L$ corresponding to a sequence of discretisation parameters $h_\ell = 2^{-\ell} h_0$ with $h_0 = 1/4$, and approximate the quantity of interest (5.1.22) by

$$Q_\ell = Q_{h_\ell} = Q_{h_\ell, N(h_\ell)} = \lambda^{h_\ell, N(h_\ell)} = \lambda^\ell .$$

We compare $\{Q_\ell\}$ to a reference solution calculated using the standard inverse iteration with $h^{-1} = 256$ and $N = 128$. We will estimate the error by computing $\mathbb{E}[|\lambda_{\text{crit}} - \lambda^\ell|]$, where the expectation is estimated using a standard Monte Carlo estimator, (2.4.8), with $1,024$ samples.

We first study the convergence of the one inner iteration method (Section 5.1.5), with respect to the number of outer iterations. The results are presented in Figure 5-2. We observe linear convergence for both the eigenvalue estimate (at least up to machine precision) and the residual (the discrete approximation to (5.1.13)), for a variety of mesh parameters.

To study the efficiency of the Monte Carlo methods, and their multilevel variants, and justify the underlyling assumptions (2.4.3), (2.4.4) and (2.4.33), we numerically estimate the parameters α, γ and β in those assumptions. In Figure 5-3 we present numerical results studying the behaviour of the bias error and cost of the eigensolver - as the mesh is refined. The convergence of $\mathbb{E}\left[|\lambda_{\text{crit}} - \lambda^\ell|\right]$ is approximately[4] $\mathcal{O}\left(h_\ell^2\right)$, and hence we estimate $\alpha \approx 2$. Likewise, we estimate $\gamma \approx 1.8$. Moreover, looking ahead to the left plot in Figure 5-4, we observe $\beta \approx 3.7$. These rates are summarised in Table 5.2.

Finally, on the right hand side of Figure 5-4 we present numerical results comparing the computational ϵ-cost of the standard and multilevel Monte Carlo methods, when estimating $\mathbb{E}[\lambda_{\text{crit}}]$.

---

[4]Note that this is twice the rate exhibited by the residual, i.e. in Figure 5-2(b). Moreover, this is consistent with what we would expect for a self-adjoint eigenvalue problem (and hence justifies the use of the $\| \cdot \|_I$ norm earlier)

(a) $\mathbb{E}\left[|\lambda_{\mathrm{crit}} - \lambda^{\ell}|\right]$

(b) $\mathbb{E}\left[\|\phi - \phi^{h_{\ell}, N(h_{\ell})}\|_{\infty}\right]$

(c) $\mathbb{E}\left[\mathcal{C}(Q_{\ell})\right]$
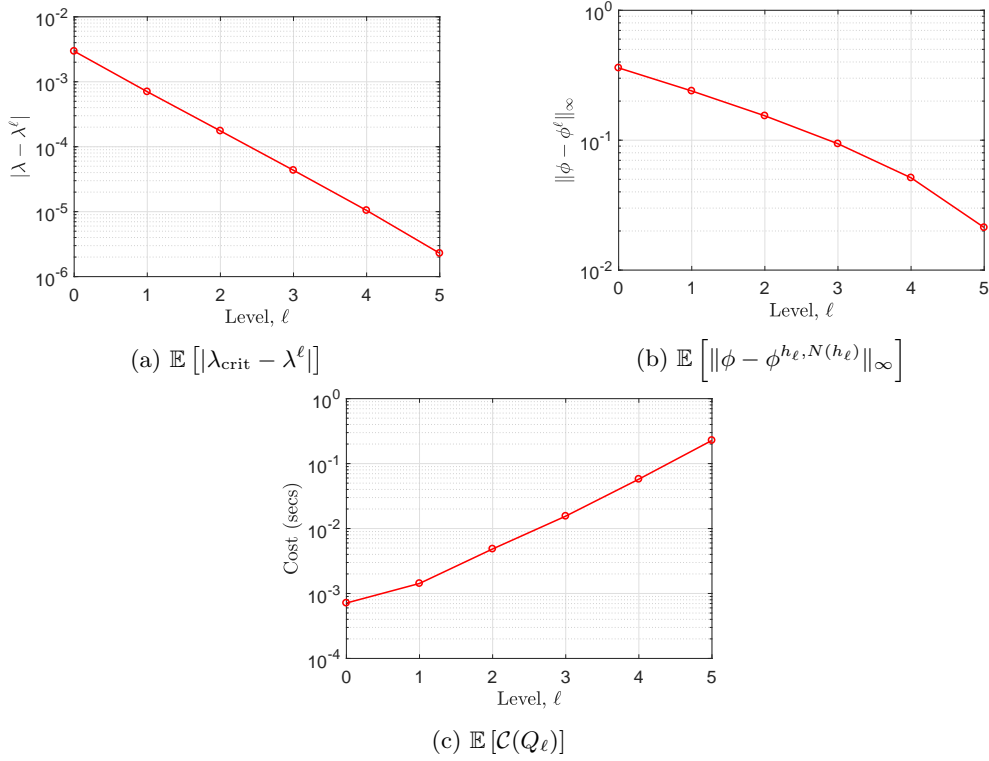
Figure 5-3: Picture (a) (respectively (b)) represents an estimate of the bias error $\mathbb{E}\left[|\lambda_{\mathrm{crit}} - \lambda^{\ell}|\right]$ (respectively $\mathbb{E}\left[\|\phi - \phi^{h_{\ell}, N(h_{\ell})}\|_{\infty}\right]$) of the one inner iteration method, as the mesh is refined. Picture (c) is an estimate of the corresponding average cost in seconds to compute the eigenpair.

In Table 5.2 we compare the rates of; the numerically estimated ϵ-cost, using the numerically observed parameters α_{obs}, β_{obs}, γ_{obs} (presented in Table 5.2) within (2.4.13) and Theorem 2.4.4, and the numerically observed ϵ-cost, estimated from the plot on the right hand side of Figure 5-4. We note very good agreement between the 'estimated' and 'observed' computational cost rates.

We conclude that for the given criticality problem, in one spatial dimension and one angular dimension (with the given discretisation), the multilevel Monte Carlo method gives us good gains over the Monte Carlo method - with an order of magnitude improvement in the observed computational ϵ-cost.

We will now turn our attention towards radiative transport (fixed source) problems in two spatial dimensions.
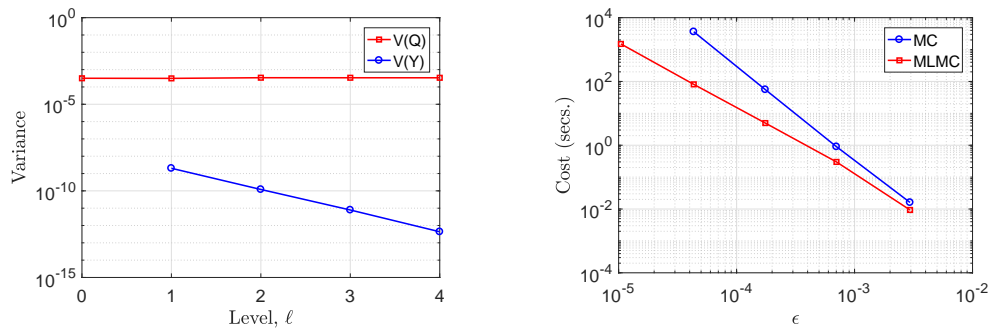
Figure 5-4: (Left) Estimate of $\mathbb{V}[Q_{\ell}]$ and $\mathbb{V}[Y_{\ell}]$; (Right) Actual cost (in seconds) of the standard and multilevel Monte Carlo methods, plotted against the achieved root-MSE accuracy. Details of the plot on the right are given in (4.3.1).

5.2 Fixed Source Problem in Two Spatial Dimensions

5.2.1 Model Problem

For the fixed source problem(s) in this section we will consider the *mono-energetic 2D-1D* problem in radiative transport, originally discussed in (1.3.5). This is an extension of the fixed source problem in Chapter 3, to a problem with a two-dimensional spatial domain and the unit circle as its angular domain. We assume that the cross-sections and source are isotropic in angle, and that there is no fission. Hence, given some random input data

$$\mathbf{Z}(\omega, \cdot) = [\sigma_S(\omega, \cdot), \sigma_A(\omega, \cdot), f(\omega, \cdot)] ,$$

we are interested in the fixed source problem: Find $\psi(\omega, \mathbf{r}, \boldsymbol{\Theta})$ such that

$$[\boldsymbol{\Theta} \cdot \nabla + \sigma(\omega, \mathbf{r})] \psi(\omega, \mathbf{r}, \boldsymbol{\Theta}) = \sigma_S(\omega, \mathbf{r})\phi(\omega, \mathbf{r}) + f(\omega, \mathbf{r}) , \qquad (5.2.1)$$

$$\text{where} \quad \phi(\omega, \mathbf{r}) = \frac{1}{2\pi} \int_{\mathbb{S}_1} \psi(\omega, \mathbf{r}, \boldsymbol{\Theta}') \, d\boldsymbol{\Theta}' , \qquad (5.2.2)$$

for any $\mathbf{r} \in \mathcal{D} = [0,1]^2$, $\boldsymbol{\Theta} \in \mathbb{S}_1 = \{v \in \mathbb{R}^2 \mid |v| = 1\}$ (the unit circle) and for almost all random realisations ω in the sample space Ω. The boundary conditions are chosen problem-specific and are given below.

We note that we previously discussed the existence and uniqueness of a solution to (5.2.1) – (5.2.2) in Remark 1.1.3.

5.2.2 Discretisation

Similarly to Section 5.1.2, we note that (5.2.1)-(5.2.2) is an integro-differential equation which is two-dimensional in space and one-dimensional in angle. Hence for ease of presentation, we suppress the dependence on ω for the moment.

We will discretise in angle by using the discrete ordinates method (see Section 1.2.2), i.e. we consider the solution of the transport equation along a finite set of N angles $\{\boldsymbol{\Theta}_k\}_{k=1}^N$ on the unit circle. We choose the angles to be uniformly spaced and define them by

$$\boldsymbol{\Theta}_k = \left(\cos \frac{2\pi k}{N}, \ \sin \frac{2\pi k}{N} \right)^T ,$$

with the corresponding quadrature weight(s) $w_k = (2\pi/N)$, for all $k = 1, \cdots, N$.

In space we use the Discontinuous Galerkin finite element method, previously detailed in Section 1.2.3. We discretise the unit square $[0,1]^2$ by the tensor product of the 1D uniform meshes $0 = x_0 < x_1 < \cdots < x_{M_x} = 1$ and $0 = y_0 < \cdots < y_{M_y} = 1$, i.e. $x_i = iM_x^{-1}$ and $y_j = jM_y^{-1}$, for $i = 0, \cdots, M_x$ and $j = 0, \cdots, M_y$.

Let $\{\mathcal{C}^h\}$ denote a family of (disjoint and open) rectangles D^h (which are parallel to the axes) such that $\overline{\mathcal{D}} = \bigcup_{D^h \in \mathcal{C}^h} \overline{D^h}$, where h denotes the maximum diameter of any D^h. Also, define the solution space

$$V^h := \{v \in L_2([0,1]^2) \mid \forall D^h \in \mathcal{C}^h, \ v|_{D^h} \in \mathcal{Q}_1(D^h)\} ,$$

where we recall $\mathcal{Q}_1(\cdot)$ denotes the set of polynomials of separate degree 1, and a set of basis functions (defined in (1.2.14) for 3D, with an analogous definition in 2D) $\{v_j(\mathbf{r}) \in V^h \mid j = 1, \cdots, M_f\}$ that span the space V^h, where $M_f := 4M_x M_y$ denotes the number of spatial degrees of freedom.

Furthermore, assume that the input data is *piecewise constant* on each element D^h, and use σ_{D^h} and $(\sigma_S)_{D^h}$ to denote the value of the (total and scattering) cross-sections on a given $D^h \in \mathcal{C}^h$.

Finally, for the numerical flux $\mathbf{F}_k$ (recall (1.2.15)) we choose the so-called *upwind numerical flux* [35, 96], i.e.

$$\mathbf{F}_k(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) := (\mathbf{\Theta}_k \cdot \mathbf{n}(\mathbf{r}))\, \psi_k^{h,N,+}(\mathbf{r}) \;, \quad \text{for all } \mathbf{r} \in \partial D^h_- \;, \tag{5.2.3}$$

for each $k = 1, \cdots, N$, where we define

$$\psi_k^{h,N,+}(\mathbf{r}) := \lim_{\epsilon \to 0} \psi_k^{h,N}(\mathbf{r} + \epsilon \mathbf{\Theta}_k) \;, \quad \text{for all } \mathbf{r} \in \partial D^h_- \;, \tag{5.2.4}$$

and the inflow boundary at each cell D^h as D^h_-, i.e. (recall (1.2.17))

$$\partial D^h_- := \{\mathbf{r} \in \partial D^h \mid \mathbf{\Theta}(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) < 0\} \;.$$

Hence, the discretised version of the weak form of (5.2.1) is: Find $\psi_k^{h,N} \in V^h$ such that for all $D^h \in \mathcal{C}^h$ and for all $v_j \in \mathcal{Q}_1(D^h)$ with support on D^h,

$$-\int_{D^h} (\mathbf{\Theta}_k \cdot \nabla v_j)\psi_k^{h,N}\, d\mathbf{r} + \sigma_{D^h} \int_{D^h} \psi_k^{h,N} v_j\, d\mathbf{r} + \int_{\partial D^h_-} \mathbf{F}_k \cdot \mathbf{n} v_j\, d\mathbf{r}$$
$$= (\sigma_S)_{D^h} \int_{D^h} \phi^{h,N} v_j\, d\mathbf{r} + \int_{D^h} f v_j\, d\mathbf{r} \tag{5.2.5}$$

for all $k = 1, \cdots, N$, where (5.2.2) is approximated by

$$\phi^{h,N} := \frac{1}{2\pi} \sum_{k'=1}^{N} w_{k'} \psi_{k'}^{h,N} \;. \tag{5.2.6}$$

The problem specific boundary conditions will be *weakly imposed* (see (1.2.16) and the surrounding discussion).

Remark 5.2.1 *The existence and uniqueness of a solution to (5.2.5) – (5.2.6) with no-inflow boundary conditions is given in [110], for a DG scheme with tetrahedral elements - assuming that the cross-sections are constant and a given relation between h and N is satisfied.*

An alternative existence and uniqueness result is given in [132, Thm. 4] but for the discrete ordinates and DG approximation of the pure transport equation (with a generic right hand side and the angle fixed, see Section 1.3). This is under the assumptions that: $\sigma \in L_\infty(\mathcal{D})$; $\sigma(\mathbf{r}) > 0$, for almost all $\mathbf{r} \in \mathcal{D}$; and the (generic) right hand side is in $L_2(\mathcal{D})$.

We note that the generalisation of (5.2.5) – (5.2.6) to the pseudo-3D problem, previously discussed in Section 1.3, is fairly straightforward.

5.2.3 Solution Methods for the Fixed Source Problem

We will now discuss how a solution to (5.2.5) – (5.2.6) can be computed. Let us begin by writing $\psi_k^{h,N} \in V^h$ as a basis expansion, i.e.

$$\psi_k^{h,N} = \sum_{j=1}^{M_f} \Psi_{j,k} v_j \;, \quad \text{for all } k = 1, \cdots, N \;, \tag{5.2.7}$$

where $\boldsymbol{\Psi} := [\Psi_{1,1}, \cdots, \Psi_{M_f,N}]^T$ denotes the vector of coefficients for $\{\psi_k^{h,N}\}_{k=1}^N$. Then, plugging (5.2.7) into (5.2.6) we can also write $\phi^{h,N}$ as an expansion in the same basis, i.e.

$$\phi^{h,N} = \frac{1}{2\pi} \sum_{k=1}^N w_k \psi_k^{h,N} = \frac{1}{2\pi} \sum_{k=1}^N w_k \sum_{j=1}^{M_f} \Psi_{j,k} v_j = \sum_{j=1}^{M_f} v_j \sum_{k=1}^N \frac{1}{2\pi} w_k \Psi_{j,k} = \sum_{j=1}^{M_f} \Phi_j v_j \; ,$$

where $\boldsymbol{\Phi} = [\Phi_1, \cdots, \Phi_{M_f}]^T$ is the vector of coefficients for $\phi^{h,N}$, with each coefficient defined by $\Phi_j = (2\pi)^{-1} \sum_{k=1}^N w_k \Psi_{j,k}$ (for all $j = 1, \cdots, M_f$).

Plugging in the basis expansion(s) for $\psi_k^{h,N}$ and $\phi^{h,N}$ into (5.2.5), allows us to re-write (5.2.5) as the following linear system:

$$\begin{pmatrix} T & -\Sigma_S \\ -W & I \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi} \\ \boldsymbol{\Phi} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix} . \tag{5.2.8}$$

The matrix T is block diagonal, with the (i,j)th component of its kth diagonal block (each of size $M_f \times M_f$) given by $-\int_{D^h} [\boldsymbol{\Theta}_k \cdot \nabla + \sigma_{D^h}] v_i v_j \; dx + \int_{\partial D_-^h} |\boldsymbol{\Theta}_k \cdot \mathbf{n}| v_i^+ v_j$, where the definition of v_i^+ is analogous to (5.2.4), i.e.[5]

$$v_i^+(\mathbf{r}) := \lim_{\epsilon \to 0} v_i(\mathbf{r} + \epsilon \boldsymbol{\Theta}_k) , \quad \text{for all } \mathbf{r} \in \partial D_-^h .$$

The $M_f N \times M_f$ matrix Σ_S contains N identical blocks (which are themselves block diagonal), with the (i,j)th component of each block given by $\int_{D^h} (\sigma_S)_{D^h} v_i v_j$. The matrix W, of size $M_f \times M_f N$, consists of N blocks, with the kth block containing $(w_k/2\pi)$ on its diagonal and zeroes everywhere else. The matrix I denotes the $M_f \times M_f$ identity matrix. Finally, $\mathbf{f}$ consists of N column vectors each of size $M_f \times 1$, where the jth component of each block is given by $\int_{D^h} f v_j$.

The ordering of the $(M_x M_y)$ cells is important to solve (5.2.8) efficiently. If we re-order the spatial cells D^h lexicographically, based on each specific angle being considered, then we can design each of the N-blocks on the diagonal of T to be block lower-diagonal themselves (allowing solves with T to be calculated in $\mathcal{O}(M_f N)$ operations). Let the notation (i,j) correspond to the unique spatial cell D^h that is both in the ith column of cells in the x-direction and jth row of cells in the y-direction. Then, for each individual angle $\boldsymbol{\Theta}$, we define an individual angle-dependent lexicographic ordering (from first to last) by:

$$\begin{aligned} (i, j) , & \quad \text{if } \boldsymbol{\Theta}^1 > 0, \boldsymbol{\Theta}^2 > 0 , \\ (M_x + 1 - i, j) , & \quad \text{if } \boldsymbol{\Theta}^1 < 0, \boldsymbol{\Theta}^2 > 0 , \\ (i, M_y + 1 - j) , & \quad \text{if } \boldsymbol{\Theta}^1 > 0, \boldsymbol{\Theta}^2 < 0 , \\ (M_x + 1 - i, M_y + 1 - j) , & \quad \text{if } \boldsymbol{\Theta}^1 < 0, \boldsymbol{\Theta}^2 < 0 . \end{aligned} \tag{5.2.9}$$

for all $i = 1, \cdots, M_x$, for each $j = 1, \cdots, M_y$. Here we used $\boldsymbol{\Theta}^1$ and $\boldsymbol{\Theta}^2$ to denote the component of $\boldsymbol{\Theta}$ in the x and y-direction(s), respectively. A similar angle-dependent ordering is also discussed in [32]. We give illustrative examples of the angle-dependent lexicographic ordering (5.2.9) are given in Figure 5-5.

Now note the similarity between the linear system (5.2.8), which approximates the solution of the (weak form of) 2D-1D transport equation, and the linear system (3.1.10), which approximates the solution of the discretised 1D radiative transport equation. Hence, we can once again apply the source iteration method, previously discussed in Section 3.1.2, to estimate $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ in (5.2.8).

---

[5]strictly speaking v_i^+ depends on the specific direction $\boldsymbol{\Theta}_k$, but for convenience we do not include this in the notation

(a) $(\mathbf{\Theta}_k)_1, (\mathbf{\Theta}_k)_2 > 0$ (b) $(\mathbf{\Theta}_k)_1 < 0, (\mathbf{\Theta}_k)_2 > 0$ (c) $(\mathbf{\Theta}_k)_1, (\mathbf{\Theta}_k)_2 < 0$
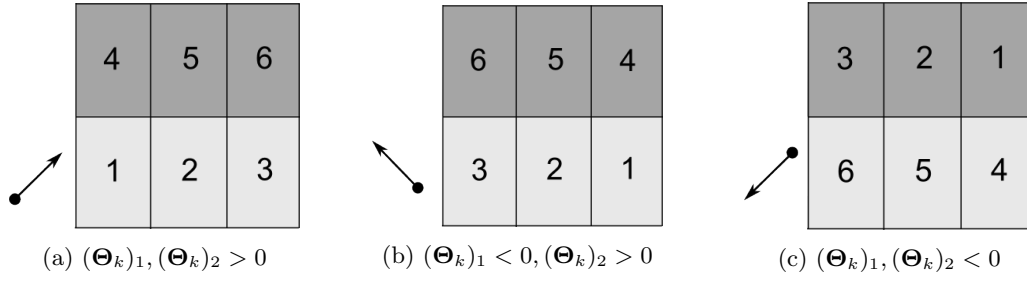
Figure 5-5: Angle-dependent lexicographic ordering of the $(M_x M_y)$ cells, when $M_x = 3$ and $M_y = 2$.

In particular, the estimate of $\mathbf{\Psi}$ at the qth iteration is given by

$$\mathbf{\Psi}^{(q)} \;=\; T^{-1}\left(\Sigma_S \mathbf{\Phi}^{(q-1)} \;+\; \mathbf{f}\right) \;, \tag{5.2.10}$$

where we assume that $\mathbf{\Phi}^{(0)} = \mathbf{0}$. Subsequently the estimate of $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi}^{(q)} \;=\; W \mathbf{\Psi}^{(q)} \;. \tag{5.2.11}$$

Moreover, we can prove the following result on the total number of floating point operations used in the source iteration method (5.2.10) – (5.2.11).

