Error Analysis and Uncertainty Quantification for the Heterogeneous Transport Equation in Slab Geometry

Ivan G. Graham[1], Matthew J. Parkinson[1] and Robert Scheichl[1,2]

April 23, 2020

[1] Dept of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.
I.G.Graham@bath.ac.uk
[2] Institut für Angewandte Mathematik, Universität Heidelberg, 69120 Heidelberg, Germany.
r.scheichl@uni-heidelberg.de

Abstract

We present an analysis of multilevel Monte Carlo techniques for the forward problem of uncertainty quantification for the radiative transport equation, when the coefficients (*cross-sections*) are heterogenous random fields. To do this, we first give a new error analysis for the combined spatial and angular discretisation in the deterministic case, with error estimates which are explicit in the coefficients (and allow for very low regularity and jumps). This detailed error analysis is done for the 1D space - 1D angle slab geometry case with classical diamond differencing. Under reasonable assumptions on the statistics of the coefficients, we then prove an error estimate for the random problem in a suitable Bochner space. Because the problem is not self-adjoint, stability can only be proved under a path-dependent mesh resolution condition. This means that, while the Bochner space error estimate is of order $\mathcal{O}(h^\eta)$ for some η, where h is a (deterministically chosen) mesh diameter, smaller mesh sizes might be needed for some realisations. We also show that the expected cost for computing a typical quantity of interest remains of the same order as for a single sample. This leads to rigorous complexity estimates for Monte Carlo and multilevel Monte Carlo: For particular linear solvers, the multilevel version gives up to two orders of magnitude improvement over Monte Carlo. We provide numerical results supporting the theory.

Keywords - Radiative Transport, Neutron Transport, Spatial heterogeneity, Random Coefficients, Error Estimate, Multilevel Monte Carlo

AMS Subject Classifications: 65N12, 65R99, 65C30, 65C05

1 Introduction

The Radiative Transport Equation (RTE) is a physically derived balance equation which models the angular flux ψ of rarefied particles (such as photons or neutrons) in a domain. Generally ψ is a function of position, direction of travel, energy and time (see e.g. [2, 9, 34]). It is assumed that the particles cannot interact with each other and that they travel along straight line paths with some energy until they interact with larger nuclei via absorption, scattering or fission. The rates σ_A, σ_S and σ_F at which these collision events occur are called the *absorption, scattering and fission cross-sections*. The RTE has many applications, for example in radiation shielding, nuclear reactor design [9, 41], astrophysics and optical tomography [17, 39].

In the context of neutron transport, the two main scenarios of interest are the so-called *fixed source problem* and the *criticality problem*. We focus on the former, which concerns the transport and scattering of particles emanating from some fixed source f. In steady state, with constant energy, assuming isotropic scattering and neglecting fission, the problem can be

written as the integro-differential equation:

$$\left[\vec{\Theta}\cdot\nabla+\sigma(\vec{r})\right]\psi(\vec{r},\vec{\Theta}) \;=\; \sigma_S(\vec{r})\phi(\vec{r}) \;+\; f(\vec{r},\vec{\Theta}) \tag{1.1}$$

with independent variables being angle $\vec{\Theta}\in\mathbb{S}^2$ (the unit sphere in $3D$), and position $\vec{r}\in V$ (the physical domain occupied by the reactor), where $\psi(\vec{r},\vec{\Theta})$ is the angular flux, and

$$\phi(\vec{r}) \;:=\; \frac{1}{4\pi}\int_{\mathbb{S}^2}\psi(\vec{r},\vec{\Theta})\,\mathrm{d}\vec{\Theta} \tag{1.2}$$

is the *scalar flux*. Equation (1.1) requires boundary conditions and here we will restrict to the zero incoming flux condition

$$\psi(\vec{r},\vec{\Theta})=0, \quad \text{when} \quad \vec{r}\in\partial V \quad \text{and} \quad \vec{n}(\vec{r})\cdot\vec{\Theta}<0, \tag{1.3}$$

with $\vec{n}(\vec{r})$ denoting the outward normal from V at a point $\vec{r}\in\partial V$. The gradient ∇ is with respect to $\vec{r}$ and the coefficient function $\sigma(\vec{r})$ is the *total cross-section* defined by

$$\sigma(\vec{r}) \;=\; \sigma_A(\vec{r})+\sigma_S(\vec{r})\;, \tag{1.4}$$

where σ_S and σ_A are, respectively, the scattering and absorption cross-sections, both assumed to be non-negative.