Lemma 5.2.2 *Consider the source iteration method* (5.2.10) *–* (5.2.11) *for computing a solution to the linear system* (5.2.8). *Assume the elements in the DG scheme are ordered according to the angle-dependent lexicographic ordering* (5.2.9). *Then, after K iterations of* (5.2.10) *–* (5.2.11), *the*

theoretical cost of source iteration $\sim\; \mathcal{O}\left(M_f\,N\,K\right)\;.$

Proof. Consider the qth iteration of (5.2.10) – (5.2.11). We illustrate below that each step requires just $\mathcal{O}(MN)$ operations:

- Each of the N-blocks in Σ_S corresponds to a $M_f \times M_f$ block diagonal matrix, with 4×4 blocks on its diagonal. Hence, the cost of computing $\Sigma_S \mathbf{\Phi}^{(q-1)}$ is $\mathcal{O}(M_f N)$ operations;

- Adding $M_f N \times 1$ sized vectors, i.e. $\Sigma_S \mathbf{\Phi}^{(q-1)}$ and $\mathbf{f}$, costs $\mathcal{O}(M_f N)$ operations;

- The matrix T is a block diagonal matrix of size $\mathcal{O}(M_f N) \times \mathcal{O}(M_f N)$ and each of the N-blocks are block lower-diagonal themselves. Hence, solves with T (i.e. $T^{-1}(\Sigma_S \mathbf{\Phi}^{(q-1)} + \mathbf{f})$) can be calculated in $\mathcal{O}(M_f N)$ operations;

- Each of the M_f rows in the matrix W contain only N terms. Hence, computing $W \mathbf{\Psi}^{(q)}$ involves $\mathcal{O}(M_f N)$ operations.

Therefore the cost for a single iteration of (5.2.10) – (5.2.11) requires $\mathcal{O}(MN)$ operations. The result follows. ∎

We considered implementing our novel hybrid algorithm, presented in Chapter 4, on this 2D example. However, due to the extra degrees of freedom in the 2D spatial domain, the cost of solves with the Schur complements vastly outweighed the cost of solves with source iteration in all but a few extreme cases. Hence we proceed only with the source iteration method.

We will now discuss the details of two problems involving uncertainty, giving numerical results for each.

	MOX	U0$_2$	Homo. Concrete
σ_S	1.33203	1.30986	0.1913
σ_A	2.63673E-01	6.52870E-02	4.714E-03

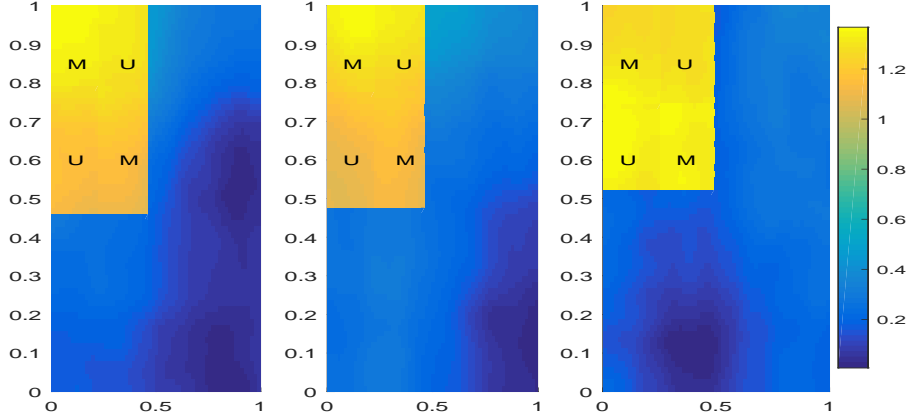Table 5.3: Assumed cross-sections for the three material types in the deterministic C5-MOX problem.

Figure 5-6: Three realisations of scattering cross-sections, for Problem 5.2.3, sampled from using the KL expansion. The MOX assemblies are denoted by 'M', and the U0$_2$ assemblies denoted by 'U'. Also recall that the C5-MOX problem models a quarter of a reactor core.

5.2.4 Random Model 1: Uniform C5-MOX

The C5-MOX problem is a common benchmark used for spatially two (and three)-dimensional neutron transport codes [43, 188, 212]. It is designed to mimic a small reactor with a heterogeneous core, by modelling a quarter of the reactor and including reflective boundary conditions on the top and left boundaries of the spatial domain (and no-inflow boundary conditions otherwise). The top left corner of the spatial domain represents a block of fuel - a two-by-two alternating arrangement of MOX and U0$_2$ fuel assemblies. The rest of the spatial domain (an L-shape) represents a shielding material, here assumed to be homogeneous concrete. See [188, Fig.1], [43, configuration C5] or Figure 5-6 for an illustration.

For the C5-MOX problem considered here, we make a number of simplifications. Firstly, as already mentioned, the C5-MOX problem includes reflective boundary conditions (i.e. (1.1.13)) - we will only consider the (vacuum) no-inflow boundary conditions, i.e.

$$\psi(\mathbf{r}, \boldsymbol{\Theta}) \; = \; 0 \; , \quad \text{if} \quad \boldsymbol{\Theta} \cdot \mathbf{n}(\mathbf{r}) < 0 \; , \quad \text{and} \quad \mathbf{r} \in \partial \mathcal{D} \; . \tag{5.2.12}$$

Secondly, the C5-MOX problem is defined on the spatial domain $[0, 64.26]^2$, whereas we consider $[0, 1]^2$. Finally, the C5-MOX problem also includes fuel assemblies with complex arrangements of fission chambers, guide tubes and U0$_2$ and MOX fuel (of varying intensities), as illustrated in [212, pg.15] and [188, Fig.3]. We simplify by ignoring the detail and instead consider a square cell of fuel, for each individual assembly - the respective cross-sections are given in Table 5.3.

For the shielding material we consider a homogenised block of concrete and present the corresponding cross-sections in Table 5.3. For details on how these cross-sections were computed, we refer the reader ahead to Section 5.2.5.

We will now introduce uncertainty into this problem. The details are given in Problem 5.2.3.

Problem 5.2.3 *Let $z_1, z_2 \sim \mathcal{U}(-\epsilon, \epsilon)$ denote uniform random variables, with $\epsilon := 0.05$. Define the*

spatial subdomains $A_{\mathrm{MOX}} = A_{\mathrm{MOX}}(\omega)$ and $A_{\mathrm{U0_2}} = A_{\mathrm{U0_2}}(\omega)$ by

$$A_{\mathrm{MOX}} := [0, 0.25 + 0.5z_1] \times [0.75 - 0.5z_2, 1] \cup [0.25 + 0.5z_1, 0.5 + z_1] \times [0.5 - z_2, 0.75 - 0.5z_2] ,$$

$$A_{\mathrm{U0_2}} := [0, 0.25 + 0.5z_1] \times [0.5 - z_2, 0.75 - 0.5z_2] \cup [0.25 + 0.5z_1, 0.5 + z_1] \times [0.75 - 0.5z_2, 1] .$$

Then, the (random) uniform C5-MOX problem is defined by (5.2.1), (5.2.2), (5.2.12) with: deterministic absorption cross-section(s) for each sub-domain A_{MOX}, $A_{\mathrm{U0_2}}$ and $\mathcal{D} \setminus (A_{\mathrm{MOX}} \cup A_{\mathrm{U0_2}})$ (as given by σ_A in Table 5.3); a (random) scattering cross-section given by a uniform random field equipped with a Matérn covariance, with parameters $\lambda_C = 1$, $\sigma_{var}^2 = 1$ and $\eta = 1.5$ (and for each sub-domain, the σ_S in Table 5.3 as its mean); and the (random) source term

$$f(\mathbf{r}, \omega) := \begin{cases} 1.5 , & \text{if } \mathbf{r} \in A_{\mathrm{MOX}}(\omega) \\ 1 , & \text{if } \mathbf{r} \in A_{\mathrm{U0_2}}(\omega) \\ 0 , & \text{otherwise} \end{cases} .$$

Three realisations of the scattering cross-section are given in Figure 5-6.

We will be interested in estimating $\mathbb{E}[Q]$, where Q is defined as

$$Q(\omega) := \frac{1}{|A^{(1)}|} \int_{A^{(1)}} \phi(\omega, \mathbf{r}) \, \mathrm{d}\mathbf{r} . \tag{5.2.13}$$

That is, for each realisation $\omega \in \Omega$, Q is the spatial average of the scalar flux over the subdomain $A^{(1)} := [0.75, 1] \times [0, 0.25] \subset \mathcal{D}$.

Numerical Results

We will now present several numerical results on the application of Monte Carlo methods to Problem 5.2.3. We consider the discretisation outlined in Section 5.2.2, with a uniform spatial mesh (in each axes, with parameters M_x and M_y) and an N-point quadrature rule defined by (1.2.11). We introduce a hierarchy of levels $\ell = 0, \cdots, L$ corresponding to a sequence of discretisation parameters defined by:

$$(M_x)_\ell := 8\left(2^\ell\right) , \quad (M_y)_\ell := 8\left(2^\ell\right) , \quad N_\ell := 18 + 10\ell , \quad d_\ell = 2\left((M_x)_\ell + (M_y)_\ell\right) . \tag{5.2.14}$$

Subsequently, we define the (spatial) mesh size for the DG method as $h_\ell := \sqrt{(M_x)_\ell^{-2} + (M_y)_\ell^{-2}}$, i.e. the (maximum) diameter of all (or any, because we assumed a uniform spatial mesh in the x and y axes) rectangles $D^{h_\ell} \in \mathcal{C}^{h_\ell}$, on the discretisation defined by level ℓ. Moreover, to sample from the uniform random field (for the scattering cross-section) in Problem 5.2.3, we use the KL expansion (defined in Section 2.3.2) truncated to d_ℓ modes (the choice of d_ℓ is again chosen empirically). An example of the resulting approximation of the scalar flux (for a single random realisation) is given in Figure 5-7.

We approximate the quantity of interest (5.2.13) by

$$Q_\ell = \frac{1}{|\mathfrak{A}|} \sum_{D^{h_\ell} \in \mathfrak{A}} \phi^\ell_{D^{h_\ell}} , \quad \text{where } \mathfrak{A} := \{D^{h_\ell} \in \mathcal{C}^{h_\ell} \mid D^{h_\ell} \subset A^{(1)}\} ,$$

where $\phi^\ell_{D^{h_\ell}}$ denotes the scalar flux approximation (computed using the discretisation parameters h_ℓ and N_ℓ) *averaged* over the four evaluations for each cell $D^{h_\ell} \in \mathcal{C}^{h_\ell}$, and $|\mathfrak{A}|$ denotes the cardinality (i.e. the number of cell elements contained within $A^{(1)}$) of $A^{(1)}$. We recall that $A^{(1)} := [0.75, 1] \times$
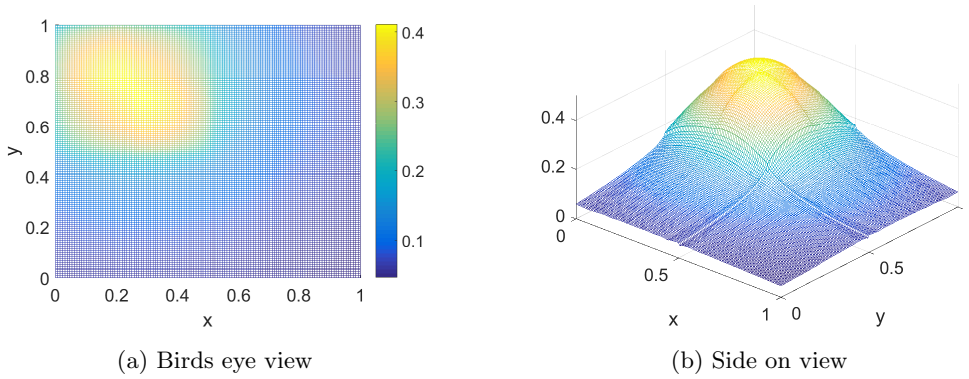
(a) Birds eye view

(b) Side on view

Figure 5-7: An approximation to the scalar flux using $M_x = M_y = 128$ and $N = 58$, for a single realisation of Problem 5.2.3. Shown at two different viewpoints.

	α_{obs}	β_{obs}	γ_{obs}	Estimated	Observed
Monte Carlo				3.9	3.3
Multilevel	1.4	2.0	2.6	2.4	2.2

Table 5.4: Summary of computational ϵ-cost rates r, where for an estimator $\widehat{Q}$, $\mathbb{E}[\mathcal{C}(\widehat{Q})] = \mathcal{O}(\epsilon^{-r})$. We estimate r in the following ways: 'Estimated' uses the numerically observed $\alpha_{\text{obs}}, \beta_{\text{obs}}, \gamma_{\text{obs}}$; 'Observed' uses the observed rates from Figure 5-9.

$[0, 0.25] \subset \mathcal{D}$.

We compare Q_ℓ to a reference solution calculated with $M_x = M_y = 128$ and $N = 58$. We will measure the error $\mathbb{E}[|Q - Q_\ell|]$, estimating the expectation using the standard Monte Carlo estimator with 512 samples.

We will now study the efficiency of the various Monte Carlo methods we have discussed. We start by numerically estimating the parameters α, γ and β from the assumptions (2.4.3), (2.4.4) and (2.4.33) respectively. The corresponding numerical results are given in Figure 5-8. From here, we observe that $\mathbb{E}[|Q - Q_\ell|] = \mathcal{O}(h_\ell^{1.4})$, i.e. $\alpha \approx 1.4$. Likewise, we estimate $\beta \approx 2.0$ and $\gamma \approx 2.6$. The rates are also summarised in Table 5.4.

Subsequently Table 5.4 gives details comparing the rates of; the numerically estimated ϵ-cost, using the numerically observed parameters $\alpha_{\text{obs}}, \beta_{\text{obs}}, \gamma_{\text{obs}}$ within Theorem 2.4.4, and the numerically observed ϵ-cost, estimated from Figure 5-9. We note reasonable agreement between the 'estimated' and 'observed' computational cost rates.

Finally, in Figure 5-9 we present numerical results comparing the computational ϵ-cost of the standard, quasi, multilevel and multilevel quasi-Monte Carlo methods, when estimating $\mathbb{E}[Q]$. The quasi-Monte Carlo samples are generated using an (extensible) randomised rank-1 lattice rule (recall Section 2.4.2), equipped with the generating vector `lattice-32001-1024-1048576.3600` (downloaded from [120]) and with $S = 8$ shifts.

We conclude that the multilevel methods (MLMC and MLQMC) and the quasi-Monte Carlo method all outperform the standard Monte Carlo method by an order of magnitude in (observed) computational ϵ-cost. However, we would generally expect the MLQMC method to outperform QMC and MLMC and we do not observe that here. We note that this is not a surprise, as to achieve (even the smallest) MSE accuracies presented in Figure 5-9, the number of QMC samples on each level are small. We conjecture that, as $\epsilon \to 0$, a further improvement for MLQMC will be observed.

We will now consider a second spatially two-dimensional fixed source problem.

(a) $\mathbb{E}\left[|Q - Q_\ell|\right]$

(b) $\mathbb{V}\left[Q_\ell\right]$ and $\mathbb{V}\left[Y_\ell\right]$

(c) $\mathbb{E}\left[\mathcal{C}(Y_\ell)\right]$
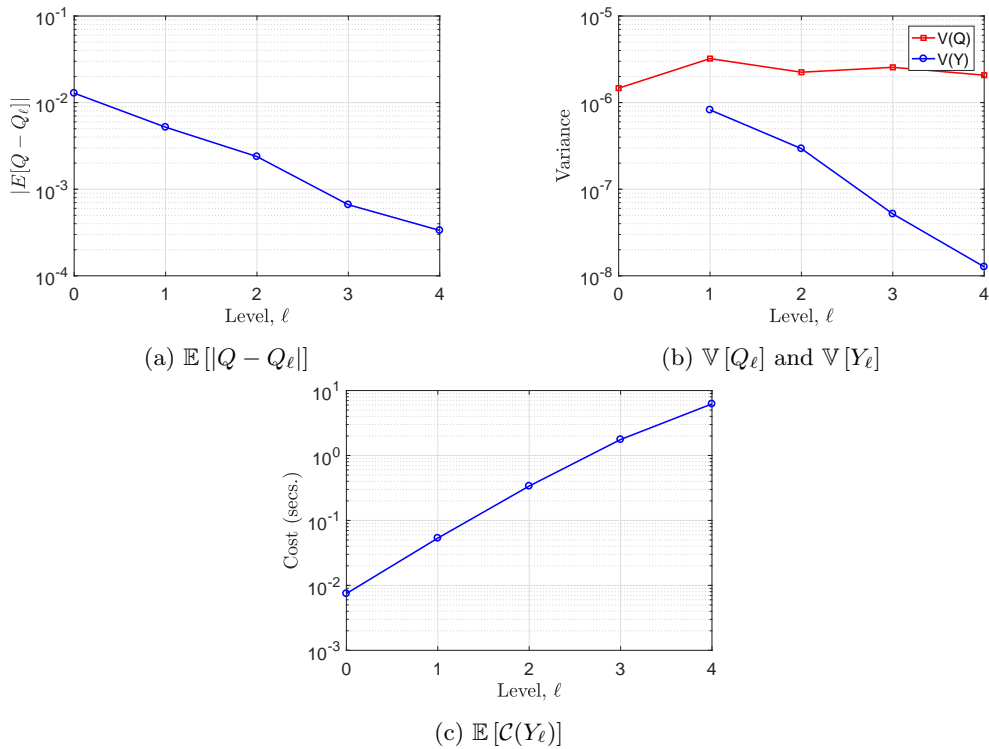
Figure 5-8: Picture (a) represents an estimate of the bias error $\mathbb{E}\left[|Q - Q_\ell|\right]$ for Problem 5.2.3. Picture (b) represents the variance of Q_ℓ and the variance of the difference process $Y_\ell = Q_\ell - Q_{\ell-1}$ (i.e. the variance reduction). Picture (c) is an estimate of the average cost, in seconds, to compute a single realisation of Y_ℓ.
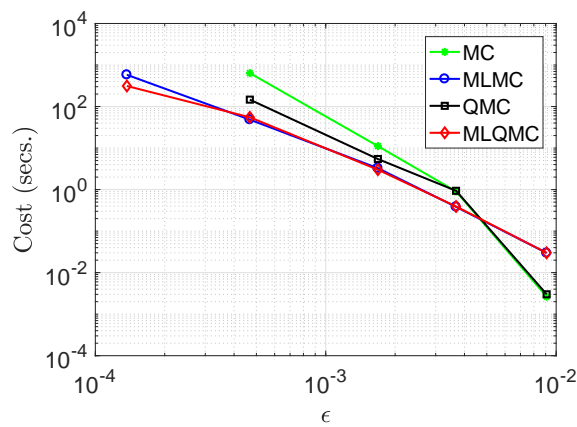
Figure 5-9: Comparison of the computational cost of standard, quasi, multilevel and multilevel quasi-Monte Carlo methods. Actual cost (in seconds) plotted against the achieved root-MSE accuracy. Details for this plot are given in (4.3.1).

Material Type	Air	Water	Granite	Cement
Volume Fraction	5%	4%	40%	51%

Table 5.5: The assumed composition of air, water, granite and cement in our model of concrete.

5.2.5 Random Model 2: Concrete Shielding

The second (fixed source) problem we consider is designed to mimic a concrete shielding problem, on a 1 metre by 1 metre block of concrete. We assume there is a fixed source of particles defined by

$$f(\mathbf{r}) \;=\; \exp\left(-5x\right) \;, \quad \text{for all } \mathbf{r} = (x, \cdot) \in \mathcal{D} = [0,1]^2 \;,$$

which for example, could be an artificial model for particles entering $\mathcal{D}$ from a nuclear reactor to the left of $\mathcal{D}$. The concrete is acting as a shield and contains a detector covering the subdomain $A^{(2)} := [0.75, 1] \times [0, 0.125] \subset \mathcal{D}$. Hence, our quantity of interest Q is defined by

$$Q(\omega) \;:=\; \frac{1}{|A^{(2)}|} \int_{A^{(2)}} \phi(\omega, \mathbf{r}) \,\mathrm{d}\mathbf{r} \;, \quad \text{for } \omega \in \Omega \;. \tag{5.2.15}$$

We also assume no-inflow vacuum boundary conditions on all sides, i.e.

$$\psi(\omega, \mathbf{r}, \mathbf{\Theta}) \;=\; 0 \;, \quad \text{if } \mathbf{\Theta} \cdot \mathbf{n}(\mathbf{r}) < 0 \;, \quad \text{and } \mathbf{r} \in \partial\mathcal{D} \;. \tag{5.2.16}$$

for almost all realisations $\omega \in \Omega$, where $\mathbf{n}(\mathbf{r})$ denotes the outward facing normal vector at $\mathbf{r}$.

The random cross-sections in this problem will come from a novel random and heterogeneous model for the cross-sections in concrete. We discuss this below, but first let us give some details on the microstructure of concrete.

Microstructure of Concrete

Concrete is a composite material that can be considered to be made up of three phases: a cement paste (a binding agent made from a mixture of water and ordinary Portland cement); an aggregate (often sand and rocks local to the reactor, such as granite or limestone); and an interfacial transition zone (e.g. containing air pores and unreacted water) [4, 104, 182]. We will assume no supplementary materials such as additives, admixtures or fly ash [4, 104], are added.

For our model of concrete we consider only components with length scales of nearly a centimetre - such a model will consist of four main components; aggregate, air, cement and water. Initially, and motivated by [14] and Figure 5-10, we assume that the concrete block has a composition of $66.\dot{6}\%$ aggregate, 5% air, $8.\dot{3}\%$ cement and 20% water. We will adjust these percentages, for reasons which are justified below. Our assumption on the final composition of the concrete is given in Table 5.5.

The smallest of the components will be the pores of air present in the interfacial transition zone. At the microstructure level there are many types of air void, see [104, Fig.11]. We will focus on 'entrapped air voids' which have typical length scales of between 0.1-1cm [104, 146].

The 'aggregate' category is commonly split into two parts, coarse aggregates (e.g. granite or limestone) and fine aggregates (e.g. sand). We will assume that the coarse aggregate is granite and that the fine aggregate is sand, and we will also assume a 3:2 ratio between coarse and fine [149]. Moreover, since both sand and cement are SiO_2 (silicon dioxide) in the majority, we will treat the fine aggregate as part of the total cement amount, as it should make little difference to the overall concrete properties.