In this paper we propose and analyse efficient multilevel Monte Carlo methods for quantifying the effect of uncertainty in the *input data* σ_A, σ_S and f, on the *output variable* ϕ. There is a growing recent interest in this question in the more general context of kinetic equations. 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.

Novel results in this paper. Since their introduction in the context of high-dimensional quadrature and SDEs in mathematical finance [28, 22], multilevel Monte Carlo methods have generated a lot of interest. While uncertainty quantification recently has become a topic of great general interest for the transport equation in both theory and practice e.g. [30, 45], to our knowledge the present work provides the first rigorous analysis of multilevel Monte Carlo methods for this problem. To allow the first results to be established, we make the simplifying assumption of one spatial and one angular dimension, the so-called "slab-geometry" case in reactor theory. We discretise with the classical discrete ordinates method, using a certain Gauss 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 $\phi^{h,N}$.

Our first set of results describes how heterogeneity in the material coefficients manifests itself in the operators underlying the RTE, and consequently in the error estimate for the numerical method. 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 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 $\|\cdot\|_{\eta,pw}$ defined below. Our first error estimate is Theorem 4.10, which shows that there are constants $\mathcal{R},\mathcal{R}'$, both dependent on σ,σ_S such that, when

$$N^{-1} \;+\; h\log N \;+\; h^\eta \;\le\; \mathcal{R}(\sigma,\sigma_S)^{-1}, \tag{1.5}$$

we have the error estimate for the scalar flux:

$$\|\phi-\phi^{h,N}\|_\infty \;\le\; \mathcal{R}'(\sigma,\sigma_S)\left(N^{-1} \;+\; h\log N \;+\; h^\eta\right)\|f\|_{\eta,pw}\;. \tag{1.6}$$

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 σ/σ_S is close to 1 anywhere in the domain or if $\|\sigma\|_\infty\|\sigma^{-1}\|_\infty\to\infty$. The proof

of (1.5), (1.6) is obtained by generalising the theory of the integral equation reformulation of (1.1) to the heterogeneous case; the homogeneous case having been studied in detail in [37]. An overview of the present work was given in [26].

The appearance of the $h \log N$ term in (1.6) reflects the fact that the transport equation in slab geometry has a singularity in its angular dependence (explained in §2). This imposes a compatibility constraint, 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 [37]. However, we will not pursue this further and thus limit our analysis to piecewise Hölder continuous data ($\eta < 1$). To ensure that the spatial and angular errors are equal order, we set $N = N(h) = \lceil h^{-\eta} \rceil$.

Our second set of results then concerns the probabilistic counterpart of (1.6). Here we have to deal with the fact that the deterministic estimate (1.6) is subject to the "mesh resolution condition" (1.5), which in turn arises from the non-self-adjointness of (1.1). In the case of coercive self-adjoint PDEs with random data and Galerkin discretisation (e.g. [13, 42]) 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 (1.5). 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 $h_\omega^{\max}$ denote (the largest) mesh diameter which satisfies the path-dependent criterion (1.5) and then set $h_\omega = \min\{h, h_\omega^{\max}\}$. Then the approximation to $\phi = \phi(\cdot,\omega)$ is taken to be $\Phi^h = \phi^{h_\omega, N(h_\omega)}$. We prove in Theorem 5.6 that

$$\|\phi - \Phi^h\|_{L_p(\Omega;L_\infty)} \;\leq\; C_{p,r}\, h^\eta \, \|f\|_{L_r(\Omega;\mathrm{C}^\eta_{pw})} \;, \tag{1.7}$$

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 5). 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 stability criterion (1.5) 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 that estimate is in $L_p(\Omega)$, for some $p > 1$, then the third main result of this paper in Lemma 5.8 is that

$$\mathbb{E}[\mathcal{C}(\Phi^h)] \;=\; \mathcal{O}\left(h^{-\gamma}\right) \;, \tag{1.8}$$

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 (w.r.t. h) as the sample-wise cost (w.r.t. 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 (1.7) and (1.8), play a crucial role in the complexity analysis of (multi-level) Monte Carlo methods for computing the expectation of (functionals of) the solution ϕ of (1.1). Suppose $Q(\phi)$ is such a functional (often called a *quantity of interest*) and to simplify notation we write this as Q (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 $\widehat{Q}_h$. 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).