	Granite	Dry Cement	Wet Cement	Water	Air
σ_A	3.29E-03	5.66E-03	6.51E-03	7.37E-03	0.00
σ_S	7.88E-02	1.06E-01	4.09E-01	7.12E-01	0.00

Table 5.6: Assumed (absorption and scattering) cross-sections for the different components of concrete. It is assumed that; granite has a 1% porosity, cement refers to Ordinary Portland cement and the wet cement has a 1:1 ratio of water to dry cement.

The cement is assumed to be ordinary Portland cement [14]. Moreover, we assume that the cement can vary in (water) saturation, from dry cement to wet (or saturated) cement (where we assume saturated means the water-cement ratio is 1:1). We will assume that this requires 80% of the water content (i.e. 16% out of the 20% total), which will be added to the total cement amount.

The assumptions outlined above leads us to our final assumed composition of concrete, as given by Table 5.5.
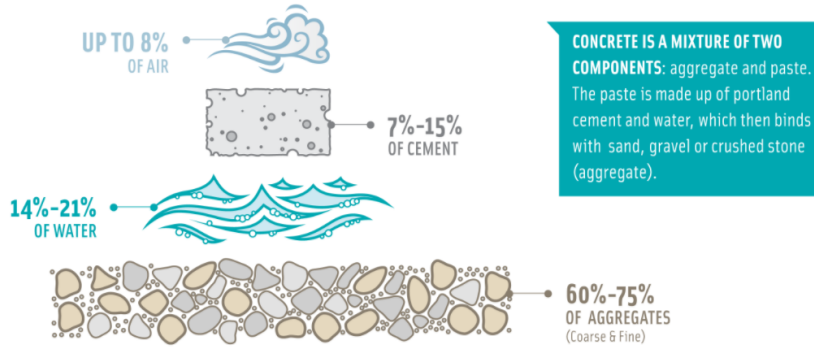
Figure 5-10: A guide to the composition of concrete. Taken from [14].

In Table 5.6 we present typical values for the absorption and scattering cross-sections of air, granite, cement and water.

Homogeneous Concrete

Given the assumed composition of our concrete in Table 5.5, we will now generate a model for the nuclear cross-sections in a homogeneous slab of concrete. This model is used in the previously discussed C5-MOX problem, see Section 5.2.4.

Considering (1.1.2) and observing the linear dependence of the (macroscopic) cross-sections on the volume fraction of each nucleus type, we suggest the following model for the (scattering) cross-section of homogeneous concrete:

$$\sigma_S(\cdot) \;=\; w^{(\mathrm{air})}\sigma_S^{(\mathrm{air})} + w^{(\mathrm{water})}\sigma_S^{(\mathrm{water})} + w^{(\mathrm{granite})}\sigma_S^{(\mathrm{granite})} + w^{(\mathrm{cement})}\sigma_S^{(\mathrm{cement})} \;. \qquad (5.2.17)$$

For each material 'm' we use $w^{(m)}$ to denote its (percentage) volume within the concrete, as given in Table 5.5, and use $\sigma_S^{(m)}$ to denote its scattering cross-section, as given in Table 5.6. The value of $\sigma_S^{(\mathrm{cement})}$ is taken to be the average of the scattering cross-sections for dry and wet cement. We apply a similar model to the absorption cross-section σ_A (with $\sigma_S^{(m)}$ replaced by $\sigma_A^{(m)}$), and compute the total cross-section $\sigma = \sigma_S + \sigma_A$ as usual.

Using the assumed material compositions in Table 5.5 and the given cross-sections in Table 5.6, it is trivial to compute the cross-sections for the homogeneous concrete using (5.2.17) - they are given in Table 5.3.

Figure 5-11: Two realisations of the material distribution within the random concrete model. Each uses a realisation of a Gaussian random field equipped with the artificial covariance (discussed in Section 2.3.2) and the parameters $\lambda_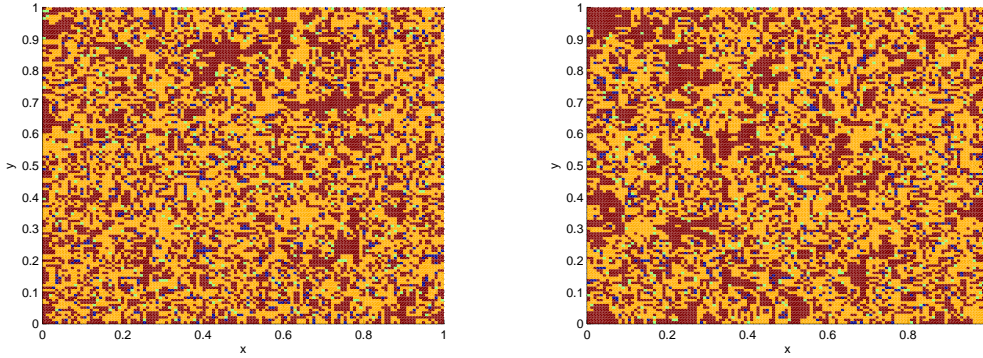C = 0.025$, $\sigma_{\mathrm{var}} = 1$, $\eta = 0.5$. We sample from the Gaussian random field using the AC expansion, truncated to $d = 3000$ modes. The colours correspond to: Dark Blue = Water; Light Blue = Air; Orange/Yellow = Wet/Dry Cement; Maroon = Coarse Aggregate.

Modelling Heterogeneity in Concrete

In the literature, the structure (and cross-sections) of concrete is often assumed to be homogeneous [6, 161, 182]. This assumption is even used in concrete mix design analysis [161], where the primary goal is to find compositions of concrete that are efficient for some purpose (e.g. shielding).

The homogeneous concrete assumption is rather surprising since concrete is a *highly heterogeneous* material [182]. Moreover, the spatial position and even the exact quantities of each material component can be unknown (they can even change over time [146, 182]). Within the literature on heterogeneous models for concrete, some papers model the aggregate as 'hard' spherical grains embedded within the cement [22]. Other papers use the *SLD method* developed in [184] which disperses the materials in the concrete uniformly, and generates the cross-sections by homogenisation. Methods that rely on some level of homogenisation, such as the SLD method, are reported to give large errors, see [211].

One other possibility is to model the heterogeneity as a realisation of a Gaussian random field, with some chosen underlying covariance function. We previously discussed this in Chapter 2, where the Karhunen-Loéve and Artificial Covariance[6] expansions were suggested as methods to sample from the (Gaussian) random field. Within the scope of concrete modelling, Gaussian fields have gained increasing popularity in the last decade, see [6, 147, 173, 214].

More generally, random fields with an underlying covariance function have been commonly used within the Uncertainty Quantification literature to represent random coefficients in a PDE, see for example [48, 123, 181]. However, each realisation of these coefficients varies continuously across the domain on which it is defined. In the context of concrete modelling, this means there are no clear interfaces between 'material of type A' and 'material of type B', i.e. water and granite.

Therefore, we propose a novel alternative which aims to generate realisations of a random field with clear and distinctive interfaces between material types (and hence cross-sections), but where the material dispersion is spatially correlated (rather than the simpler method of uniformly distributing each material type with no correlation). Simply put, this is achieved by the following

---

[6]we recall that this assumed our random field was equipped with some unknown (artificial covariance) with eigenvalues ξ_i and eigenfunctions $\eta_i(\mathbf{r}) = \cos\left(2\pi\rho_i^1 r_1\right)\cos\left(2\pi\rho_i^2 r_2\right)$

$\Upsilon(g) \in$	$[0, 0.1)$	$[0.1, 0.1125)$	$[0.1125, 0.3675)$	$[0.3675, 0.38)$	$[0.38, 0.48)$	$[0.48, 0.50]$
Material	Granite	Air	Cement	Air	Granite	Water

Table 5.7: Chosen split of the domain of the standard normal distribution (for $g \leq 0$) which defines the map $\mathcal{M}_m$, where $\Upsilon(\cdot)$ denotes the cumulative standard normal distribution.

sequence of maps $\mathcal{M}_c \circ \mathcal{M}_m \circ G(\cdot, \mathbf{r}) : \Omega \mapsto \mathbb{R}$, for all $\mathbf{r} \in \mathcal{D}$:

$$G(\cdot, \mathbf{r}) : \Omega \mapsto \mathbb{R} , \quad \mathcal{M}_m : \mathbb{R} \mapsto \mathbb{M} , \qquad\qquad \mathcal{M}_c : \mathbb{M} \mapsto \mathbb{R} .$$

Gaussian RV,   Gaussian RV to Material Type,   Material Type to Cross-Section.

Simply put this means at each spatial point $\mathbf{r}$, we sample from a Gaussian random variable (G), associate this value with a material within the concrete $(\mathcal{M}_m)$, and then map the material type to a deterministic cross-section $(\mathcal{M}_c)$. We will now give details on each of these steps.

We begin by generating a realisation of a Gaussian random field, equipped with some covariance function, $G(\cdot, \mathbf{r}) : \Omega \mapsto \mathbb{R}$, for all $\mathbf{r} \in \mathcal{D}$ (or at least finitely many points in $\mathcal{D}$ prescribed by some mesh). This gives us a notion of spatial correlation in the model. The remaining steps consider a specific spatial point $\mathbf{r} \in \mathcal{D}$, where we use $g = G(\omega, \mathbf{r})$ to denote a realisation of the random variable $G(\cdot, \mathbf{r})$, for $\omega \in \Omega$.

The second step converts the (realisation of a) random variable g to the set $\mathbb{M} \subset \mathbb{R}^+$, where each element of the set corresponds to a particular material type within the concrete - we note that it is important to retain the underlying spatial correlation in this step. Intuitively, this is achieved by assigning the material type based on where g lands on the domain of a univariate standard normal distribution, as illustrated in Figure 5-12 and Table 5.7. More formally this is given by the mapping $\mathcal{M}_m : \mathbb{R} \mapsto \mathbb{M}$, defined below in (5.2.18), which maps $g \in \mathbb{R}$ onto the set $\mathbb{M} := \{[0, 1], 2, 3, 4\}$. The elements of the set $\mathbb{M}$ each denote a material type within the concrete. The integers $\{2, 3, 4\}$ correspond to the spatial point being of $\{air, water, granite\}$ type, respectively. Whereas the interval $[0, 1]$ corresponds to the spatial point being of cement type, with $\mathcal{M}_m(g) = 0$ corresponding to dry cement, $\mathcal{M}_m(g) = 1$ corresponding to saturated cement and $\mathcal{M}_m(g) \in (0, 1)$ corresponding to a combination of $\mathcal{M}_m(g)$ dry cement and $(1 - \mathcal{M}_m(g))$ saturated cement.

When $g \leq \Upsilon^{-1}(0.5) = 0$, we define the map $\mathcal{M}_m$ by:

$$\mathcal{M}_m(g) = \begin{cases} \frac{g - \Upsilon^{-1}(0.1125)}{\Upsilon^{-1}(0.3675) - \Upsilon^{-1}(0.1125)} , & \text{if } g \in \left[\Upsilon^{-1}(0.1125), \Upsilon^{-1}(0.3675)\right) \\ 2 , & \text{if } g \in \left[\Upsilon^{-1}(0.1), \Upsilon^{-1}(0.1125)\right) \cup \left[\Upsilon^{-1}(0.3675), \Upsilon^{-1}(0.38)\right) \\ 3 , & \text{if } g \in \left[\Upsilon^{-1}(0.48), \Upsilon^{-1}(0.5)\right] \\ 4 , & \text{if } g \in \left(-\infty, \Upsilon^{-1}(0.1)\right) \cup \left[\Upsilon^{-1}(0.38), \Upsilon^{-1}(0.48)\right) , \end{cases}$$

$$(5.2.18)$$

where $\Upsilon^{-1}(\cdot)$ denotes the inverse cumulative normal distribution, with zero mean and unit variance (recall that we use non-standard notation for the standard normal cumulative distribution function (usually Φ), denoting it here by Υ, see Section 2.4.2). The definition is taken analogously for $g > 0$. The particular split (also in Table 5.7) used in (5.2.18) is chosen so that: there is a small interfacial transition zone of air pockets where cement paste does not bind properly with granite; the water is trapped by clusters of granite rocks; the final interval $g \in \left(-\infty, \Upsilon^{-1}(0.1)\right)$ allows the shape of granite to become more rounded.

The final step is to assign the cross-sections at each spatial point, by assigning the value of $\mathcal{M}_m(g)$ to the cross-section of that material. Formally this is given by the map $\mathcal{M}_c$, but it is a
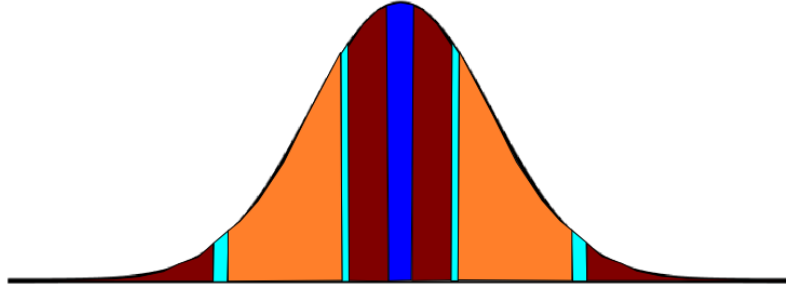
Figure 5-12: An illustration of the connection between the split of the domain of a univariate standard normal distribution and its assignment to a material type. Details are given in Table 5.7. The colours correspond to: Dark Blue = Water; Light Blue = Air; Orange = Cement (of varying saturation); Maroon = Coarse Aggregate.

simple assignment[7] from material type to cross-section(s), as given in Table 5.6.

We emphasise that the composition in Table 5.5 will now be *on average*, as (5.2.18) and Table 5.7 allow for fluctuations in the percentage of each material in the concrete.

Numerical Results

We will now present numerical results relating to the concrete model discussed in this Section. We will begin by describing the realisations of the concrete model, given in Figure 5-11, before considering the effectiveness of Monte Carlo sampling method for the Uncertainty Quantification problem.

Consider Figure 5-11 and recall that the colours correspond to: Dark Blue = Water; Light Blue = Air; Orange/Yellow = Wet/Dry Cement; Maroon = Coarse Aggregate. We begin by noting that the domain primarily consists of a cement mixture binding blocks of aggregate (of varying sizes). The aggregate is well-distributed, non-spherical and takes a variety of sizes; from small pieces of aggregate up to larger (or at least groups of small pieces) pieces. Moreover, there are a few small air pores distributed throughout the cement, and a number of small pockets of water trapped within larger blocks of the aggregate - where we recall that other non-trapped water is absorbed into the cement. We note that, whilst it is not clear in the figure, the cement varies in saturation of water (i.e. the orange/yellow colour is non-uniform).

To finish we now present numerical results on the application of Monte Carlo methods to the concrete shielding problem discussed in Section 5.2.5. Unfortunately, the microscale detail (e.g. air pores of size 0.1-1cm) on the modelled 1 metre by 1 metre domain for the concrete means that our solver struggles to converge on the meshes available to us. To illustrate the Monte Carlo algorithms on a similar problem we consider the following instead.

Problem 5.2.4 *Consider a block of concrete containing air, water, granite and cement, with the (average) percentage composition of each material given by Table 5.5. Moreover, define their respective (deterministic) cross-sections by Table 5.6.*

Then, the (random) concrete shielding problem is defined by (5.2.5), (5.2.6), (5.2.16), where the (absorption, scattering and total) cross-sections are given by the sequence of maps $\mathcal{M}_c \circ \mathcal{M}_m \circ G(\cdot, \mathbf{r}) : \Omega \mapsto \mathbb{R}$ (defined in Section 5.2.5), for all $\mathbf{r} \in \mathcal{D}$. We assume that the Gaussian random field $G(\cdot, \mathbf{r})$, for all $\mathbf{r} \in \mathcal{D}$, is equipped with the artificial covariance (discussed in Section 2.3.2) and uses the parameters; $\lambda_C = 1$, $\sigma_{var} = 1$ and $\eta = 0.5$.

---

[7]strictly speaking, we should have separate notation (for $\mathcal{M}_c$) to distinguish between the assignment of scattering and absorption cross-sections, but we avoid this technicality.
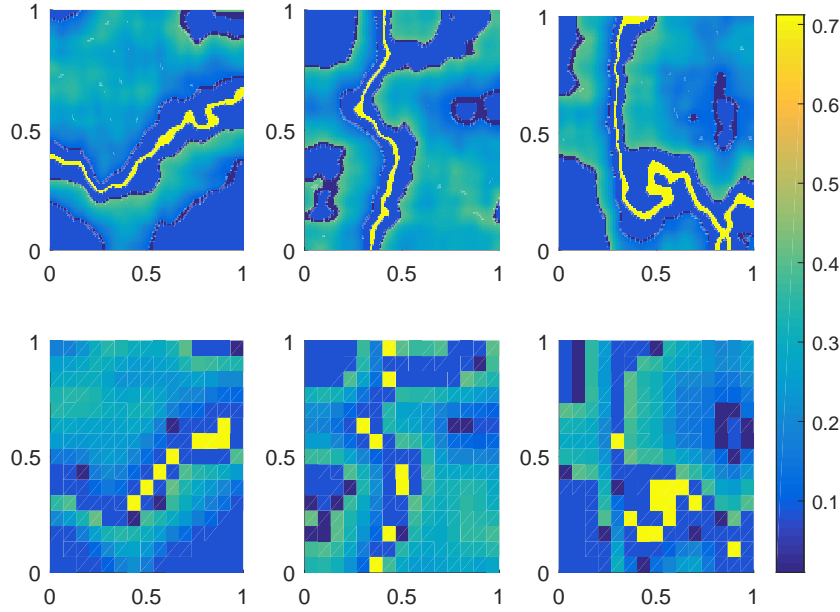
Figure 5-13: Three realisations (left to right) of the scattering cross-sections from Problem 5.2.4, presented on two spatial meshes; (Top) $M_x = M_y = 128$, (Bottom) $M_x = M_y = 16$. The colours correspond approximately to: Yellow = Water; Dark Blue = Air; Lighter Blue = Dry/Wet Cement; Blue = Coarse Aggregate.

To sample from the Gaussian random field we use the AC expansion (see Section 2.3.2). The truncation parameter in the AC expansion is defined in (5.2.14).

That is, we use the same parameters to define the random field that were also used to generate the realisations in Figure 5-11, *except* the correlation length is changed from $\lambda_C = 0.025$ to $\lambda_C = 1$. Intuitively, this corresponds to 'zooming in' on a smaller part of the 1 metre by 1 metre spatial domain (although formally we still define (5.2.1) – (5.2.2) over $\mathcal{D} := [0, 1]^2$). Three realisations, on two different meshes, of the random field for the scattering cross-section are shown in Figure 5-13.

We consider the same discretisation outlined in Section 5.2.2, and also used in the numerical results of Section 5.2.4 - with spatial parameters (M_x, M_y) and N angles. Moreover, we introduce a hierarchy of levels $\ell = 0, \cdots, L$ corresponding to a sequence of discretisation parameters defined in (5.2.14) (and the spatial mesh size h_ℓ defined below). In Figure 5-14 we present an example of an approximation to the scalar flux, for a single realisation of Problem 5.2.4.

The quantity of interest (5.2.15) is approximated by

$$Q_\ell \ = \ \frac{1}{|\mathfrak{A}|} \sum_{D^{h_\ell} \in \mathfrak{A}} \phi^\ell_{D^{h_\ell}} \, , \quad \text{where} \quad \mathfrak{A} \ := \ \{ D^{h_\ell} \in \mathcal{C}^{h_\ell} \mid D^{h_\ell} \subset A^{(2)} \} \, ,$$

where $\phi^\ell_{D^{h_\ell}}$ denotes the scalar flux approximation (computed using the discretisation parameters h_ℓ and N_ℓ) averaged over the four evaluations on each cell $D^{h_\ell} \in \mathcal{C}^{h_\ell}$, and $|\mathfrak{A}|$ denotes the cardinality of $\mathfrak{A}$. We recall that $A^{(2)} := [0.75, 1] \times [0, 0.125] \subset \mathcal{D}$.

We will compare $\{Q_\ell\}$ to a reference solution calculated with $M_x = M_y = 128$ and $N = 58$. We measure the error $\mathbb{E}[|Q - Q_\ell|]$ and estimate the expectation using the standard Monte Carlo estimator with 256 samples.

We will now numerically estimate the parameters α, γ and β in the assumptions (2.4.3), (2.4.4)

(a) Birds eye view   (b) Side on view
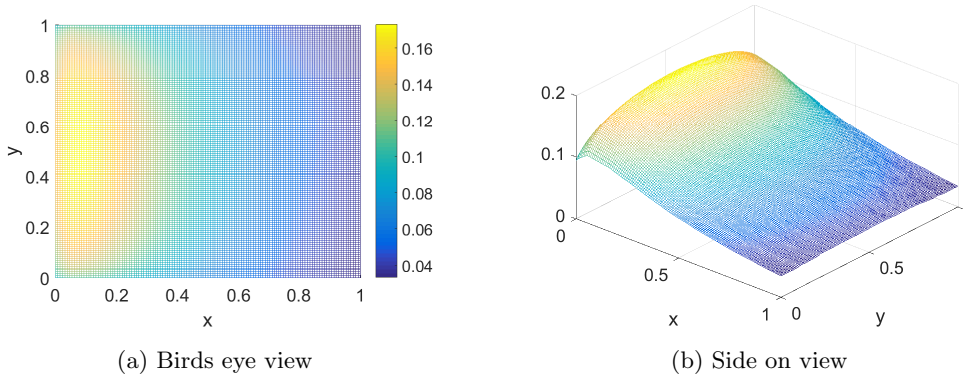
Figure 5-14: An approximation to the scalar flux for a single random realisation of Problem prob:conc. Shown at two different viewpoints.

	α_{obs}	β_{obs}	γ_{obs}	Estimated	Observed
Monte Carlo				4.0	3.9
Multilevel	1.3	2.1	2.6	2.4	2.5

Table 5.8: Summary of computational ϵ-cost rates r, where for an estimator $\widehat{Q}$, $\mathbb{E}[\mathcal{C}(\widehat{Q})] = \mathcal{O}(\epsilon^{-r})$. We estimate r in the following ways: 'Estimated' uses the numerically observed α_{obs}, β_{obs}, γ_{obs}; 'Observed' uses the observed rates from Figure 5-16.

and (2.4.33) respectively. The corresponding numerical results are given in Figure 5-15. From here, we estimate $\alpha \approx 1.3$, $\beta \approx 2.1$ and $\gamma \approx 2.6$, with these rates also summarised in Table 5.8.

Furthermore, we give details in Table 5.8 comparing the rates of; the numerically estimated ϵ-cost, using α_{obs}, β_{obs}, γ_{obs} and Theorem 2.4.4, and the numerically observed ϵ-cost, estimated from Figure 5-16. We again note excellent agreement between the 'estimated' and 'observed' computational cost rates.

Finally, we present numerical results in Figure 5-16 which compare the computational ϵ-cost of the standard, quasi, multilevel and multilevel quasi-Monte Carlo methods, when estimating $\mathbb{E}[Q]$. Again, the quasi-Monte Carlo samples are generated using an (extensible) randomised rank-1 lattice rule equipped with the generating vector `lattice-32001-1024-1048576.3600` [120], and $S = 8$ shifts.

We conclude that the multilevel methods outperform the single level methods (standard Monte Carlo and quasi-Monte Carlo) by an order (and a half) of magnitude in (observed) computational ϵ-cost. Moreover, the QMC rules do not appear to give us any gains over standard Monte Carlo (although we again note that the number of QMC samples required to achieve the MSE accuracies presented in Figure 5-16 were small and we may observe gains in the asymptotic limit $\epsilon \to 0$).

As an aside, we now make the following remark regarding the use of the Multi-Index Monte Carlo methods (which is discussed in Appendix D) to the Uncertainty Quantification problems considered here.

Remark 5.2.5 *In Figure 5-17 we present numerical results relating to the assumptions we made for MIMC, i.e. Assumption D.0.1. We introduce a multi-index $\boldsymbol{\ell} = [\ell_1, \ell_2, \ell_3] \in \mathcal{I}(L)$, the full tensor set* (D.0.1)*, corresponding to a sequence of discretisation parameters defined by:*

$$(M_x)_{\ell_1} := 8\left(2^{\ell_1}\right) , \quad (M_y)_{\ell_2} := 8\left(2^{\ell_2}\right) , \quad N_{\ell_3} := 18 + 10\ell_3 ,$$

(a) $\mathbb{E}\left[|Q - Q_\ell|\right]$

(b) $\mathbb{V}\left[Y_\ell\right]$

(c) $\mathbb{E}\left[\mathcal{C}(Q_\ell)\right]$
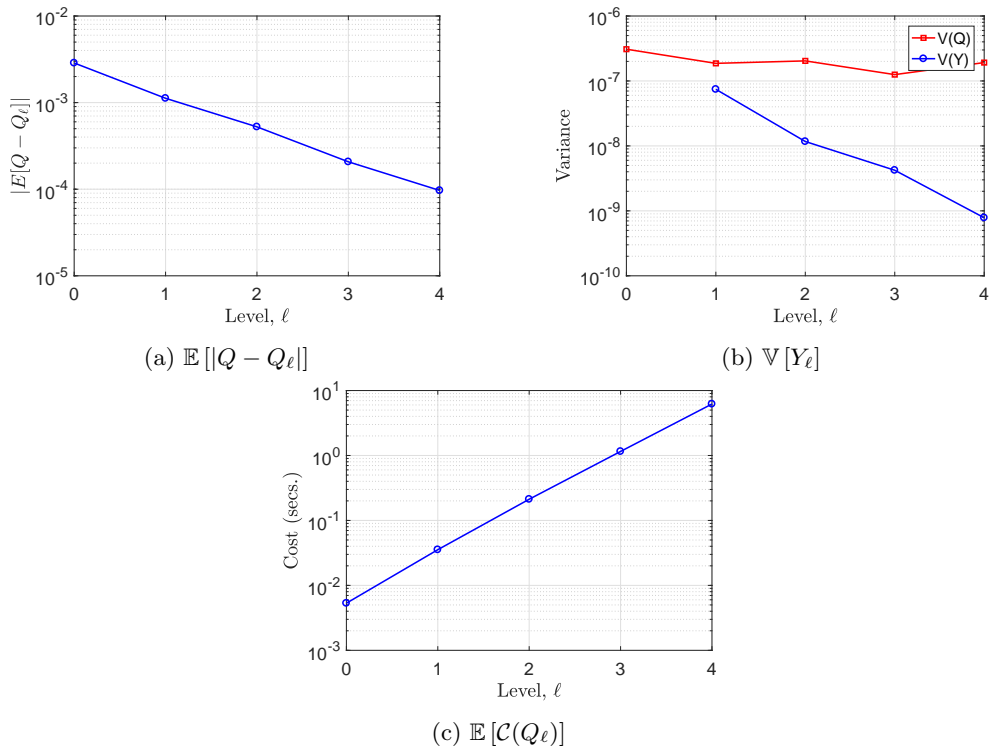
Figure 5-15: Picture (a) represents an estimate of the bias error $\mathbb{E}\left[|Q - Q_\ell|\right]$ for Problem 5.2.4. Picture (b) represents the variance of Q_ℓ and the variance of the difference process $Y_\ell = Q_\ell - Q_{\ell-1}$ (i.e. the variance reduction). Picture (c) is an estimate of the average cost, in seconds, to compute a single realisation of Y_ℓ.
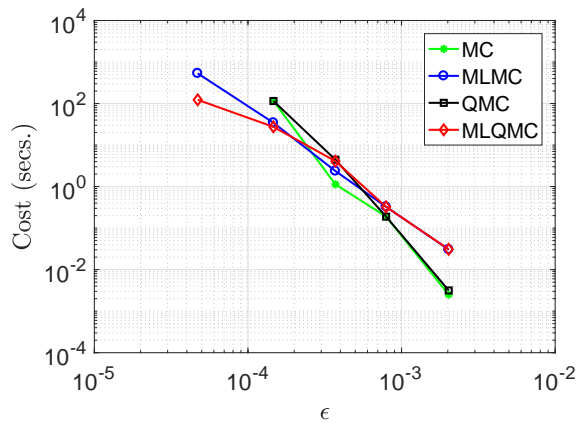
Figure 5-16: Comparison of the computational cost of standard, quasi, multilevel and multilevel quasi-Monte Carlo. Actual cost (in seconds) plotted against the achieved root-MSE accuracy. Details for this plot are given in (4.3.1).

(a) $\mathbb{E}\left[|Q - Q_{\boldsymbol{\ell}}|\right]$, with $\ell_3 = 3$

(b) $\mathbb{E}\left[\mathcal{C}(Q_{\boldsymbol{\ell}})\right]$, with $\ell_3 = 3$

(c) $\mathbb{V}\left[Y_{\boldsymbol{\ell}}\right]$, with $\ell_3 = 3$
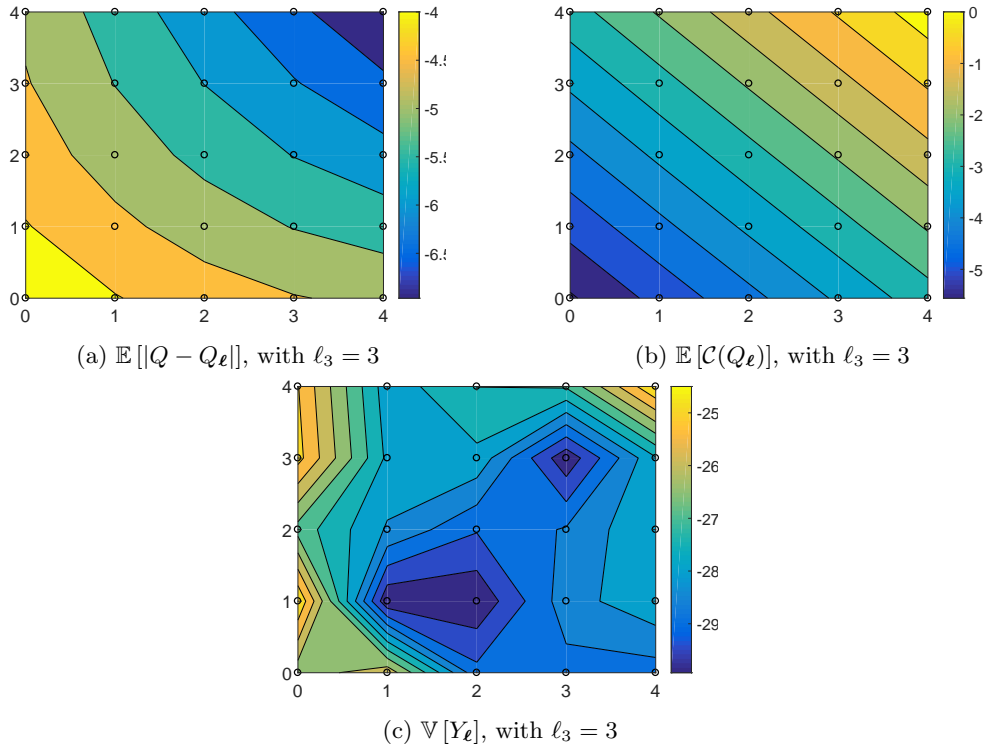
Figure 5-17: Estimates of the bias error, variance reduction and cost growth with increasing levels MIMC, for the C5-MOX problem from Section 5.2.4. $\ell_3 = 3$ corresponds to $N = 48$.

and where we subsequently define the number of KL modes by

$$d_{\ell_1,\ell_2} = 2\left((M_x)_{\ell_1} + (M_y)_{\ell_2}\right) \ .$$

The plots given in Figure 5-17 estimate the bias error, variance reduction and cost growth with respect to the multi-index $\boldsymbol{\ell}$. Increasing 'levels' in the x-axis corresponds to increasing the ℓ_1 parameter (i.e. M_x), whilst keeping all others fixed. Likewise, the y-axis corresponds to increasing ℓ_2 (i.e. (M_y)). For these figures, we have fixed the number of angles to $N = 48$ (i.e. $\ell_3 = 3$).

We note that we considered fixed angles here, because there is an additional difficulty when N varies also - due to the stability link between the number of angles and the spatial mesh (e.g. recall Theorem 3.3.11 in the spatially one-dimensional problem).

We observe that the bias error and cost plots behave as we perhaps might expect - with the bias error hinting towards the optimal choice of $\mathcal{I}$ being either the total degree set, i.e. (D.0.14), or the hyperbolic cross set, i.e. (D.0.3). However, the variance reduction does not exhibit the same clear convergence rate.

This is perhaps not surprising since MIMC requires more regularity than MLMC and it is well known that the scalar flux has limited regularity, regardless of the smoothness of the coefficients [162]. For example, even for infinitely-differentiable cross-sections on $\mathcal{D} \subset \mathbb{R}^2$, $\phi \in H^{3/2-\epsilon}(\mathcal{D})$, for $0 < \epsilon \ll 1$, where we define

$$H^r(\mathcal{D}) \ := \ \left\{q : \mathcal{D} \mapsto \mathbb{R} \ | \ \max_{|\mathbf{i}|_\infty \leq r}\left\|\frac{\partial q^{|\mathbf{i}|_1}}{\partial y_1^{i_1}\partial y_2^{i_2}}\right\|_\infty < \infty\right\} \ , \tag{5.2.19}$$

for some $r \geq 0$. We note that the paper [162] proves a more general regularity result than $\phi \in H^{3/2-\epsilon}$ (i.e. for the radiative transport equation) - with its results extending to solution(s) of

Fredholm integral equation(s) of the second kind, with at most a weakly singular kernel.

It is left as future work to analyse the MIMC variance reduction assumption (D.0.11) *in the context of radiative transport, and more generally in the context of hyperbolic integro-differential equations.*

Appendices

Appendix A

The Transport Equation

A.1   Weak Form of the Pure Transport Equation

We now prove why the weak form of (1.2.12) is given by (1.2.13). We recall that we consider the problem: Find $\psi(\mathbf{r}, \boldsymbol{\Theta})$ such that

$$[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r})]\, \psi(\mathbf{r}, \boldsymbol{\Theta}) \;=\; f(\mathbf{r}, \boldsymbol{\Theta}) \;, \tag{A.1.1}$$

for all $\mathbf{r} \in \mathcal{D}$, and for a given parameter $\boldsymbol{\Theta}$.

Throughout this section, we will use the notation set out in Section 1.2.3. We begin by multiplying (A.1.1) by a test function v (defined over a cuboid D^h), and integrating the resulting expression over the support (D^h) of v. Hence, we can write

$$
\begin{aligned}
\int_{D^h} f v \; d\mathbf{r} - \int_{D^h} \sigma \psi v \; d\mathbf{r} \;&=\; \int_{D^h} [\boldsymbol{\Theta} \cdot \nabla \psi]\, v \; d\mathbf{r} \\
&=\; \int_{D^h} v \,[\nabla \cdot \boldsymbol{\Theta} \psi] \; d\mathbf{r} \\
&=\; \int_{\partial D^h} v\, ((\boldsymbol{\Theta} \psi) \cdot \mathbf{n}) \; d\mathbf{r} - \int_{D^h} [(\boldsymbol{\Theta} \psi) \cdot \nabla v] \; d\mathbf{r} \;,
\end{aligned}
$$

where the second equality holds because $[\boldsymbol{\Theta} \cdot \nabla \psi]\, v = v\,[\boldsymbol{\Theta} \cdot \nabla \psi] = v\,[\nabla \cdot \boldsymbol{\Theta} \psi]$, and the third equality follows by Green's first identity.

Therefore, we can construct the weak form of (A.1.1): Find $\psi(\mathbf{r}, \boldsymbol{\Theta})$, such that for each D^h

$$-\int_{D^h} (\boldsymbol{\Theta} \cdot \nabla v)\psi \; d\mathbf{r} + \int_{D^h} \sigma \psi v \; d\mathbf{r} \;+\; \int_{\partial D^h} (\mathbf{F} \cdot \mathbf{n})\, v \; d\mathbf{r} \;=\; \int_{D^h} f v \; d\mathbf{r} \;,$$

for all v defined on D^h. We note that we have introduced the numerical flux $\mathbf{F}$ in place of $(\boldsymbol{\Theta} \psi)$. A natural choice of $\mathbf{F}$ would be the upwind flux, defined in (5.2.3), however many other choice can be considered. This is discussed further in Section 1.2.3.

A.2   Analytic Solution of Pure Transport equation

Consider: Find $u(\mathbf{r})$ such that

$$[\boldsymbol{\Theta} \cdot \nabla + \sigma(\mathbf{r})]\, u(\mathbf{r}) \;=\; g(\mathbf{r}) \;, \tag{A.2.1}$$

subject to the no-inflow boundary condition

$$u(\mathbf{r}) \; = \; 0 \;, \quad \text{if} \; \; \boldsymbol{\Theta} \cdot \mathbf{n}(\mathbf{r}) \; < \; 0 \;,$$

where $\boldsymbol{\Theta}$ is a given parameter. This is analogous to the no-inflow boundary condition (1.1.11), under the time-independent and mono-energetic assumptions. We present a derivation of the analytic expression for the solution below, although it is also given in [32].

To find an expression for the solution to (A.2.1) with these boundary conditions, we will use the method of characteristics. Let us re-write (A.2.1) as

$$\left[\Theta_x \frac{\partial}{\partial x} + \Theta_y \frac{\partial}{\partial y} + \Theta_z \frac{\partial}{\partial z} \right] u(\mathbf{r}) \; = \; g(\mathbf{r}) - \sigma(\mathbf{r}) u(\mathbf{r}) \;,$$

which we consider along characteristics (where we abuse notation by writing $g(s) = g(\mathbf{r}(s))$, and similarly for $\sigma(s)$ and $u(s)$) as

$$\left[\frac{\partial x}{\partial s} \frac{\partial}{\partial x} + \frac{\partial y}{\partial s} \frac{\partial}{\partial y} + \frac{\partial z}{\partial s} \frac{\partial}{\partial z} \right] u(s) \; = \; \frac{du}{ds} \; = \; g(s) - \sigma(s) u(s) \;, \tag{A.2.2}$$

with $\mathbf{r} = \mathbf{r}(s) = [x(s), y(s), z(s)]$, where s denotes the length along the characteristic and where we assume that

$$\begin{cases} \frac{\partial x}{\partial s}(s) = \Theta_x \\ \frac{\partial y}{\partial s}(s) = \Theta_y \\ \frac{\partial z}{\partial s}(s) = \Theta_z \end{cases} \Rightarrow \begin{cases} x(s) = x_0 + s\Theta_x \\ y(s) = y_0 + s\Theta_y \\ z(s) = z_0 + s\Theta_z \end{cases} \Rightarrow \mathbf{r}(s) \; = \; \mathbf{r}_0 + s\boldsymbol{\Theta} \;, \tag{A.2.3}$$

with $\mathbf{r_0} := (x_0, y_0, z_0) \in \partial \mathcal{D}$. The equation $\mathbf{r}(s) = \mathbf{r}_0 + s\boldsymbol{\Theta}$ in (A.2.3) implies the characteristics are straight lines in the direction $\boldsymbol{\Theta}$. Now, using (A.2.2) we can also write

$$\frac{du}{ds} + \sigma(s) u(s) \; = \; g(s) \;,$$

which is a simple first-order ODE that can be solved using the integrating factor method. The integrating factor can be computed as

$$I_s \; = \; \exp\left(\tau(\mathbf{r}_0, \mathbf{r}(s)) \right) \;, \tag{A.2.4}$$

where we define

$$\tau(\mathbf{r}(s_1), \mathbf{r}(s_2)) \; := \; \int_0^{s_2 - s_1} \sigma(\mathbf{r}(s_1) + s\boldsymbol{\Theta}) \, ds \;, \tag{A.2.5}$$

i.e. the integral is defined along the path of the characteristic. The quantity $|\tau(\mathbf{r}(s_1), \mathbf{r}(s_2))|$ is often called the 'optical length' or 'optical path' [27].

Hence, using the integrating factor I_s in (A.2.4), we can write

$$\begin{aligned} u(\mathbf{r}(s)) \; &= \; \exp\left[-\tau(\mathbf{r}_0, \mathbf{r}(s)) \right] \int_0^{d(\mathbf{r}, \boldsymbol{\Theta})} \exp\left[\tau(\mathbf{r}_0, \mathbf{r}(s')) \right] g(\mathbf{r}(s')) \, ds' \\ &= \; \int_0^{d(\mathbf{r}, \boldsymbol{\Theta})} \exp\left[-\tau(\mathbf{r}(s'), \mathbf{r}(s)) \right] g(\mathbf{r}(s')) \, ds' \;, \end{aligned}$$

where we define

$$d(\mathbf{r}, \boldsymbol{\Theta}) \; := \; \inf\{ s > 0 \mid \mathbf{r} - s\boldsymbol{\Theta} = \mathbf{r_0} \in \partial \mathcal{D} \} \;, \tag{A.2.6}$$
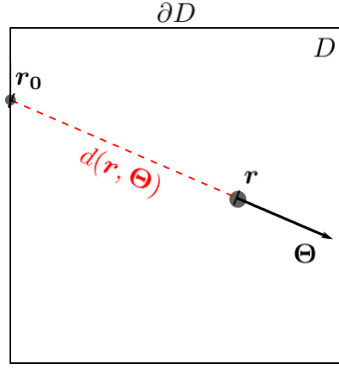
Figure A-1: A representation of the distance $d(\mathbf{r}, \boldsymbol{\Theta})$, defined in (A.2.6).

as the distance from a point $\mathbf{r}$ with a direction $\boldsymbol{\Theta}$, to the boundary in the opposite direction (i.e. $-\boldsymbol{\Theta}$). This is illustrated in Figure A-1. More explicitly, in 3D we can write $d(\mathbf{r}, \boldsymbol{\Theta})$ as

$$d(\mathbf{r}, \boldsymbol{\Theta}) \;=\; \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} \;,$$

where at least one of x_0, y_0 and z_0 will be zero, dependent on $\mathbf{r}_0 \in \partial \mathcal{D}$.

Finally, we note that for problems where, in general, $|\boldsymbol{\Theta}| \neq 1$ (such as the 1D slab geometry where $\mu \in [-1, 1]$), then we can write

$$u(\mathbf{r}(s)) \;=\; \mathcal{S}_{\boldsymbol{\Theta}} g(\mathbf{r}(s)) \;:=\; |\boldsymbol{\Theta}|^{-1} \int_0^{d(\mathbf{r}, \boldsymbol{\Theta})} \exp\left[-\boldsymbol{\Theta}^{-1} \tau(\mathbf{r}(s'), \mathbf{r}(s))\right] g(\mathbf{r}(s')) \, ds' \;, \qquad \text{(A.2.7)}$$

where the definition of $d(\mathbf{r}, \boldsymbol{\Theta})$ in (A.2.6) changes to

$$d(\mathbf{r}, \boldsymbol{\Theta}) \;:=\; \inf\{s > 0 \mid \mathbf{r} - s(\boldsymbol{\Theta}/|\boldsymbol{\Theta}|) = \mathbf{r}_0 \in \partial \mathcal{D}\} \;,$$

and we have introduced the solution operator $\mathcal{S}_{\boldsymbol{\Theta}}$. The solution operator $\mathcal{S}_{\boldsymbol{\Theta}}$ is at the heart of why the pure transport problem (1.3.8) is useful in radiative transport. It provides a closed-form solution for the angular flux (for each given angle), as illustrated by the following Proposition.

Proposition A.2.1 *Let ψ be a solution to (1.3.1), (1.1.11). Then, ψ is uniquely determined by*

$$\psi(\mathbf{r}, \boldsymbol{\Theta}) \;=\; \mathcal{S}_{\boldsymbol{\Theta}}\left((\sigma_S + \nu \sigma_F)\phi + f\right)(\mathbf{r}) \;, \qquad \text{(A.2.8)}$$

where ϕ is the scalar flux, defined in (1.3.2).

Appendix B

Fundamentals of Monte Carlo

The following results can be found in any graduate text or lectures notes, but since they are at the heart of Monte Carlo sampling we outline them below for completeness. For clarity, we will alter the notation used in the main body of this thesis. In particular, we replace the notation $\widehat{Q}_h$ for the estimator of $\mathbb{E}[Q]$ by $\widehat{Q}^N$, i.e. we remove the notion of approximating Q (removing the discretisation parameter h) and emphasise the dependence of the estimator on the number of samples N. Such a notation change ensures that the (N-sample) MC estimator (the sample mean) given in (2.4.8), becomes

$$\widehat{Q}^N := \frac{1}{N} \sum_{n=1}^{N} Q(\mathbf{Z}^{(n)}) \,,$$

where $\{\mathbf{Z}^{(n)}\}_{n=1}^{N}$ denotes a sequence of (possibly random fields corresponding to) i.i.d. MC samples.