To compare methods in UQ, the *computational ϵ-cost* $\mathcal{C}_\epsilon(\widehat{Q}_h)$ of an estimator $\widehat{Q}_h$ is often considered. If ϵ denotes a desired accuracy (in the sense of root mean-squared error), 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 a general theory in [13], the ϵ-cost of standard and multilevel Monte Carlo methods can be computed in terms of the parameter η in (1.7) (related to the regularity of the data), the parameter γ in (1.8) (related to the cost per sample), as well as another parameter β that quantifies the speed of variance reduction between levels of the multilevel scheme and can also be derived from (1.7). In the fourth main result of this paper in Theorem 5.14, we prove rigorously that

$$\beta \;\geq\; 2\eta \;. \tag{1.9}$$

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 in [26] are given in Example 5.10. In particular, for the asymptotically cheaper one of the two methods, which is an iterative procedure called source iteration, we have $\gamma \leq 1+\eta$. The general theory in [13] 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 $\widehat{Q}_h^{MLMC}$:

$$\mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\Big] \;=\; \mathcal{O}\big(\epsilon^{-\left(4+\frac{1-\eta}{\eta}\right)}\big) \quad \text{and} \quad \mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\Big] \;=\; \mathcal{O}\big(\epsilon^{-\left(2+\frac{1-\eta}{\eta}\right)}\big)\;,$$

i.e., a theoretical gain of up to two orders of magnitude in ϵ^{-1}. However, we will see in the numerical section that this estimate of improvement 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 in practice.

Related literature. 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 [32]. 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 have enjoyed a massive recent renaissance and the neutron transport problem was one of the motivations behind the original introduction of DG [38].

The fundamental paper on the analysis of the discrete ordinates method for the transport equation is [37], 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 (1.6)). Later research e.g. [4], [5], [6] 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. [33], [43], and [3]) allow for variation in cross-sections, but error estimates explicit in this data are not available there. A method for tackling directly the integral equation reformulation of the RTE is given in [21]. Again the analysis is not explicit in the heterogeneity.

While the coefficient-explicit analysis given here can in principle be extended to higher spatial and angular dimensions (a start is contained in [11]), it is clear that the details will be quite formidable, so for the full analysis we restrict here to the 1D case. However we note that there is substantial contemporary interest in practical 3D modelling in the heterogeneous case. For example, the recent thesis [24] solves the multigroup approximation of the P_N angular approximation of the transport equation using a non-conforming spatial discretization for a highly heterogeneous reactor, using a domain decomposition approach.

The field of UQ has grown very quickly in recent years and its application to radiative transport theory is currently of considerable interest. There are a number of groups that already work on UQ in radiative and neutron transport, e.g. [7, 20, 23] and references therein. The most recent research has focussed on using the polynomial chaos expansion (PCE), combined with a collocation method to estimate the coefficients in the expansion. The main disadvantage

of standard PCE is that the number of terms grow exponentially in the number of stochastic dimensions and in the order of the PCE, the so-called *curse of dimensionality*. A variety of techniques have been used to remedy this, including (adaptive) sparse grids [23], hybrid mixtures of polynomials [8] and expanding the quantity of interest in terms of low-dimensional subspaces of the stochastic variable [7]. We note however that none of these papers provide any rigorous error or cost analysis.

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. [29], which has led to recent work on UQ in this context (e.g. [44]). Recently modern operator compression and adaptive techniques have been applied to efficiently attack the high-dimensional aspects of the transport problem [14]. For further general discussion on the transport equation, see [15, 34].

By contrast our work focusses on multilevel Monte Carlo and sampling methods [26]. Monte Carlo is inherently dimension independent and Quasi-Monte Carlo can be proved to be so under certain conditions, e.g. [25]. As far as we know, these methods have not been applied to radiative transport until now. Our previous paper [26] gave an overview of this topic and also investigated the multilevel quasi-Monte Carlo method for the RTE. Further details, are in [36].

Structure of paper and notation. In Section 2, we introduce the model problem; the Radiative Transport equation in slab geometry with spatially heterogeneous cross-sections and its discretisation. To set up the error analysis, Section 3 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 4, the elements are brought together to prove (1.6). We introduce uncertainty into the input data in Section 5, and we extend the error estimate of Section 4 to the probabilistic error estimate (1.7) and subsequently prove (1.8). Numerical results are given, with the cross-sections assumed to be log-normal random fields equipped with the Matérn class of covariances, and represented by a Karhunen-Loève expansion. An overview of the results of this paper, without detailed analysis was previously presented in [26].