Below, we outline the theoretical justification of why the Monte Carlo estimator converges. Primarily, the convergence occurs because of the law of large numbers, although the Central Limit Theorem allows us to state stronger statements under the additional assumption that $Q \in L_2(\Omega)$, defined in (2.1.3). We recall that we have a complete probability space $(\Omega, \mathcal{G}, \mathbb{P})$, with sample space Ω, probability measure $\mathbb{P} : \mathcal{G} \mapsto [0, 1]$.

B.1 Monte Carlo Convergence

Theorem B.1.1 ((Weak and Strong) Law of Large Numbers) *Assume that $\{Q(\mathbf{Z}^{(n)})\}_{n=1}^{N}$ are independent and identically distributed, with finite mean $\mathbb{E}\left[Q(\mathbf{Z}^{(n)})\right] = \mu$, for all $n = 1, \cdots, N$. Then,*

$$\widehat{Q}^N \to \mu \,, \quad as \ N \to \infty \,.$$

The weak law states that the convergence $\widehat{Q}^N \to \mu$ is in probability, i.e. for all $\epsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[|\widehat{Q}^N - \mu| < \epsilon\right] = 1 \,. \tag{B.1.1}$$

A stronger statement, that the convergence is almost surely, i.e.

$$\mathbb{P}\left[\lim_{N\to\infty} \widehat{Q}^N = \mu\right] = 1 \,, \tag{B.1.2}$$

is given by the strong law.

Whilst the statements for the strong and weak law are slightly different, provided the random variables have finite mean then the general statement is:

131

- the distribution of $\widehat{Q}^N$ gets closer to the deterministic value $\mathbb{E}[Q] = \mu$, as $N \to \infty$.

The Central Limit Theorem, which we present next, takes the above statement further - stating that, under the additional assumption that the random variables have finite variance:

- $\widehat{Q}^N$ is *normally distributed* around μ;

- with standard deviation $\sqrt{N}$.

The last of these motivates the Monte Carlo convergence result, presented in (2.4.10).

Theorem B.1.2 (Central Limit Theorem) *Assume that $\{Q(\mathbf{Z}^{(n)})\}_{n=1}^N$ are independent and identically distributed, with finite mean $\mathbb{E}\left[Q(\mathbf{Z}^{(n)})\right] = \mu$ and finite variance $\mathbb{V}\left[Q(\mathbf{Z}^{(n)})\right] = \sigma^2$, for $n = 1, \cdots, N$. Then,*

$$\sqrt{N} \, \frac{\widehat{Q}^N - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \ ,$$

where $\xrightarrow{d}$ denotes convergence in distribution. That is, if F_N denotes the cumulative distribution of $\widehat{Q}^N$, then for every x such that Υ is continuous at x,

$$\lim_{N \to \infty} F_N(x) = \Upsilon(x) \ , \tag{B.1.3}$$

where Υ denotes the cumulative standard normal distribution, defined in (2.4.16). Recall that we use non-standard notation for the cumulative standard normal distribution (usually Φ), denoting it here by Υ.

B.2 Unbiased Estimators

Now that we know the Monte Carlo estimator (2.4.8) converges to $\mathbb{E}[Q]$, we will prove that the estimator is unbiased, i.e. $\mathbb{E}\left[\widehat{Q}^N\right] = \mathbb{E}[Q]$. We also prove that the sample variance, introduced in (2.4.12), is an unbiased estimator of $\mathbb{V}\left[\widehat{Q}^N\right]$.

As above, we assume that $\{Q(\mathbf{Z}^{(n)})\}_{n=1}^N$ are independent and identically distributed, with finite mean $\mathbb{E}\left[Q(\mathbf{Z}^{(n)})\right] = \mu$ and finite variance $\mathbb{V}\left[Q(\mathbf{Z}^{(n)})\right] = \sigma^2$, for all $n = 1, \cdots, N$.

Sample Mean

The proof that the MC estimator is unbiased is simple, by the linearity of the expectation and the independence of $\mathbf{Z}^{(n)}$, for all $n = 1, \cdots, N$:

$$\mathbb{E}\left[\widehat{Q}^N\right] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N Q(\mathbf{Z}^{(n)})\right] = \frac{1}{N} \sum_{n=1}^N \mathbb{E}\left[Q(\mathbf{Z}^{(n)})\right] = \frac{N}{N} \mu = \mu \ .$$

Sample Variance

Recall the definition of the sample variance in (2.4.12), i.e.

$$\widehat{v}^N := \frac{1}{N(N-1)} \sum_{n=1}^N \left(Q(\mathbf{Z}^{(n)}) - \widehat{Q}^N\right)^2 \ ,$$

To see that $\widehat{v}^N$ is an unbiased estimator of $\mathbb{V}\left[\widehat{Q}^N\right]$, i.e. $\mathbb{E}\left[\widehat{v}^N\right] = \mathbb{V}\left[\widehat{Q}^N\right] = \sigma^2/N$, write

$$
\begin{aligned}
N(N-1)\widehat{v}^N &= \sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \widehat{Q}^N\right)^2 \\
&= \sum_{n=1}^{N} \left(\left(Q(\mathbf{Z}^{(n)}) - \mu\right) - \left(\widehat{Q}^N - \mu\right)\right)^2 \\
&= \sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2 - 2\left(Q(\mathbf{Z}^{(n)}) - \mu\right)\left(\widehat{Q}^N - \mu\right) + \left(\widehat{Q}^N - \mu\right)^2 \\
&= \left(\sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2\right) - 2\left(\widehat{Q}^N - \mu\right)\sum_{n=1}^{N}\left(Q(\mathbf{Z}^{(n)}) - \mu\right) + N\left(\widehat{Q}^N - \mu\right)^2 \\
&= \left(\sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2\right) - 2N\left(\widehat{Q}^N - \mu\right)^2 + N\left(\widehat{Q}^N - \mu\right)^2 \\
&= \left(\sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2\right) - N\left(\widehat{Q}^N - \mu\right)^2 \\
&= \sum_{n=1}^{N} \left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2 - \left(\widehat{Q}^N - \mu\right)^2 \ ,
\end{aligned}
$$

where the fifth equality holds because $\sum_{n=1}^{N} Q(\mathbf{Z}^{(n)}) = N\widehat{Q}^N$, by (2.4.8). Then, applying the expectation operator, we have

$$
\begin{aligned}
\mathbb{E}[\widehat{v}^N] &= \frac{1}{N(N-1)}\mathbb{E}\left[\sum_{n=1}^{N}\left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2 - \left(\widehat{Q}^N - \mu\right)^2\right] \\
&= \frac{1}{N(N-1)}\left(\sum_{n=1}^{N}\mathbb{E}\left[\left(Q(\mathbf{Z}^{(n)}) - \mu\right)^2\right] - \mathbb{E}\left[\left(\widehat{Q}^N - \mu\right)^2\right]\right) \\
&= \frac{1}{N(N-1)}\left(\sum_{n=1}^{N}\mathbb{V}\left[Q(\mathbf{Z}^{(n)}) - \mu\right] - \mathbb{V}\left[\widehat{Q}^N - \mu\right]\right) \\
&= \frac{1}{N(N-1)}\left(\sum_{n=1}^{N}\mathbb{V}\left[Q(\mathbf{Z}^{(n)})\right] - \mathbb{V}\left[\widehat{Q}^N\right]\right) \\
&= \frac{1}{N(N-1)}\sum_{n=1}^{N}\left(\sigma^2 - \frac{\sigma^2}{N}\right) = \frac{\sigma^2}{N} \ ,
\end{aligned}
$$

by the linearity of $\mathbb{E}[\cdot]$ and since $\mathbb{E}\left[X^2\right] = \mathbb{V}[X] + \mathbb{E}[X]^2 = \mathbb{V}[X]$, for a zero-mean random variable X.

Appendix C

Key Proofs within Multilevel Monte Carlo

We will now present a proof (and a sketch of a proof) for two key results at the heart of multilevel Monte Carlo - previously discussed in Section 2.4.3.

C.0.1 Proof of Theorem 2.4.2

We begin by proving the optimal (in the sense of minimising the cost of the estimator, whilst achieving a given accuracy) distribution of the samples N_ℓ across the levels $\ell = 0, \cdots, L$, i.e. the result of Theorem 2.4.2. We then present a sketch of the proof of the complexity theory for multilevel Monte Carlo, i.e. Theorem 2.4.4.

Proof of Theorem 2.4.2. The statement in the lemma is equivalent to the following constrained optimisation problem:

$$\text{Minimize} \quad \sum_{\ell=0}^{L} N_\ell \mathcal{C}_\ell \ , \tag{C.0.1}$$

$$\text{subject to} \quad \sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Y_\ell] \ = \ \frac{1}{2}\epsilon^2 \ . \tag{C.0.2}$$

since the overall cost of the MLMC estimator is $\sum_{\ell=0}^{L} N_\ell \mathcal{C}_\ell$ (i.e. the number of samples multiplied by the cost of each, summed over all levels) and the variance of the MLMC estimator was given in (2.4.30).

To find the optimal $\{N_\ell\}$ we use the method of Lagrange multipliers. We begin by defining the Lagrangian $\mathscr{L}$, with an associated Lagrange multiplier μ^2, as:

$$\mathscr{L}\left(\{N_\ell\}; \mu\right) \ := \ \sum_{\ell=0}^{L} N_\ell \mathcal{C}_\ell + \mu^2 \left(\sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Y_\ell] - \frac{1}{2}\epsilon^2 \right) \ .$$

We seek the μ and the $\{N_\ell\}$ that minimise $\mathscr{L}\left(\{N_\ell\}; \mu\right)$.

Consider the derivative of $\mathscr{L}$ with respect to N_ℓ, for each $\ell = 0, \cdots, L$, and set it to zero, i.e.

$$0 \ = \ \frac{\partial \mathscr{L}}{\partial N_\ell} \ = \ \mathcal{C}_\ell - \mu^2 \mathbb{V}[Y_\ell] N_\ell^{-2} \ ,$$

then simple algebra implies that, for all $\ell = 0, \cdots, L$,

$$N_\ell = \mu \sqrt{\frac{\mathbb{V}[Y_\ell]}{\mathcal{C}_\ell}} . \tag{C.0.3}$$

Similarly, consider the derivative of $\mathscr{L}$ with respect to μ and set it to zero, i.e.

$$0 = \frac{\partial \mathscr{L}}{\partial \mu} = 2\mu \left(\sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Y_\ell] \right) - \mu \epsilon^2 = 2 \left(\sum_{\ell=0}^{L} \sqrt{\mathcal{C}_\ell \mathbb{V}[Y_\ell]} \right) - \mu \epsilon^2 ,$$

where we have used (C.0.3). This implies that

$$\mu = 2\epsilon^{-2} \sum_{\ell=0}^{L} \sqrt{\mathcal{C}_\ell \mathbb{V}[Y_\ell]} . \tag{C.0.4}$$

The proof finishes by combining (C.0.3) with (C.0.4) and taking the ceiling function of the subsequent N_ℓ. The ceiling function ensures that the number of samples is an integer and also ensures that the sampling error is sufficient small, i.e. $\sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Y_\ell] \leq \frac{1}{2}\epsilon^2$.

We also note that it is simple to see this choice of N_ℓ is a minimum, rather than a maximum, by considering the second derivative of $\mathscr{L}$ with respect to each N_ℓ - and then observing that μ, $\mathbb{V}[Y_\ell]$, $N_\ell > 0$, for all $\ell = 0, \cdots, L$.

∎

C.0.2   Sketch of the Proof of Theorem 2.4.4

The following proof is a sketch of those outlined in [80, 48].

Sketch of the Proof of Theorem 2.4.4.   Assume for simplicity that $h_0 = 1$ and that there exists a $\mathfrak{c} \in \mathbb{N} \setminus \{1\}$ such that

$$h_\ell = \mathfrak{c}^{-1} h_{\ell-1} , \quad \text{for all} \ \ell = 1, \cdots, L .$$

We note that this condition is not restrictive to this proof, and the results above generalise provided that $\{h_\ell\}_{\ell=1}^{L}$ satisfies $c_1 \leq (h_{\ell-1}/h_\ell) \leq c_2$, for all $\ell = 1, \cdots, L$ and some constants $1 < c_1 \leq c_2 < \infty$ [197].

There are two main parts to the proof. Firstly, we show that there exists an L which ensures the squared bias error is less than $\epsilon^2/2$. Let us define

$$L := \left\lceil \alpha^{-1} \log_{\mathfrak{c}} \left(\sqrt{2} c_3 \epsilon^{-1} \right) \right\rceil , \tag{C.0.5}$$

where $\lceil \cdot \rceil$ denotes the ceiling function and c_3 denotes the hidden constant in (2.4.3). Moreover, let $\widehat{Q}_h^{MLMC}$ denote the MLMC estimator defined in (2.4.29). Then, by noting that (C.0.5) implies $c_3 h_L^\alpha = c_3 \mathfrak{c}^{L\alpha} \leq (1/\sqrt{2})\epsilon$, as well as using $\mathbb{E}\left[\widehat{Q}_h^{MLMC} \right] = \mathbb{E}[Q_h]$ and the assumption (2.4.3), we can write

$$\left(\mathbb{E}\left[\widehat{Q}_h^{MLMC} \right] - \mathbb{E}[Q] \right)^2 = (\mathbb{E}[Q_h] - \mathbb{E}[Q])^2 = c_3^2 h_L^{2\alpha} \leq \frac{\epsilon^2}{2} .$$

The second steps involves the consideration of three possible cases: $\beta > \gamma$; $\beta = \gamma$; and $\beta < \gamma$. For each case, we compute the optimal N_ℓ by using (2.4.32), from which it follows that the sampling error $\mathbb{V}\left[\widehat{Q}_h^{MLMC} \right] \leq \epsilon^2/2$ (see e.g. the proof of Lemma 2.4.2). Moreover, recalling (2.4.32) and

using the assumptions (2.4.4) and (2.4.33), we can write

$$
N_\ell - 1 \;=\; \epsilon^{-2} \left(\sum_{\ell=0}^{L} \sqrt{\mathbb{V}[Y_\ell]\mathcal{C}_\ell} \right) \sqrt{\frac{\mathbb{V}[Y_\ell]}{\mathcal{C}_\ell}} \;\leq\; c\epsilon^{-2} \left(\sum_{\ell=0}^{L} h_\ell^{\frac{1}{2}(\beta-\gamma)} \right) h_\ell^{\frac{1}{2}(\beta+\gamma)} \;, \tag{C.0.6}
$$

for some constant $c > 0$. Therefore, the total cost of the MLMC estimator is $\sum_\ell N_\ell \mathcal{C}_\ell \leq c\epsilon^{-2} \left(\sum_\ell h_\ell^{\frac{1}{2}(\beta-\gamma)} \right)^2$ and

$$
\sum_\ell h_\ell^{\frac{1}{2}(\beta-\gamma)} \;\leq\;
\begin{cases}
(L+1)h_0^{\frac{1}{2}(\beta-\gamma)} \;=\; (L+1)\mathfrak{c}^L h_L^{\frac{1}{2}(\beta-\gamma)} \;, & \text{if } \beta > \gamma \\
(L+1) \;, & \text{if } \beta = \gamma \\
(L+1)h_L^{\frac{1}{2}(\beta-\gamma)} \;, & \text{if } \beta < \gamma
\end{cases}
$$

Then the resulting bound on the computational cost, i.e. (2.4.34), follows by using $h_L \sim \epsilon^{1/\alpha}$ and the assumptions $\alpha \geq \frac{1}{2}\min\{\beta,\gamma\}$ and $\epsilon < \exp(-1)$.

∎

Appendix D

Further Variance Reduction: Multi-Index Monte Carlo

For MLMC we considered first-order differences $Y_\ell = \mathcal{Q}_\ell - \mathcal{Q}_{\ell-1}$, with the levels parameterised by a scalar $\ell \in \mathbb{N}$. The differencing allowed cheap and inaccurate estimates of $\mathbb{E}[Q_h]$ to be corrected by estimators of decreasing variance (as ℓ increases). Multi-Index Monte Carlo (MIMC) aims to extend this idea, by considering higher-order differences across multi-dimensional levels (formally a multi-index) with the aim of achieving improved variance reduction over MLMC.

The theoretical foundations of MIMC were only very recently introduced in [99]. Since then, the method has been successfully applied to problems concerning elliptic PDEs [172] and stochastic differential equations [101]. Also, like MLMC, the ideas are complimentary to QMC estimators [172] and stochastic collocation [98].

As MIMC uses a hierarchy of discretisations, in much the same way as MLMC, it is perhaps easiest to explain MIMC by comparing it to MLMC and noting the two key differences. The remaining formulae within MIMC are clear extensions of those in MLMC. For simplicity (see ahead to Remark D.0.6), let $s = s^{\mathrm{det}}$ denote the dimensionality of the deterministic model (e.g. PDE)[1]. In the case of the full time-dependent RTE (1.1.5), which is defined on a subset of $\mathbb{R}^3 \times \mathbb{S}_2 \times \mathbb{R}^+ \times \mathbb{R}^+$, then $s = s^{\mathrm{det}} = 7$.

The first difference arises from the choice of *levels*. In MLMC we considered a hierarchy of scalar-valued levels $\mathbb{N} \ni \ell = 0, \cdots, L$, corresponding to a decreasing sequence of discretisation parameters $h_0 > \cdots > h_L$, with $h_\ell \in \mathbb{R}$ for all $\ell = 0, \cdots, L$. For example, in the context of a mesh-based solver, this could correspond to a $h_\ell^{-1} \times \cdots \times h_\ell^{-1}$ mesh. Hence, if we increase the level (i.e. $\ell \to \ell + 1$) the corresponding mesh is typically $h_{\ell+1}^{-1} \times \cdots \times h_{\ell+1}^{-1}$, for $h_\ell^{-1} < h_{\ell+1}^{-1}$, i.e. the mesh is *refined for all s dimensions*.

On the other hand, MIMC defines levels by a *multi-index*, i.e. a vector $\boldsymbol{\ell} = [\ell_1, \cdots, \ell_s] \in \mathbb{N}^s$. That is, the levels $\boldsymbol{\ell}$ are s-dimensional and correspond to a set of discretisation parameters $h_{\boldsymbol{\ell}} \in \mathbb{R}^s$. The additional flexibility of $h_{\boldsymbol{\ell}} \in \mathbb{R}^s$ (over $h_\ell \in \mathbb{R}$) allows us to selectively choose which directions we want to refine in, without needing to refine elsewhere. For the example of a mesh-based method (and assuming $s = 2$), the discretisation parameter $h_{\boldsymbol{\ell}}$ corresponds to a $h_{\ell_1}^{-1} \times h_{\ell_2}^{-1}$ mesh. Then, when refining we can consider the meshes $h_{\ell_1}^{-1} \times h_{\ell_2}^{-1}$ and $h_{\ell_1+1}^{-1} \times h_{\ell_2}^{-1}$ (i.e. $\ell_1 \to \ell_1 + 1$ only), without needing to consider $h_{\ell_1+1}^{-1} \times h_{\ell_2+1}^{-1}$ (i.e. both $\ell_1 \to \ell_1 + 1$ and $\ell_2 \to \ell_2 + 1$).

Of course, there is an upper limit for the possible discretisation in each dimension. In MLMC this was given by the scalar $L = L^{MLMC} \in \mathbb{N}$. For MIMC we consider another multi-index

---

[1]The s here is not to be confused with the number of shifts S for randomised QMC estimators

$\mathbf{L} = [L_1, \cdots, L_s] \in \mathbb{N}^s$ which, along with some given rule, subsequently defines an index set $\mathcal{I}(\mathbf{L})$. We are then restricted to levels (i.e. multi-indices) $\boldsymbol{\ell} \in \mathcal{I}(\mathbf{L})$. For simplicity in the remaining discussion, we will only consider the case where $L_i = L \in \mathbb{N}$, for a given $L \in \mathbb{N}$, for all $i = 1, \cdots, s$. Hence, we replace the notation $\mathbf{L} \in \mathbb{N}^s$ and $\mathcal{I}(\mathbf{L})$ in MIMC, with L and $\mathcal{I}(L)$.

The choice of $\mathcal{I} = \mathcal{I}(L)$ is fundamental to the effectiveness of MIMC. For now, let us consider perhaps the simplest $\mathcal{I}$, the *full tensor index set* defined by

$$\mathcal{I}(L) := \{ \boldsymbol{\ell} \in \mathbb{N}_0^s \mid \ell_i \leq L, \text{ for all } i = 1, \cdots, s \} . \tag{D.0.1}$$

Other common examples of index set include: the (weighted) total degree index set [99]

$$\mathcal{I}(L) := \{ \boldsymbol{\ell} \in \mathbb{N}_0^s \mid \boldsymbol{\ell} \cdot \boldsymbol{\delta} \leq L \} , \tag{D.0.2}$$

where $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_s] \in (0,1]^s$ denotes a vector of given weights satisfying $\boldsymbol{\delta} \cdot \mathbf{1} = 1$; and the (weighted) hyperbolic cross index set [19, 172]

$$\mathcal{I}(L) := \left\{ \boldsymbol{\ell} \in \mathbb{N}_0^s \mid \prod_{i=1}^{s} \max\{1, \delta_i \ell_i\} \leq L \right\} , \tag{D.0.3}$$

where $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_s]$ denotes a vector of given weights satisfying $\prod_{i=1}^{s} \delta_i = 1$ and $\delta_i > 0$, for all $i = 1, \cdots, s$. There is also scope for algorithms that adaptively build index set(s), e.g. [171] which is based on an adaptive algorithm in the field of sparse grids [77]. Hence we have a plethora of possibilities for $\mathcal{I}$ - and the only requirement is that $\mathcal{I}$ is a *downward closed set* [172]. That is, if

$$\boldsymbol{\ell} \in \mathcal{I} , \quad \text{and} \quad \mathbf{k} \leq \boldsymbol{\ell} \quad \Rightarrow \quad \mathbf{k} \in \mathcal{I} ,$$

where here $\mathbf{k} \leq \boldsymbol{\ell}$ denotes elementwise inequality, i.e. $k_i \leq \ell_i$, for all $i = 1, \cdots, s$.

The second difference between MLMC and MIMC lies in the *choice of differencing*. For MLMC, we considered a first order difference between neighbouring levels ℓ and $(\ell-1)$, i.e. $Q_\ell - Q_{\ell-1}$. For MIMC we also consider first-order differences between neighbouring levels but, due to the multi-dimensional structure of the index set $\mathcal{I}$, this has a different meaning. In particular, we use the first-order *mixed* difference operator $\Delta := \otimes_{i=1}^{s} \Delta_i$, defined as the tensor product of the first-order differences Δ_i, i.e.

$$\Delta_i Q_{\boldsymbol{\ell}} := \begin{cases} Q_{\boldsymbol{\ell}} - Q_{\boldsymbol{\ell}-\mathbf{e}_i} , & \text{if } \ell_i \geq 1 \\ Q_{\boldsymbol{\ell}} , & \text{if } \ell_i = 0 \end{cases} , \tag{D.0.4}$$

where $\mathbf{e}_i$ denotes a vector of zeroes, with a single one at the ith component. For an example of $\Delta Q_{\boldsymbol{\ell}}$ when $s = 2$, see [99, pg.4] or [172, pg.5]. An example for $s = 3$ is given below:

$$\begin{aligned}
\Delta Q_{[\ell_1, \ell_2, \ell_3]} &= \Delta_3 \Delta_2 \Delta_1 Q_{[\ell_1, \ell_2, \ell_3]} = \Delta_3 \Delta_2 \left(Q_{[\ell_1, \ell_2, \ell_3]} - Q_{[\ell_1-1, \ell_2, \ell_3]} \right) \\
&= \Delta_3 \left[\left(Q_{[\ell_1, \ell_2, \ell_3]} - Q_{[\ell_1-1, \ell_2, \ell_3]} \right) - \left(Q_{[\ell_1, \ell_2-1, \ell_3]} - Q_{[\ell_1-1, \ell_2-1, \ell_3]} \right) \right] \\
&= \left[\left(Q_{[\ell_1, \ell_2, \ell_3]} - Q_{[\ell_1-1, \ell_2, \ell_3]} \right) - \left(Q_{[\ell_1, \ell_2-1, \ell_3]} - Q_{[\ell_1-1, \ell_2-1, \ell_3]} \right) \right] \\
&\quad - \left[\left(Q_{[\ell_1, \ell_2, \ell_3-1]} - Q_{[\ell_1-1, \ell_2, \ell_3-1]} \right) - \left(Q_{[\ell_1, \ell_2-1, \ell_3-1]} - Q_{[\ell_1-1, \ell_2-1, \ell_3-1]} \right) \right] .
\end{aligned}$$

Note that we now require (up to) 2^s estimates of Q, over a variety of meshes.

Once these changes in levels and differencing have been established, the methodology follows MLMC. Consider the full tensor index set $\mathcal{I}$, defined in (D.0.1), then we can take advantage of
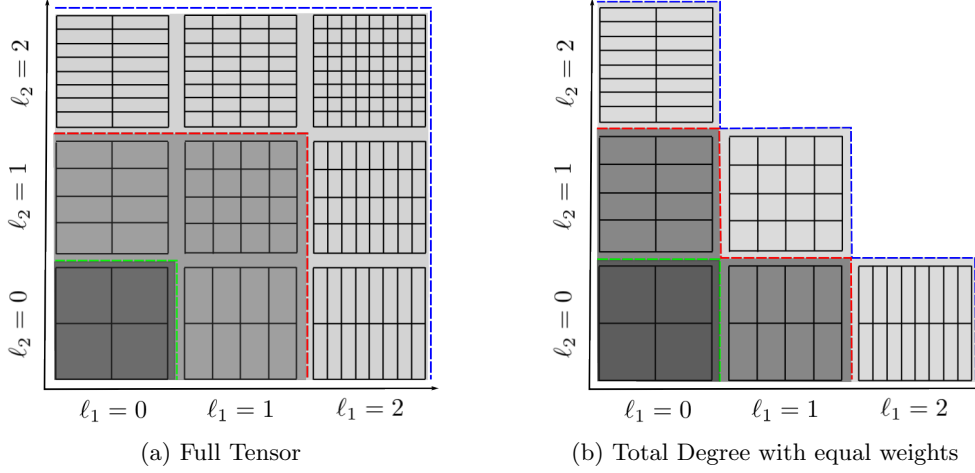
Figure D-1: The allowed $s = 2$ dimensional meshes for two types of index sets $\mathcal{I}$. The selected meshes for increasing L are those to the bottom left of the green, red and blue dotted lines, respectively.

linearity of the expectation by writing

$$\mathbb{E}[Q_h] = \sum_{\boldsymbol{\ell} \in \mathcal{I}} \mathbb{E}[Y_{\boldsymbol{\ell}}] , \quad \text{with } Y_{\boldsymbol{\ell}} := \Delta Q_{\boldsymbol{\ell}} , \tag{D.0.5}$$

and $Q_{\boldsymbol{\ell}} := 0$ when $\ell_i = -1$, for any $i = 1, \cdots, s$. Each $\mathbb{E}[Y_{\boldsymbol{\ell}}]$ is then estimated individually. For example, using a standard MC estimator with $N_{\boldsymbol{\ell}}$ samples to estimate $\mathbb{E}[Y_{\boldsymbol{\ell}}]$, leads to the MIMC estimator

$$\widehat{Q}_h^{MIMC} := \sum_{\boldsymbol{\ell} \in \mathcal{I}} \widehat{Y}_{\boldsymbol{\ell}}^{MC} = \sum_{\boldsymbol{\ell} \in \mathcal{I}} \frac{1}{N_{\boldsymbol{\ell}}} \sum_{n=1}^{N_{\boldsymbol{\ell}}} Y_{\boldsymbol{\ell}} \left(\mathbf{Z}^{(\boldsymbol{\ell},n)} \right) , \tag{D.0.6}$$

where $\{\mathbf{Z}^{(\boldsymbol{\ell},n)}\}_{n=1}^{N_{\boldsymbol{\ell}}}$ denote samples of the random field corresponding to i.i.d. MC samples at $\boldsymbol{\ell} \in \mathcal{I}$, chosen independently from samples on other levels. As with (2.4.31), the independence allows us to re-write the MSE (2.4.2) for the MIMC estimator (D.0.6) as:

$$e(\widehat{Q}_h^{MIMC})^2 = (\mathbb{E}[Q - Q_h])^2 + \sum_{\boldsymbol{\ell} \in \mathcal{I}} N_{\boldsymbol{\ell}}^{-1} \mathbb{V}[Y_{\boldsymbol{\ell}}] . \tag{D.0.7}$$

Moreover, we can again use the method of Lagrange Multipliers to show the sequence $\{N_{\boldsymbol{\ell}}\}$, defined by [99, eq.(10)]

$$N_{\boldsymbol{\ell}} = \left\lceil 2\epsilon^{-2} \left(\sum_{\boldsymbol{\ell} \in \mathcal{I}} \sqrt{\mathbb{V}[Y_{\boldsymbol{\ell}}] \mathcal{C}(Y_{\boldsymbol{\ell}})} \right) \sqrt{\frac{\mathbb{V}[Y_{\boldsymbol{\ell}}]}{\mathcal{C}(Y_{\boldsymbol{\ell}})}} \right\rceil , \quad \text{for all } \boldsymbol{\ell} \in \mathcal{I} , \tag{D.0.8}$$

minimises the cost of (D.0.6) whilst ensuring the sampling error $\sum_{\boldsymbol{\ell} \in \mathcal{I}} N_{\boldsymbol{\ell}}^{-1} \mathbb{V}[\Delta Q_{\boldsymbol{\ell}}] \le \epsilon^2/2$. The proof is analogous to (2.4.32). Furthermore, as in the MLMC case, we note that each of the terms in $Y_{\boldsymbol{\ell}}$ are computed using the same realisation(s) $\mathbf{Z}^{(\boldsymbol{\ell},n)}$, for $n = 1, \cdots, N_{\boldsymbol{\ell}}$.

To bound the computational ϵ-cost for the MIMC estimator, we make the following assumptions (Assumption D.0.1) on the expectation, variance and cost of the difference process $Y_{\boldsymbol{\ell}} = \Delta Q_{\boldsymbol{\ell}}$. We note that, for simplicity in the presentation, we will assume that

$$h_{\ell_i} = \left(\frac{1}{2} \right)^{\ell_i} , \quad \text{for all } \ell_i = 0, \cdots, L , \quad \text{and for all } i = 1, \cdots, s . \tag{D.0.9}$$

Assumption D.0.1 *Assume there exists vectors* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in (\mathbb{R}^+)^s$ *and* $\beta_i \leq 2\alpha_i$, *for all* $i = 1, \cdots, s$, *such that*

$$|\mathbb{E}[\Delta Q_{\boldsymbol{\ell}}]| = \mathcal{O}\left(\prod_{i=1}^{s}\left(\frac{1}{2}\right)^{\ell_i \alpha_i}\right) , \qquad (D.0.10)$$

$$\mathbb{V}[\Delta Q_{\boldsymbol{\ell}}] = \mathcal{O}\left(\prod_{i=1}^{s}\left(\frac{1}{2}\right)^{\ell_i \beta_i}\right) , \qquad (D.0.11)$$

$$\mathcal{C}(\Delta Q_{\boldsymbol{\ell}}) = \mathcal{O}\left(\prod_{i=1}^{s}\left(\frac{1}{2}\right)^{-\ell_i \gamma_i}\right) . \qquad (D.0.12)$$

These assumptions are analogous to (2.4.3), (2.4.33) and (2.4.4), although we emphasise that the multi-index parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are different to the corresponding α, β, γ parameters in MLMC. Moreover, note that Assumption D.0.1 implies that MIMC requires *mixed regularity* of a certain order, whereas MLMC requires only ordinary regularity of the same order [99].

Under the assumptions (D.0.10) – (D.0.12), and with $\mathcal{I}$ taken as the full tensor index set (D.0.1), [99, Thm 2.1] proved the following theoretical complexity estimate. We also note that the proceeding theory also applies to the more general case where Assumptions D.0.1 hold for an unbiased estimator of $\Delta Q_{\boldsymbol{\ell}}$.

Theorem D.0.2 *Consider Assumption D.0.1 for parameters* $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ *and* $\boldsymbol{\ell} \in \mathcal{I}(L)$, *the full tensor index set defined in* (D.0.1). *Further assume that*

$$\sum_{i=1}^{s} \frac{\min\{\beta_i, \gamma_i\}}{\alpha_i} < 2 .$$

Then, for any $\epsilon > 0$, *there exists an* $L \sim \log(\epsilon^{-1})$ *and a sequence* $\{N_{\boldsymbol{\ell}}\}_{\boldsymbol{\ell} \in \mathcal{I}}$ *such that* $e\left(\widehat{Q}_h^{MIMC}\right) \leq \epsilon^2$ *and*

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MIMC})\right] = \mathcal{O}\left(\epsilon^{-2-\sum_{i=1}^{s} r_i}\right) , \qquad (D.0.13)$$

where for each $i = 1, \cdots, s$ *we define*

$$r_i := \begin{cases} 0 , & \text{if } \beta_i \geq \gamma_i \\ \frac{\gamma_i - \beta_i}{\alpha_i} , & \text{if } \beta_i < \gamma_i \end{cases} .$$

For each $i \in \{1, \cdots, s\}$ *such that* $\beta_i = \gamma_i$, *there is an additional* $\log(\epsilon^{-1})^2$ *factor on the right hand side of* (D.0.13).

As we have already mentioned, the choice of index set $\mathcal{I}$ is fundamental to the efficiency of MIMC. In general the full tensor index set is *not the optimal index set*, and the cost in (D.0.13) can be higher than the cost of the MLMC estimator. Subsequently, the paper [99] analytically computes a (quasi-)optimal index set (following the knapsack methodology in [153]), under the assumptions (D.0.10) – (D.0.12). They show that a (quasi-)optimal index set is of (weighted) total degree type ((D.0.2)), i.e.

$$\mathcal{I}(L) := \left\{\boldsymbol{\ell} \in \mathbb{N}_0^s \mid \boldsymbol{\delta} \cdot \boldsymbol{\ell} \leq L \right\} , \qquad (D.0.14)$$

where $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_s] \in (0, 1]^s$ denotes a vector of (quasi-optimal and positive) weights, defined by

$$\delta_i := c_\delta \log(2)\left(\alpha_i + \frac{\gamma_i - \beta_i}{2}\right) , \quad \text{for } i = 1, \cdots, s , \qquad (D.0.15)$$

where c_δ denotes a normalisation constant to ensure $\sum_{i=1}^s \delta_i = 1$. The weights are positive and normalisable due to the assumption that $\gamma_i > 0$ and $\beta_i \leq 2\alpha_i$ (from Assumption D.0.1), for all $i = 1, \cdots, s$.

For simplicity of presentation in the remaining chapter, we will assume that the set $\{i \in \{1, \cdots, s\} \mid \delta_i^{-1}\alpha_i = \min_j \delta_j^{-1}\alpha_j\}$ contains only a single element, i.e. $\mathfrak{e} = 1$ (using the $\mathfrak{e}$ notation as defined in [99, eq.(34a)]). We refer the reader to [99, eq.(34) and eq.(36)] for further details.

Remark D.0.3 *The $L = L^{MIMC-QO}$ used in* (D.0.14) *is only proportional to the $L = L^{MLMC}$ used in MLMC (and MIMC equipped with full tensor set* (D.0.1)*). In particular, $L^{MIMC-QO} \sim \kappa L^{MLMC}$, where we define*

$$\kappa := \max_{i \in \{1, \cdots, s\}} \delta_i \left(\log(2)\alpha_i\right)^{-1} .$$

For the case $\alpha_i = \alpha_j$, $\beta_i = \beta_j$ and $\gamma_i = \gamma_j$, for all $i, j = 1, \cdots, s$ (the so-called isotropic case), then $L^{MIMC-QO} \sim s^{-1}L^{MLMC}$.

In [99, Lemma 2.2], a theoretical complexity estimate is proven using total degree sets for arbitrary choices of $\boldsymbol{\delta}$, i.e. (D.0.2), but we do not give details here. Subsequently, a complexity estimate for the (quasi)-optimal index set (D.0.14) is proven in [99, Thm 2.2] and we present it below.

Theorem D.0.4 *Consider Assumption D.0.1 for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\ell} \in \mathcal{I}(L)$, the particular weighted total degree index set defined in* (D.0.14). *Then, for any $\epsilon > 0$, there exists an $L \sim \kappa \log(\epsilon^{-1})$ (with κ defined in Remark D.0.3) and a sequence $\{N_{\boldsymbol{\ell}}\}_{\boldsymbol{\ell} \in \mathcal{I}}$ such that $e\left(\widehat{Q}_h^{MIMC}\right) \leq \epsilon^2$ and*

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MIMC})\right] = \mathcal{O}\left(\epsilon^{-2-2\max\{0,\chi\}} \log(\epsilon^{-1})^{\mathfrak{p}}\right) , \qquad (D.0.16)$$

where we define

$$\chi := \frac{1}{2} \max_{i \in \{1, \cdots, s\}} \left(\frac{\gamma_i - \beta_i}{\alpha_i}\right) ,$$

and where $\mathfrak{p}$ is given in Table D.1 - which uses the following parameters $\mathfrak{x}$, $\mathfrak{y}$ and $\mathfrak{z}$:

$$\mathfrak{x} := \left| \{i \in \{1, \cdots, s\} \mid (2\alpha_i)^{-1}(\gamma_i - \beta_i) = \chi\} \right| ,$$

$$\mathfrak{y} := \left| \{i \in \{1, \cdots, s\} \mid \beta_i = \gamma_i\} \right| , \qquad \mathfrak{z} := \min_{i \in \{1, \cdots, s\}} \gamma_i^{-1} (2\alpha_i - \beta_i) \geq 0 .$$

χ	$\chi < 0$	$\chi = 0$		$\chi > 0$		
$\mathfrak{z}$	$\mathfrak{z} \in \mathbb{R}^+$	$\mathfrak{z} = 0$		$\mathfrak{z} > 0$	$\mathfrak{z} = 0$	$\mathfrak{z} > 0$
s	$s \in \mathbb{N}$	$s \leq 2$	$s > 2$	$s \in \mathbb{N}$	$s \in \mathbb{N}$	$s \in \mathbb{N}$
$\mathfrak{p}$	$2\mathfrak{y}$	$2\mathfrak{y}$	$2\mathfrak{y} + s - 3$	$2\mathfrak{y}$	$s - 1 + 2(\mathfrak{x}-1)(\chi+1)$	$2(\mathfrak{x}-1)(\chi+1)$

Table D.1: The value of $\mathfrak{p}$, given s and the parameters χ, $\mathfrak{x}$, $\mathfrak{y}$ and $\mathfrak{z}$ defined in Theorem D.0.4.

We can immediately see the benefit of using (D.0.14) over the full tensor set (D.0.1). Specifically, the rate of the computational ϵ-cost for the full tensor set *grows additively for each $i = 1, \cdots, s$ where $\gamma_i - \beta_i \geq 0$* (see the sum in the exponent in (D.0.13)). In comparison, the rate for the quasi-optimal set (D.0.14) is only affected by the 'worst direction', i.e. $\arg\max_{i \in \{1, \cdots, s\}}(\gamma_i - \beta_i)/\alpha_i$.

However, there are a number of subtleties to the use of the quasi-optimal index set (D.0.14) within a MIMC estimator. The first was previously discussed in Remark D.0.3 and we discuss a second in Remark D.0.5.

Remark D.0.5 *The equality in* (D.0.5) *no longer holds for* $\mathcal{I} = \mathcal{I}_{QO}$, *the quasi-optimal set* (D.0.14). *This means that the MIMC estimator associated with* $\mathcal{I}_{QO}$ *is a biased estimator.*

For example, consider the case where $L = L^{MIMC-QO} = L^{MLMC}$ *and let* $\mathcal{I}_{FT}$ *denote the full-tensor set* (D.0.1). *Then, the additional error is given by*

$$\sum_{\boldsymbol{\ell} \in (\mathcal{I}_{FT}(L) \backslash \mathcal{I}_{QO}(L))} \mathbb{E}\left[Y_{\boldsymbol{\ell}}\right] \; \approx \; \sum_{\boldsymbol{\ell} \in (\mathcal{I}_{FT}(L) \backslash \mathcal{I}_{QO}(L))} \widehat{Y}_{\boldsymbol{\ell}}^{MC} \; .$$

Therefore, we must choose $L = L^{MIMC-QO} \geq L^{MLMC}$ *sufficiently large to ensure that the bias constraint* $(\mathbb{E}[Q - Q_L])^2 \leq \epsilon^2/2$ *is satisfied for* $\mathcal{I}_{QO}$. *The required* $L^{MIMC-QO}$, *for general total degree sets* (D.0.2), *is proven in [99, Lemma 2.2] (this is where the definition of* κ *in Remark D.0.3 comes from).*

However, $L^{MIMC-QO}$ *is difficult to estimate in practice and hence [99, pg.23] proposes a heuristic way (for a MIMC estimator associated with any index set* $\mathcal{I}$) *of ensuring that the bias is sufficiently small - by estimating*

$$\left| \mathbb{E}\left[Q - Q_L\right] \right| \; \approx \; \left| \sum_{\boldsymbol{\ell} \in \partial\mathcal{I}(L)} \widehat{Y}_{\boldsymbol{\ell}}^{MC} \right| \; = \; \left| \sum_{\boldsymbol{\ell} \in \partial\mathcal{I}(L)} \frac{1}{N_{\boldsymbol{\ell}}} \sum_{n=1}^{N_{\boldsymbol{\ell}}} Y_{\boldsymbol{\ell}}\left(\mathbf{Z}^{(\boldsymbol{\ell},n)}\right) \right| \; , \tag{D.0.17}$$

and iteratively increasing L *(by one) until the bias constraint is satisfied (for a given* ϵ). *Here* $\partial\mathcal{I}$ *denotes the upper[2] boundary of the index set.*

The lack of equality in (D.0.5) and the sparsity of $\mathcal{I}_{\mathrm{QO}}$ compared with $\mathcal{I}_{\mathrm{FT}}$, could lead one to worry that we require $L^{MIMC-QO} \gg L^{MLMC}$ and that no real cost gains can be achieved. However, it is reasonable (under certain assumptions on Q) to expect otherwise, if we study the theoretical foundations of *sparse grids* [189, 76, 39]. We will briefly explain below, in the context of interpolating some function Q.

The initial theory into sparse grids compared the standard tensor product construction (e.g. see [39, eq.(3.17)] and the surrounding discussion) with the Smolyak sparse grid construction [189] (e.g. [39, eq.(3.61)][3]).

The standard tensor product selects $\boldsymbol{\ell}$ such that $|\boldsymbol{\ell}|_{\infty} := \max\{|\ell_i|\}_{i=1}^s \leq L$ - that is, it is comparable to the full tensor set (D.0.1). This comparison can be also be made intuitively by comparing the left plot in Figure D-1 with [39, Fig. 3.4]. It can easily be shown that the associated cost with this set grows exponentially with s, i.e. $\mathcal{O}(L^s)$ if we consider the isotropic case with $\mathcal{O}(L)$ terms in each direction. Moreover, provided the function Q has integrable mixed derivatives (with respect to its argument) up to order r, then the convergence of the error can be shown to be $\mathcal{O}(L^{-r})$ [76] (see also [39] for a simpler proof when $r = 2$).

On the other hand, the standard sparse grid construction selects $\boldsymbol{\ell}$ such that $|\boldsymbol{\ell}|_1 := \sum_{i=1}^s |\ell_i| \leq (L + s)$ - that is, it is comparable to the total degree set (D.0.14) with equal weights. This comparison can be also be made intuitively by comparing the right plot in Figure D-1 with [39, Fig. 3.5]. In this case, it can be shown that the associated cost is $\mathcal{O}\left(L \log(L)^{(d-1)(r-1)}\right)$, that is, the cost is almost dimension-independent (except for a logarithmic term). Furthermore, despite the substantial cost reduction (for large d), it can be shown that the convergence of the error, with respect to L, is only affected by a dimension-dependent *logarithmic term*, i.e. $\mathcal{O}\left(L^{-r} \log(L)^{(d-1)(r-1)}\right)$. These results are given in [189, 76], and proven for the case $r = 2$ in [39].