2 The Model Problem

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

$$\mu\frac{\partial\psi}{\partial x}(x,\mu) \;+\; \sigma(x)\psi(x,\mu) \;=\; \sigma_S(x)\phi(x) \;+\; f(x)\;, \quad x\in(0,1), \quad \mu\in[-1,1], \tag{2.1}$$

$$\text{where} \qquad \phi(x) \;=\; \frac{1}{2}\int_{-1}^{1}\psi(x,\mu')\;d\mu' \tag{2.2}$$

denotes the scalar flux, subject to zero incoming flux:

$$\psi(0,\mu) \;=\; 0, \quad \text{for} \quad \mu>0 \quad \text{and} \quad \psi(1,\mu)=0, \quad \text{for} \quad \mu<0\;. \tag{2.3}$$

The total cross-section $\sigma(x)$ is given by $\sigma=\sigma_S+\sigma_A$. The problem (2.1) – (2.3) is obtained from (1.1) – (1.3) when the input data is constant in two of the spatial dimensions (here assumed to be y and z). Note that (2.1) degenerates at $\mu=0$, which corresponds to particles moving perpendicular to the x-direction.

Notation 2.1. When working on the spatial domain $(0,1)$, for $1\le p\le\infty$, we will denote the standard Lebesgue spaces as L_p with norm $\|\cdot\|_p$. For any interval $I\subset[0,1]$, we denote by $\mathrm{C}(I)$ the space of uniformly continuous functions on I, equipped with norm $\|\cdot\|_\infty$. Any function

$g \in \mathrm{C}(I)$ has a unique continuous extension to $\overline{I}$. 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

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

When $I = [0,1]$, we write for short $\mathrm{C} = \mathrm{C}(I)$, $\mathrm{C}^\xi = \mathrm{C}^\xi(I)$ and $\|f\|_\xi = \|f\|_{\mathrm{C}^\xi(I)}$. Finally, for any normed spaces X and Y, we write $\|.\|_{X\mapsto Y}$ 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 a priori defined partition

$$0 = c_1 < ... < c_J = 1, \tag{2.4}$$

with $J \geq 2$. We denote the corresponding space of piecewise continuous functions by

$$\mathrm{C}_{pw} := \big\{ g \in L_\infty[0,1] : g|_{(c_j,c_{j+1})} \in C(c_j, c_{j+1}), \text{ for each } j = 1,\ldots,J-1 \big\}.$$

For definiteness we will assume that the value of $g(c_j)$ is taken to be the limit from the right for $j = 1,\ldots,J-1$ and the limit from the left for $j = J$. The space C_{pw} is equipped with the usual uniform norm $\|\cdot\|_\infty$. Similarly, for any $\xi \in (0,1]$, let

$$\mathrm{C}^\xi_{pw} := \big\{ g \in \mathrm{C}_{pw} : g|_{(c_j,c_{j+1})} \in \mathrm{C}^\xi(c_j, c_{j+1}), \text{ for each } j = 1,\ldots,J-1 \big\}$$

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

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

Assumption 2.2. (Input Data)

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

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

2. There exists a partition (2.4) and $\eta \in (0,1]$, such that $\sigma, \sigma_S, f \in \mathrm{C}^\eta_{pw}$.

2.1 Discretisation

To discretise (2.1) – (2.3) in angle, we use a $2N$-point quadrature rule

$$\int_{-1}^1 g(\mu)d\mu \approx \sum_{|k|=1}^N w_k g(\mu_k)\,, \tag{2.5}$$

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

$$0 = x_0 < x_1 < \ldots < x_M = 1, \tag{2.6}$$

which is assumed to resolve the break points $\{c_j\}$ introduced in (2.4). We set $h_j = x_j - x_{j-1}$. Further assumptions on the quadrature rule and mesh will be added in Section 4.