---

[2] [99] refer to it as the 'outer boundary', which we feel is misleading, e.g. it does not generally include $\boldsymbol{\ell} \in \mathcal{I}$, where $\ell_i = 0$, for any $i = 1, \cdots, s$

[3] we will alter the notation in [39] to make it comparable to our notation, i.e. $d \to s$ and $n \to L+1$

Therefore, for our problem of estimating $\mathbb{E}[Q]$ for some random variable Q, it would be reasonable to expect substantial cost gains when using the total degree set (D.0.14) instead of the full tensor set (D.0.1), provided Q has bounded mixed derivatives of a certain order. This forms part of Assumption D.0.1.

Remark D.0.6 *Strictly speaking, s denotes a chosen dimensionality for MIMC, where s can take any value between 1 and s^{det}, the dimension of the deterministic model (e.g. PDE). For the preceding discussion we assumed $s = s^{det}$. For the case $s < s^{det}$, we must select dimensions within the deterministic model for which we want the mesh refinement to be coupled. An extreme example involves the coupling of all s^{det} dimensions (i.e. $s = 1$), this is simply multilevel Monte Carlo. Standard Monte Carlo is then a further extension where the 'differencing' operator is replaced by the identity operator.*

Appendix E

Numerical Analysis of the Transport Equation

In this appendix, we will prove the results that were stated without proof in Chapters 3 and 4. In particular, we prove the bound on the function Ein($\cdot$) used in Theorem 3.2.7, we prove that the left and right sides of (3.3.27) are equivalent, we prove a simple bound on a norm of the operator $\left(I - \mathcal{P}^h\right)$ i.e. (3.3.9), and finally that the measure of the set of 'bad' samples Ω_{bad} goes to zero as the mesh size h goes to zero - which is discussed in Section 4.1.1.

E.1   Upper bound on Ein(.)

Lemma E.1.1 *For Ein($\cdot$) defined by (3.2.11) and [1, footnote on pg.228], i.e.*

$$Ein(z) \;=\; \int_0^z \frac{1}{t} \left(1 - \exp(-t)\right) \; dt \;, \quad \textit{for all } z > 0 \;,$$

then,

$$0 \;\leq\; Ein(z) \;\leq\; z \;.$$

Proof. By the Leibniz integral rule, the first two derivatives of Ein are

$$\text{Ein}'(z) = z^{-1}(1 - \exp(-z)) \;, \quad \text{and} \quad \text{Ein}''(z) = z^{-2} \exp(-z) \left(z - \exp(z) + 1\right) \;,$$

which are differentiable functions on $(0, \infty)$. Moreover, since

$$\lim_{z \to 0} \text{Ein}'(z) \;=\; 1 \;,$$

where for small z, we use $\exp(-z) \approx 1 - z$, and

$$\lim_{z \to \infty} \text{Ein}'(z) \;=\; 0 \;,$$

and $\text{Ein}''(z) < 0$, for all $z > 0$, then Ein' must be non-vanishing on $(0, \infty)$. Hence,

$$0 \;\leq\; \text{Ein}'(z) \;\leq\; 1 \;, \quad \text{for all} \;\; z \in (0, \infty) \;.$$

Therefore by the Mean Value Theorem then, for all $z > 0$ there exists $c \in (0, z)$ such that

$$\text{Ein}(z) - \text{Ein}(0) = z\text{Ein}'(c) .$$

The desired result follows since $\text{Ein}(0) = 0$ and $\text{Ein}'(c) \in [0, 1]$, for all $c > 0$. ∎

E.2 Equivalence of the Stability Operator

Lemma E.2.1 *Let $\mathcal{K}^{h,N}$ and $\mathcal{P}^h$ be the operators defined in Chapter 3. Then, the following equality holds*

$$\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1} = I + \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S . \tag{E.2.1}$$

Proof. Consider the following trivial statement, and then manipulate:

$$
\begin{aligned}
& \mathcal{K}^{h,N} - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\mathcal{K}^{h,N} &&= \mathcal{K}^{h,N} - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\mathcal{K}^{h,N} \\
\Rightarrow\quad & \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right) &&= \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)\mathcal{K}^{h,N} \\
\Rightarrow\quad & \mathcal{K}^{h,N} &&= \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)\mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1} \\
\Rightarrow\quad & \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S &&= \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)\mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S \\
\Rightarrow\quad & \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right) - I &&= -\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)\mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S \\
\Rightarrow\quad & I - \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1} &&= -\mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S \\
\Rightarrow\quad & \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1} &&= I + \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S ,
\end{aligned}
$$

which is the result.

∎

E.3 Upper bound on $(I - \mathcal{P}^h)$

Lemma E.3.1 *Recall the definition of $\mathcal{P}^h$ from Section 3.1.3. Then, $\left(I - \mathcal{P}^h\right) : C_{pw}^\xi \mapsto C_{pw}$ with bound*

$$\|(I - \mathcal{P}^h)g\|_\infty \leq h^\xi \|g\|_{\xi,pw} ,$$

for any $g \in \mathrm{C}_{pw}^\xi$, with $0 < \xi \leq 1$.

Proof. Re-write

$$
\begin{aligned}
\|(I - \mathcal{P}^h)g\|_\infty &\leq \max_j \sup_{x \in I_j} |g(x) - g(x_{j-1/2})| \\
&\leq \max_j \sup_{x \in I_j} |x - x_{j-1/2}|^\xi \frac{|g(x) - g(x_{j-1/2})|}{|x - x_{j-1/2}|^\xi} \\
&\leq h^\xi \|g\|_{\xi,pw} ,
\end{aligned}
$$

where $h := \max_j h_j$ and we define the piecewise Hölder norm, $\|\cdot\|_{\xi,pw}$, in Section 3.1. The proof that $(I - \mathcal{P}^h)$ maps into C_{pw} follows since $\mathcal{P}^h : C_{pw}^\xi \mapsto C_{pw}$. ∎

E.4 Measure of Ω_{bad}, with respect to h

We will now justify the comment we made, just above Theorem 4.1.3, that "Due to Theorem 4.1.1(a) the measure of the set Ω_{bad} converges to 0 in the limit, as $h \to 0$". That is: Show that

$$\lim_{h \to 0} \mathbb{P}(\Omega_{bad}(h)) \ \to \ 0 \ , \tag{E.4.1}$$

where we note that we can partition our sample space Ω by

$$\Omega \ = \ \{\omega \in \Omega \mid \mathcal{R}_3\left(\sigma(\omega, \cdot), \sigma_S(\omega, \cdot)\right) \leq h^{-\eta}\} \cup \underbrace{\{\omega \in \Omega \mid \mathcal{R}_3\left(\sigma(\omega, \cdot), \sigma_S(\omega, \cdot)\right) > h^{-\eta}\}}_{=: \ \Omega_{bad}(h)} \ .$$

The $\mathcal{R}_3(\sigma, \sigma_S)$ is defined in (3.3.28) and we have proven that $\mathcal{R}_3(\sigma, \sigma_S) \in L_p(\Omega)$, for all $p \in [1, \infty)$, in Theorem 4.1.2. Moreover, for simplicity we again assume (4.1.4), i.e.

$$N \ = \ N(h) \ = \ \max\left\{\lceil ch^{-\min\{1, \eta\}}\rceil \ , \ 4\right\} \ ,$$

for some constant $c > 0$ independent of h and ω. See also (4.1.5).

Let us consider the more general problem, where we have a non-empty set X, an associated measure μ and a function $f \in L_1(X)$. Consider the following problem: Show that

$$\lim_{H \to \infty} \mu\left(\{x \in X \mid f(x) > H\}\right) \ \to \ 0 \ , \tag{E.4.2}$$

which is equivalent to showing that

$$\lim_{H \to \infty} \mu\left(\{x \in X \mid f(x) \leq H\}\right) \ \to \ \mu(X) \ . \tag{E.4.3}$$

Then, the problem (E.4.1) is a specific case of problem (E.4.3), where $X = \Omega$, $\mu = \mathbb{P}$, $f = \mathcal{R}_3(\sigma, \sigma_S)$ and $H = h^{-\eta}$. We will now prove (E.4.3) holds.

Since $f \in L_1(X)$, the function f must be almost surely finite. That is, there exists a constant $M < \infty$ and a set $F \subset X$ with zero-measure, i.e. $\mu(F) = 0$, such that

$$f(x) \ \leq \ \begin{cases} M \ , & \text{for all } x \in (X \setminus F) \ , \\ \infty \ , & \text{otherwise} \ . \end{cases} \tag{E.4.4}$$

Hence, we have the trivial result that

$$\begin{aligned} \mu(X \setminus F) \ &= \ \mu\left(\{x \in X \mid f(x) \leq M\}\right) \\ &\leq \ \lim_{H \to 0} \mu\left(\{x \in X \mid f(x) \leq H\}\right) \\ &\leq \ \mu(X) \ , \end{aligned} \tag{E.4.5}$$

since the set is non-decreasing. Moreover,

$$\mu(X \setminus F) \ = \ \mu(X \cap F^c) \ = \ \mu(F^c) \ = \ \mu(X) - \mu(F) \ = \ \mu(X) \ ,$$

and therefore using the bounds either side of (E.4.5),

$$\mu(X) \ \leq \ \lim_{H \to 0} \mu\left(\{x \in X \mid f(x) \leq H\}\right) \ \leq \ \mu(X) \ .$$

Hence, we have proven that (E.4.3) holds - and therefore (E.4.1).

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] B.T. Adams and J.E. Morel. A two-grid acceleration scheme for the multigroup SN equations with neutron upscattering. *Nucl. Sci. Eng*, 115(3):253–264, 1993.

[3] M.L. Adams and E.W. Larsen. Fast iterative methods for discrete-ordinates particle transport calculations. *Prog. Nucl. Energy*, 40(1):3–159, 2002.

[4] T. Akçaoğlu, M. Tokyay, and T. Çelik. Effect of coarse aggregate size and matrix quality on ITZ and failure behavior of concrete under uniaxial compression. *Cement Concrete Comp.*, 26(6):633–638, 2004.

[5] R.E. Alcouffe. Diffusion synthetic acceleration methods for the diamond-differenced discrete-ordinates equations. *Nucl. Sci. Eng*, 64(2):344–355, 1977.

[6] D.L. Allaix and V.I. Carbone. Discretization of 2d random fields: A genetic algorithm approach. *Eng. Struct.*, 31(5):1111–1119, 2009.

[7] E.J. Allen, H.D. Victory Jr, and K. Ganguly. On the convergence of finite-differenced multi-group, discrete-ordinates methods for anisotropically scattered slab media. *SIAM J. Numer. Anal.*, 26(1):88–106, 1989.

[8] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[9] M. Asadzadeh. Analysis of a fully discrete scheme for neutron transport in two-dimensional geometry. *SIAM J. Numer. Anal.*, 23(3):543–561, 1986.

[10] M. Asadzadeh. Lp and eigenvalue error estimates for the discrete ordinates method for two-dimensional neutron transport. *SIAM J. Numer. Anal.*, 26(1):66–87, 1989.

[11] M. Asadzadeh. A finite element method for the neutron transport equation in an infinite cylindrical domain. *SIAM J. Numer. Anal.*, 35(4):1299–1314, 1998.

[12] M. Asadzadeh and L. Thevenot. *On discontinuous Galerkin and discrete ordinates approximations for neutron transport equation and the critical eigenvalue.* Department of Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, 2009.

[13] J.R. Askew. A characteristics formulation of the neutron transport equation in complicated geometries. Technical report, United Kingdom Atomic Energy Authority, 1972.

[14] Portland Cement Association. Portland cement association. `www.cement.org`. Accessed: 2017-12-21.

[15] K.E. Atkinson. *An introduction to numerical analysis*. John Wiley & Sons, 2008.

[16] D. Ayres and M.D. Eaton. Uncertainty quantification in nuclear criticality modelling using a high dimensional model representation. *Ann. Nucl. Energy*, 80:379–402, 2015.

[17] D. Ayres, S. Park, and M.D. Eaton. Propagation of input model uncertainties with different marginal distributions using a hybrid polynomial chaos expansion. *Ann. Nucl. Energy*, 66:1–4, 2014.

[18] D.A.F. Ayres, M.D. Eaton, A.W. Hagues, and M.M.R. Williams. Uncertainty quantification in neutron transport with generalized polynomial chaos using the method of characteristics. *Ann. Nucl. Energy*, 45:14–28, 2012.

[19] K.I. Babenko. Approximation by trigonometric polynomials in a certain class of periodic functions of several variables. In *Dokl. Akad. Nauk*, volume 132, pages 982–985. Russian Academy of Sciences, 1960.

[20] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034, 2007.

[21] I. Babuška and M. Suri. On locking and robustness in the finite element method. *SIAM J. Numer. Anal.*, 29(5):1261–1293, 1992.

[22] F. Ballani, D.J. Daley, and D. Stoyan. Modelling the microstructure of concrete with spherical grains. *Comput. Mater. Sci.*, 35(4):399–407, 2006.

[23] A.R. Barron and C.H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Stat.*, pages 1347–1369, 1991.

[24] D.E.G. Barroso. A numerical solution of the time-dependent neutron transport equation using the characteristic method. Applications to ICF and to hybrid fission-fusion systems. *Preprint arXiv:1704.02861*, 2017.

[25] A. Barth and F.G. Fuchs. Uncertainty quantification for linear hyperbolic equations with stochastic process or random field coefficients. *Appl. Numer. Math.*, 121:38–51, 2017.

[26] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.

[27] G.I. Bell and S. Glasstone. *Nuclear Reactor Theory*, volume 252. Van Nostrand Reinhold, New York, 1970.

[28] R. Bellman. *Dynamic Programming*. Courier Corporation, 2013.

[29] S.E. Benzley, E. Perry, K. Merkley, B. Clark, and G. Sjaardama. A comparison of all hexagonal and all tetrahedral finite element meshes for elastic and elasto-plastic analysis. In *Proceedings, 4th international meshing roundtable*, volume 17, pages 179–191. Sandia National Laboratories Albuquerque, NM, 1995.

[30] A. Bespalov, C.E. Powell, and D. Silvester. A priori error analysis of stochastic Galerkin mixed approximations of elliptic PDEs with random data. *SIAM J. Numer. Anal.*, 50(4):2039–2063, 2012.

[31] C. Bierig and A. Chernov. Approximation of probability density functions by the multilevel Monte Carlo maximum entropy method. *J. Comput. Phys.*, 314:661–681, 2016.

[32] J. Blake. *Domain Decomposition Methods for Nuclear Reactor Modelling with Diffusion Acceleration*. PhD thesis, University of Bath, 2016.

[33] T.E. Booth. Monte Carlo variance comparison for expected-value versus sampled splitting. *Nucl. Sci. Eng*, 89(4):305–309, 1985.

[34] J.M. Borwein and A.S. Lewis. Convergence of best entropy estimates. *SIAM J. Optim.*, 1(2):191–205, 1991.

[35] F. Brezzi, B. Cockburn, L.D. Marini, and E. Süli. Stabilization mechanisms in discontinuous Galerkin finite element methods. *Comput. Methods Appl. Mech. Eng.*, 195(25-28):3293–3310, 2006.

[36] L.L. Briggs, W.F. Miller Jr, and E.E. Lewis. Ray-effect mitigation in discrete ordinate-like angular finite element approximations in neutron transport. *Nucl. Sci. Eng*, 57(3):205–217, 1975.

[37] R.J. Brissenden and A.R. Garlick. Biases in the estimation of keff and its error by Monte Carlo methods. *Ann. Nucl. Energy*, 13(2):63–83, 1986.

[38] A.G. Buchan, A.A. Calloo, M.G. Goffin, S. Dargaville, F. Fang, C.C. Pain, and I.M. Navon. A POD reduced order model for resolving angular direction in neutron/photon transport problems. *J. Comput. Phys.*, 296:138–157, 2015.

[39] H.J. Bungartz and M. Griebel. Sparse grids. *Act. Num.*, 13:147–269, 2004.

[40] R. Butler, T.J. Dodwell, R.T. Haftka, N.H. Kim, T. Kim, S. Kynaston, and R. Scheichl. Uncertainty quantification of composite structures with defects using multilevel Monte Carlo simulations. In *17th AIAA Non-Deterministic Approaches Conference*, page 1598, 2015.

[41] R.E. Caflisch, W.J. Morokoff, and A.B. Owen. *Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension*. Department of Mathematics, University of California, Los Angeles, 1997.

[42] A.D. Carlson, V.G. Pronyaev, D.L. Smith, N.M. Larson, Z. Chen, G.M. Hale, F.J. Hambsch, E.V. Gai, S.Y. Oh, S.A. Badikov, et al. International evaluation of neutron cross section standards. *Nucl. Data Sheets*, 110(12):3215–3324, 2009.

[43] C. Cavarec, J.F. Perron, D. Verwaerde, and J.P. West. Benchmark calculations of power distribution within assemblies. Technical report, Nuclear Energy Agency, 1994.

[44] M.B. Chadwick, P. Obložinskỳ, M. Herman, N.M. Greene, R.D. McKnight, D.L. Smith, P.G. Young, R.E. MacFarlane, G.M. Hale, S.C. Frankle, et al. ENDF/B-VII. 0: Next generation evaluated nuclear data library for nuclear science and technology. *Nucl. Data Sheets*, 107(12):2931–3060, 2006.

[45] S. Chandrasekhar. *Radiative Transfer*. Courier Corporation, 2013.

[46] B. Chang, T. Manteuffel, S. McCormick, J. Ruge, and B. Sheehan. Spatial multigrid for isotropic neutron transport. *SIAM J. Sci. Comput.*, 29(5):1900–1917, 2007.

[47] J. Charrier, R. Scheichl, and A.L. Teckentrup. Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. *SIAM J. Numer. Anal.*, 51(1):322–352, 2013.

[48] K.A. Cliffe, M.B. Giles, R. Scheichl, and A.L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3, 2011.

[49] B. Cockburn, G.E. Karniadakis, and C.W. Shu. The development of discontinuous Galerkin methods. In *Discontinuous Galerkin Methods*, pages 3–50. Springer, 2000.

[50] B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. *J. Comput. Phys.*, 141(2):199–224, 1998.

[51] N. Collier, A.L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone. A continuation multilevel Monte Carlo algorithm. *BIT*, 55(2):399–432, 2015.

[52] R. Dautray and J. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 6 Evolution Problems II.* Berlin: Springer-Verlag, 2000.

[53] B. Davison and J.B. Sykes. *Neutron transport theory.* Clarendon Press, Oxford, 1957.

[54] S. Dereich and F. Heidenreich. A multilevel Monte Carlo algorithm for lévy-driven stochastic differential equations. *Stoch. Proc. Appl.*, 121(7):1565–1587, 2011.

[55] R.A. DeVore and L.R. Scott. Error bounds for gaussian quadrature and weighted-l^1 polynomial approximation. *SIAM J. Numer. Anal.*, 21(2):400–412, 1984.

[56] J. Dick, R.N. Gantner, Quoc T.L. Gia, and C. Schwab. Higher order quasi-Monte Carlo integration for Bayesian estimation. *Preprint arXiv:1602.07363*, 2016.

[57] J. Dick, F.Y. Kuo, Qu.T. Le Gia, D. Nuyens, and C. Schwab. Higher order QMC petrov–Galerkin discretization for affine parametric operator equations with random field inputs. *SIAM J. Numer. Anal.*, 52(6):2676–2702, 2014.

[58] J. Dick, F.Y. Kuo, and I.H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013.

[59] S. Disney and I.H. Sloan. Lattice integration rules of maximal rank formed by copying rank 1 rules. *SIAM J. Numer. Anal.*, 29(2):566–577, 1992.

[60] S. Dolgov and R. Scheichl. A hybrid alternating least squares–TT Cross algorithm for parametric PDEs. *Preprint arXiv:1707.04562*, 2017.

[61] T. Durduran, R. Choe, W.B. Baker, and A.G. Yodh. Diffuse optics for tissue monitoring and tomography. *Rep. Prog. Phys.*, 73(7):076701, 2010.

[62] M. Eiermann, O.G. Ernst, and E. Ullmann. Computational aspects of the stochastic finite element method. *Comput. Vis. Sci.*, 10(1):3–15, 2007.

[63] M. Eigel, C.J. Gittelson, C. Schwab, and E. Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Eng.*, 270:247–269, 2014.

[64] M. Eigel, M. Pfeffer, and R. Schneider. Adaptive stochastic Galerkin FEM with hierarchical tensor representations. *Numer. Math.*, 136(3):765–803, 2017.

[65] M. Eldred and J. Burkardt. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. In *47th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition*, page 976, 2009.

[66] D. Elfverson, D.J. Estep, F. Hellman, and A. Målqvist. Uncertainty quantification for approximate p-quantiles for physical models with stochastic inputs. *JUQ*, 2(1):826–850, 2014.

[67] D. Elfverson, F. Hellman, and A. Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *JUQ*, 4(1):312–330, 2016.

[68] K.F. Evans. The spherical harmonics discrete ordinate method for three-dimensional atmospheric radiative transfer. *J. Atmos. Sci.*, 55(3):429–446, 1998.

[69] E.D. Fichtl and A.K. Prinja. The stochastic collocation method for radiation transport in random media. *J. Quant. Spectrosc. Radiat. Transfer*, 112(4):646–659, 2011.

[70] Nuclear Power for Everybody. Nuclear power for everybody. `www.nuclear-power.net`. Accessed: 2018-04-09.

[71] M.A. Freitag and A. Spence. Convergence of inexact inverse iteration with application to preconditioned iterative solves. *BIT*, 47(1):27–44, 2007.

[72] B.D. Ganapol. Analytical Benchmarks for Nuclear Engineering Applications. *Nuclear Energy Agency*, 2008.

[73] R.N. Gantner. Higher order quasi-Monte Carlo rules. `www.sam.math.ethz.ch/HOQMC`. Accessed: 2017-12-20.

[74] R.N. Gantner and C. Schwab. Computational higher order quasi-Monte Carlo integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 271–288. Springer, 2016.

[75] Z.G. Ge, Z.X. Zhao, H.H. Xia, Y.X. Zhuang, T.J. Liu, J.S. Zhang, and H.C. Wu. The updated version of chinese evaluated nuclear data library (CENDL-3.1). *J. Korean Phys. Soc.*, 59(2):1052–1056, 2011.

[76] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3-4):209, 1998.

[77] T. Gerstner and M. Griebel. Dimension–adaptive tensor–product quadrature. *Computing*, 71(1):65–87, 2003.

[78] R.G. Ghanem and P.D. Spanos. Stochastic finite element method: Response statistics. In *Stochastic Finite Elements: A Spectral Approach*, pages 101–119. Springer, 1991.

[79] A.D. Gilbert, I.G. Graham, F.Y. Kuo, R. Scheichl, and I.H. Sloan. Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients. *Preprint arXiv:1808.02639*, 2018.

[80] M.B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[81] M.B. Giles. Multilevel Monte Carlo methods. *Act. Num.*, 24:259–328, 2015.

[82] M.B. Giles, T. Nagapetyan, and K. Ritter. Multilevel Monte Carlo approximation of distribution functions and densities. *JUQ*, 3(1):267–295, 2015.

[83] M.B. Giles and C. Reisinger. Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance. *SIAM J. Financial Math.*, 3(1):572–592, 2012.

[84] M.B. Giles and B.J. Waterhouse. Multilevel quasi-Monte Carlo path simulation. *Advanced Financial Modelling, Radon Series on Computational and Applied Mathematics*, pages 165–181, 2009.

[85] L. Gilli, D. Lathouwers, J.L. Kloosterman, T.H.J.J. van der Hagen, A.J. Koning, and D. Rochman. Uncertainty quantification for criticality problems using non-intrusive and adaptive polynomial chaos techniques. *Ann. Nucl. Energy*, 56:71–80, 2013.

[86] T. Goda. Good interlaced polynomial lattice rules for numerical integration in weighted Walsh spaces. *J. Comput. Appl. Math.*, 285:279–294, 2015.

[87] T. Goda and J. Dick. Construction of interlaced scrambled polynomial lattice rules of arbitrary high order. *Found. Comput. Math.*, 15(5):1245–1278, 2015.

[88] G.H. Golub and Q. Ye. Inexact inverse iteration for generalized eigenvalue problems. *BIT*, 40(4):671–684, 2000.

[89] I.G. Graham, F.Y. Kuo, J.A. Nichols, R. Scheichl, C. Schwab, and I.H. Sloan. Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. *Numer. Math.*, 131(2):329–368, 2015.

[90] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *J. Comput. Phys.*, 230(10):3668–3694, 2011.

[91] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Analysis of circulant embedding methods for sampling stationary random fields. *SIAM J. Numer. Anal.*, 56(3):1871–1895, 2018.

[92] I.G. Graham, M.J. Parkinson, and R. Scheichl. Error analysis and uncertainty quantification for the heterogeneous transport equation in slab geometry. *In Preparation*, 2018.

[93] I.G. Graham, M.J. Parkinson, and R. Scheichl. Modern Monte Carlo variants for uncertainty quantification in neutron transport. In *Contemporary Computational Mathematics – A Celebration of the 80th Birthday of Ian Sloan, eds. J. Dick, F.Y. Kuo, and H. Wozniakowski*, pages 455–481. Springer, 2018.

[94] A. Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17. SIAM, 1997.

[95] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[96] J.L. Guermond, G. Kanschat, and J.C. Ragusa. Discontinuous Galerkin for the radiative transport equation. In *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, pages 181–193. Springer, 2014.

[97] A. Haghighat and J.C. Wagner. Monte Carlo variance reduction with deterministic importance functions. *Prog. Nucl. Energy*, 42(1):25–53, 2003.

[98] A.L. Haji-Ali, F. Nobile, L. Tamellini, and R. Tempone. Multi-index stochastic collocation for random PDEs. *Comput. Methods in Appl. Mech. Eng.*, 306:95–122, 2016.

[99] A.L. Haji-Ali, F. Nobile, and R. Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numer. Math.*, 132(4):767–806, 2016.

[100] A.L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone. Optimization of mesh hierarchies in multilevel Monte Carlo samplers. *Stoch. Partial Differ. Equ. Anal. Comput.*, 4(1):76–112, 2016.

[101] A.L. Haji-Ali and R. Tempone. Multilevel and multi-index Monte Carlo methods for McKean-Vlasov equations. *Preprint arXiv:1610.09934*, 2016.

[102] DOE Fundamentals Handbook. Nuclear Physics and Reactor theory. *Washington DC: Department of Energy*, 1993.

[103] S. Heinrich. Multilevel Monte Carlo methods. In *Lect. Notes Comput. Sc.*, pages 58–67. Springer, 2001.

[104] A.A. Hilal. Microstructure of concrete. In *High Performance Concrete Technology and Applications*. InTech, 2016.

[105] H. Hoel, K.J.H. Law, and R. Tempone. Multilevel ensemble Kalman filtering. *SIAM J. Numer. Anal.*, 54(3):1813–1839, 2016.

[106] M. Holtz. *Sparse Grid Quadrature in High Dimensions With Applications in Finance and Insurance*, volume 77. Springer Science & Business Media, 2010.

[107] A. Isotalo. *Computational Methods for Burnup Calculations with Monte Carlo Neutronics*. PhD thesis, Aalto University, 2013.

[108] K. Ivanov. Nuce 521 neutron transport theory lecture notes. Accessed: 2018-07-17.

[109] S. Jin and D. Levermore. The discrete-ordinate method in diffusive regimes. *Transport Theor. Stat.*, 20(5-6):413–439, 1991.

[110] C. Johnson and J. Pitkäranta. Convergence of a fully discrete scheme for two-dimensional neutron transport. *SIAM J. Numer. Anal.*, 20(5):951–966, 1983.

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

[112] G. Katsiolides, E.H. Müller, R. Scheichl, T. Shardlow, M.B. Giles, and D.J. Thomson. Multilevel Monte Carlo and improved timestepping methods in atmospheric dispersion modelling. *J. Comput. Phys.*, 354:320–343, 2018.

[113] D. Kaushik, M. Smith, A. Wollaber, B. Smith, A. Siegel, and W.S. Yang. Enabling high-fidelity neutron transport simulations on petascale architectures. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, page 67. ACM, 2009.

[114] H.B. Keller. On the pointwise convergence of the discrete-ordinate method. *SIAM J. Appl. Math.*, 8(4):560–567, 1960.

[115] B.N. Khoromskij, A. Litvinenko, and H.G. Matthies. Application of hierarchical matrices for computing the Karhunen–Loeve expansion. *Computing*, 84(1-2):49–67, 2009.

[116] A. Klenke. *Probability Theory: A Comprehensive Course.* Springer Science & Business Media, 2013.

[117] R. Kornhuber, C. Schwab, and M.W. Wolf. Multilevel Monte Carlo finite element methods for stochastic elliptic variational inequalities. *SIAM J. Numer. Anal.*, 52(3):1243–1268, 2014.

[118] S. Krumscheid and F. Nobile. Multilevel Monte Carlo approximation of functions. Technical report, École Polytechnique F'ed'erale de Lausanne, 2017.

[119] S. Kucherenko, D. Albrecht, and A. Saltelli. Exploring multi-dimensional spaces: A comparison of latin hypercube and quasi Monte Carlo sampling techniques. *Preprint arXiv:1505.02350*, 2015.

[120] F.Y. Kuo. Quasi-Monte Carlo generating vectors and construction algorithms. `http://web.maths.unsw.edu.au/~fkuo/`. Accessed: 2017-12-20.

[121] F.Y. Kuo, W.T.M. Dunsmuir, I.H. Sloan, M.P. Wand, and R.S. Womersley. Quasi-Monte Carlo for highly structured generalised response models. *Methodol. Comput. Appl.*, 10(2):239, 2008.

[122] F.Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients: a survey of analysis and implementation. *Found. Comput. Math.*, 16(6):1631–1696, 2016.

[123] F.Y. Kuo, R. Scheichl, C. Schwab, I.H. Sloan, and E. Ullmann. Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. *Preprint arXiv:1507.01090*, 2015.

[124] F.Y. Kuo, C. Schwab, and I.H. Sloan. Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(6):3351–3374, 2012.

[125] F.Y. Kuo, C. Schwab, and I.H. Sloan. Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. *Found. Comput. Math.*, 15(2):411–449, 2015.

[126] Y.L. Lai, K.Y. Lin, and W.W. Lin. An inexact inverse iteration for large sparse eigenvalue problems. *Numer. Linear Algebr.*, 4(5):425–437, 1997.

[127] B. Lapeyre, E. Pardoux, and R. Sentis. *Introduction to Monte-Carlo Methods for Transport and Diffusion Equations*, volume 6. Oxford University Press, 2003.

[128] E.W. Larsen and P. Nelson. Finite-difference approximations and superconvergence for the discrete-ordinate equations in slab geometry. *SIAM J. Numer. Anal.*, 19(2):334–348, 1982.

[129] P. L'Ecuyer. Randomized quasi-Monte Carlo: an introduction for practitioners. In *12th International conference on Monte Carlo and quasi-Monte Carlo methods in scientific computing (MCQMC 2016)*, 2016.

[130] V. Lemaire and G. Pagès. Multilevel Richardson–Romberg extrapolation. *Bernoulli*, 23(4A):2643–2692, 2017.

[131] G. Leobacher and F. Pillichshammer. *Introduction to quasi-Monte Carlo integration and applications.* Springer, 2014.

[132] P. Lesaint and P.A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Elsevier, 1974.

[133] E.E. Lewis and W.F. Miller. *Computational Methods of Neutron Transport*. John Wiley and Sons, Inc., and New York, NY, 1984.

[134] G. Longoni. *Advanced Quadrature Sets, Acceleration and Preconditioning techniques for the discrete ordinates method in parallel computing environments*. PhD thesis, University of Florida, 2004.

[135] G.J. Lord, C.E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, 2014.

[136] I. Lux and L. Koblinger. *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*, volume 102. CRC Press, 1991.

[137] R.C. Martin, J.B. Knauer, and P.A. Balo. Production, distribution and applications of Californium-252 neutron sources. *Appl. Radiat. Isot.*, 53(4-5):785–792, 2000.

[138] W.R. Martin. *The Application of the Finite Element Method to the Neutron Transport Equation*. PhD thesis, University of Michigan, 1976.

[139] R.G. McClarren, M.L. Adams, P.A. Vaquer, and C. Strack. The asymptotic drift-diffusion limit of thermal neutrons. *Journal of Computational and Theoretical Transport*, 43(1-7):402–417, 2014.

[140] M.D. McKay, R.J. Beckman, and W.J. Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[141] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Am. Stat. Assoc.*, 44(247):335–341, 1949.

[142] J. Mika. Existence and uniqueness of the solution to the critical problem in neutron transport theory. *Stud. Math.*, 37(3):213–225, 1971.

[143] S. Mishra, C. Schwab, and J. Sukys. Multilevel Monte Carlo finite volume methods for shallow water equations with uncertain topography in multi-dimensions. *SIAM J. Sci. Comput.*, 34(6):B761–B784, 2012.

[144] S. Mishra, C. Schwab, and J. Šukys. Multi-level Monte Carlo finite volume methods for uncertainty quantification of acoustic wave propagation in random heterogeneous layered medium. *J. Comput. Phys.*, 312:192–217, 2016.

[145] G. Monegato and V. Colombo. Product integration for the linear transport equation in slab geometry. *Numer. Math.*, 52(2):219–240, 1988.

[146] Paulo J.M. Monteiro. Scaling and saturation laws for the expansion of concrete exposed to sulfate attack. *P. Natl. Acad. Sci. USA*, 103(31):11467–11472, 2006.

[147] T. Most and C. Bucher. Probabilistic analysis of concrete cracking using neural networks and random fields. *Probabilist. Eng. Mech.*, 22(2):219–229, 2007.

[148] S. Moustafa. *Massively Parallel Cartesian Discrete Ordinates Method for Neutron Transport Simulation*. PhD thesis, Université de Bordeaux, 2015.

[149] I.B. Muhit, S. Haque, and M.R. Alam. Influence of crushed coarse aggregates on properties of concrete. *AJCE*, 1(5):103–106, 2013.

[150] S. Murphy. *Methods for Solving Discontinuous-Galerkin Finite Element Equations with Application to Neutron Transport*. PhD thesis, École Doctorale Mathématiques, Informatique et Télécommunications (Toulouse), 2015.

[151] Y. Nakatsukasa. Approximate and integrate: Variance reduction in Monte Carlo integration via function approximation. *Preprint arXiv:1806.05492*, 2018.

[152] H. Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *B. AM. Math. Soc.*, 84(6):957–1041, 1978.

[153] F. Nobile, L. Tamellini, and R. Tempone. Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: Application to random elliptic PDEs. *Numer. Math.*, 134(2):343–388, 2016.

[154] D. Nuyens and R. Cools. Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Math. Comput*, 75(254):903–920, 2006.

[155] E.J. Nyström. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Math.*, 54(1):185–204, 1930.

[156] B. OMalley, J. Kópházi, M.D. Eaton, V. Badalassi, P. Warner, and A. Copestake. Discontinuous Galerkin spatial discretisation of the neutron transport equation with pyramid finite elements and a discrete ordinate (s_n) angular approximation. *Ann. Nucl. Energy*, 113:526–535, 2018.

[157] A.B. Owen. *Monte Carlo Theory, Methods and Examples*. Unpublished, available online at https://statweb.stanford.edu/ owen/mc/, 2013.

[158] H. Park, D.A. Knoll, and C.K. Newman. Nonlinear acceleration of transport criticality problems. *Nucl. Sci. Eng*, 172(1):52–65, 2012.

[159] M.H. Park and M.V. Tretyakov. A block circulant embedding method for simulation of stationary Gaussian random fields on block-regular grids. *Int. J. Uncertain. Quan.*, 5(6), 2015.

[160] T.E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28(1):133–140, 1991.

[161] T. Piotrowski, D. Tefelski, A. Polański, and J. Skubalski. Monte Carlo simulations for optimization of neutron shielding concrete. *Open Engineering*, 2(2):296–303, 2012.

[162] J. Pitkäranta. Estimates for the derivatives of solutions to weakly singular Fredholm integral equations. *SIAM J. Math. Anal.*, 11(6):952–968, 1980.

[163] J. Pitkäranta and L.R. Scott. Error estimates for the combined spatial and angular approximations of the transport equation for slab geometry. *SIAM J. Numer. Anal.*, 20(5):922–950, 1983.

[164] J.J. Powers, N. George, G.I. Maldonado, and A. Worrall. Report on reactor physics assessment of candidate accident tolerant fuel cladding materials in LWRs. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2015.

[165] C.E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[166] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[167] H.M. Regan, M. Colyvan, and M.A. Burgman. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecol. Appl.*, 12(2):618–628, 2002.

[168] K. Ren. Recent developments in numerical techniques for transport-based medical imaging methods. *Commun. Comput. Phys.*, 8(1):1–50, 2010.

[169] C.H. Rhee and P.W. Glynn. Unbiased estimation with square root convergence for sde models. *Oper. Res.*, 63(5):1026–1043, 2015.

[170] G. Richter. On the order of convergence of the discontinuous Galerkin method for hyperbolic equations. *Math. Comput.*, 77(264):1871–1885, 2008.

[171] P. Robbe, D. Nuyens, and S. Vandewalle. A dimension-adaptive multi-index Monte Carlo method applied to a model of a heat exchanger. *Preprint arXiv:1708.04959*, 2017.

[172] P. Robbe, D. Nuyens, and S. Vandewalle. A multi-index quasi–Monte Carlo algorithm for lognormal diffusion problems. *SIAM J. Sci. Comput.*, 39(5):S851–S872, 2017.

[173] E. Roubin, J.B. Colliat, and N. Benkemoun. Meso-scale modeling of concrete: A morphological description based on excursion sets of random fields. *Comput. Mater. Sci.*, 102:183–195, 2015.

[174] Y. Saad. *Numerical methods for large eigenvalue problems.* Manchester University Press, 1992.

[175] R. Sanchez, L. Mao, and S. Santandrea. Treatment of boundary conditions in trajectory-based deterministic transport methods. *Nucl. Sci. Eng.*, 140(1):23–50, 2002.

[176] R. Sanchez and N.J. McCormick. Review of neutron transport approximations. *Nucl. Sci. Eng.*, 80(4), 1982.

[177] A. Santamarina, D. Bernard, N. Dos Santos, O. Leray, C. Vaglio, and L.C. Leal. Re-estimation of nuclear data and JEFF3.1.1 uncertainty calculations. Technical report, Oak Ridge National Laboratory (ORNL), 2012.

[178] F. Scheben. *Iterative Methods for Criticality Computations in Neutron Transport Theory.* PhD thesis, University of Bath, 2011.

[179] R. Scheichl, A.M. Stuart, and A.L. Teckentrup. Quasi-Monte Carlo and multilevel Monte Carlo methods for computing posterior expectations in elliptic inverse problems. *JUQ*, 5(1):493–518, 2017.

[180] H. Schraube, V. Mares, S. Roesler, and W. Heinrich. Experimental verification and calculation of aviation route doses. *Radiat. Prot. Dosim.*, 86(4):309–315, 1999.

[181] C. Schwab and R.A. Todor. Karhunen–Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122, 2006.

[182] K.L. Scrivener. The microstructure of concrete. *Mat. Sci. Series*, 1989.

[183] S. Seuret and J.L. Véhel. The local Hölder function of a continuous function. *Appl. Comput. Harmonic Anal.*, 13(3):263–276, 2002.

[184] V.M. Shmakov, V.D. Lyutov, and V.F. Dean. Effective cross sections for calculations of criticality of dispersed media. In *Proc. PHYSOR*, pages 7–11, 2000.

[185] R.N. Slaybaugh, T.M. Evans, G.G. Davidson, and Paul P.H. Wilson. Multigrid in energy preconditioner for Krylov solvers. *J. Comput. Phys.*, 242:405–419, 2013.

[186] I.H. Sloan, F.Y. Kuo, and S. Joe. Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM J. Numer. Anal.*, 40(5):1650–1665, 2002.

[187] I.H. Sloan and A. Reztsov. Component-by-component construction of good lattice rules. *Math. Comput.*, 71(237):263–273, 2002.

[188] M.A. Smith, E.E. Lewis, and B.C. Na. Benchmark on deterministic transport calculations without spatial homogenization: A 2-d/3-d mox fuel assembly 3-d benchmark, 2003.

[189] S.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Dokl. Akad. Nauk*, volume 148, pages 1042–1045. Russian Academy of Sciences, 1963.

[190] A. Sood, R.A. Forster, and D.K. Parsons. Analytical benchmark test set for criticality code verification. *Prog. Nucl. Energy*, 42(1):55–106, 2003.

[191] A. Sotoodehnia and R.C. Erdmann. Two-dimensional transport in a slab. *SIAM J. Appl. Math.*, 18(1):160–171, 1970.

[192] J. Spanier and E.M. Gelbard. *Monte Carlo Principles and Neutron Transport Problems.* Courier Corporation, 2008.

[193] M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

[194] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science & Business Media, 2012.

[195] K. Takeuchi. A numerical method for solving the neutron transport equation in finite cylindrical geometry. *Nucl. Sci. & Techn.*, 6(8):466–473, 1969.

[196] B. Tang. Selecting latin hypercubes using correlation criteria. *Statistica Sinica*, pages 965–977, 1998.

[197] A.L. Teckentrup. *Multilevel Monte Carlo methods for elliptic PDEs with random coeffcients.* PhD thesis, University of Bath, 2013.

[198] A.L. Teckentrup, P. Jantsch, C.G. Webster, and M. Gunzburger. A multilevel stochastic collocation method for partial differential equations with random input data. *JUQ*, 3(1):1046–1074, 2015.

[199] A.L. Teckentrup, R. Scheichl, M.B. Giles, and E. Ullmann. Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.*, 125(3):569–600, 2013.

[200] C.P. Thurgood, A. Pollard, and H.A. Becker. The TN quadrature set for the discrete ordinates method. *J. Heat. Transf.*, 117(4):1068–1070, 1995.

[201] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. SIAM, 1997.

[202] N. Tsoulfanidis. *Nuclear energy: selected entries from the encyclopedia of sustainability science and technology.* Springer Science & Business Media, 2012.

[203] S. Ukai. Solution of multi-dimensional neutron transport equation by finite element method. *J. Nucl. Sci. Technol.*, 9(6):366–373, 1972.

[204] H.D. Victory Jr. Convergence of the multigroup approximations for subcritical slab media with applications to shielding calculations. *Adv. Appl. Math.*, 5(3):227–259, 1984.

[205] V.S. Vladimirov. Mathematical problems in the one-velocity theory of particle transport. Technical report, Atomic Energy of Canada Limited, 1963.

[206] J.S. Warsa, T.A. Wareing, J.E. Morel, J.M. McGhee, and R.B. Lehoucq. Krylov subspace iterations for deterministic k-eigenvalue calculations. *Nucl. Sci. Eng*, 147(1):26–42, 2004.

[207] E. Woodcock, T. Murphy, P. Hemmings, and S. Longworth. Techniques used in the GEM code for Monte Carlo neutronics calculations in reactors and other systems of complex geometry. In *Proc. Conf. Applications of Computing Methods to Reactor Problems*, pages 557–579, 1965.

[208] F. Xiong, Y. Xiong, W. Chen, and S. Yang. Optimizing latin hypercube design for sequential sampling of computer experiments. *Eng. Optimiz.*, 41(8):793–810, 2009.

[209] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.

[210] D. Xiu and G.E. Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.

[211] T. Yamamoto. Heterogeneity effect on neutron shielding in borated concrete and Monte Carlo-based cross section homogenization method for particle dispersed media. *Progress in Nuclear Science and Technology (Atomic Energy Society of Japan)*, 4:404–407, 2014.

[212] M. Young. *Orthogonal-Mesh, 3D SN with Embedded 2-D Method of Characteristics for Whole-Core, Pin-Resolved Reactor Analysis.* PhD thesis, University of Michigan, 2016.

[213] B. Zhang, H. Liu, and S. Jin. An asymptotic preserving Monte Carlo method for the multi-species Boltzmann equation. *J. Comput. Phys.*, 305:575–588, 2016.

[214] L. Zhao and G.J. Yun. Probabilistic service life of reinforced concrete structures with randomly distributed corrosion-induced cracks. *Structure and Infrastructure Engineering*, pages 1–15, 2017.