Our discrete scheme for (2.1) – (2.3) is then

$$\mu_k \frac{\psi^{h,N}_{k,j} - \psi^{h,N}_{k,j-1}}{h_j} \ + \ \sigma_{j-1/2} \frac{\psi^{h,N}_{k,j} + \psi^{h,N}_{k,j-1}}{2} \ = \ \sigma_{S,j-1/2}\phi^{h,N}_{j-1/2} \ + \ f_{j-1/2}\,, \tag{2.7}$$

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

$$\phi^{h,N}_{j-1/2} \ = \ \frac{1}{2} \sum_{|k|=1}^{N} w_k \frac{\psi^{h,N}_{k,j} \ + \ \psi^{h,N}_{k,j-1}}{2} \ , \quad j = 1, ..., M \ , \tag{2.8}$$

and with

$$\psi^{h,N}_{k,0} = 0, \quad \text{for} \quad k > 0 \quad \text{and} \quad \psi^{h,N}_{k,M} = 0, \quad \text{for} \quad k < 0 \ . \tag{2.9}$$

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 $f_{j-1/2}$.

2.2 Abstract form of the method

As preparation for analysing (2.1) – (2.3) and its discretisation, (2.7) – (2.9), consider first the *pure transport problem*: For fixed $\mu \in [-1, 1]$, find $u = u(x)$, $x \in (0, 1)$, such that

$$\mu \frac{du}{dx} + \sigma u = g, \quad \text{with } u(0) = 0, \text{ when } \mu > 0 \quad \text{and } u(1) = 0 \text{ when } \mu < 0, \tag{2.10}$$

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.) It is easy to show that the unique solution of this problem is $u := \mathcal{S}_\mu g$, where

$$\mathcal{S}_\mu g(x) \ = \ \begin{cases} \mu^{-1} \int_0^x \exp\left(\mu^{-1}\tau(x,y)\right) g(y) \ dy \ , & \mu > 0 \\ \sigma^{-1}(x) \ g(x) \ , & \mu = 0 \\ -\mu^{-1} \int_x^1 \exp\left(\mu^{-1}\tau(x,y)\right) g(y) \ dy \ , & \mu < 0 \end{cases} \ , \tag{2.11}$$

and

$$\tau(x, y) \ := \ \int_x^y \sigma(s) \ ds \ . \tag{2.12}$$

The quantity $|\tau(x, y)|$ is often called the 'optical length' or 'optical path' [9]. To mimic the averaging process in (2.2) it is natural to also consider the integral operator:

$$\mathcal{K}g(x) \ := \ \frac{1}{2} \int_{-1}^{1} \mathcal{S}_\mu g(x) \ d\mu \ = \ \frac{1}{2} \int_0^1 E_1(|\tau(x, y)|) g(y) \ dy \ , \tag{2.13}$$

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

$$E_1(z) \ := \ \int_1^\infty \exp(-tz) \frac{\mathrm{d}t}{t} \ = \ \int_0^1 \exp(-z/s) \frac{\mathrm{d}s}{s} \ . \tag{2.14}$$

The operators $\mathcal{S}_\mu$ and $\mathcal{K}$ relate to (2.1) – (2.3) by the following proposition.

Proposition 2.3. *Let ψ be a solution to* (2.1) – (2.3)*. Then,*

$$\psi(x, \mu) \ = \ \mathcal{S}_\mu \left(\sigma_S \phi \ + \ f\right)(x) \tag{2.15}$$

and hence, ϕ solves the integral equation

$$\phi \ = \ \mathcal{K} \left(\sigma_S \phi \ + \ f\right) \ . \tag{2.16}$$

We shall see later that (2.16) has a unique solution and this ensures that (2.1) – (2.3) has a unique solution. Analogously we can consider the discrete system (2.7) – (2.9). Let V^h denote the space of continuous piecewise-linear functions with respect to the mesh $\{x_j\}_{j=0}^M$, and for any

$v \in \mathrm{C}$, let $\mathcal{P}^h v$ denote the piecewise constant function which interpolates v at the mid-points of subintervals. Then consider the discretisation of (2.10) defined by seeking $u^h \in V^h$ to satisfy

$$\int_{I_j} \left(\mu \frac{\mathrm{d}u^h}{\mathrm{d}x} + \mathcal{P}^h \sigma \; u^h \right) \; = \; \int_{I_j} g \; , \quad \text{with} \quad I_j = (x_{j-1}, x_j) \; , \quad j = 1, \ldots, M \; , \tag{2.17}$$

with $u^h(0) = 0$ when $\mu > 0$ and $u^h(1) = 0$ when $\mu < 0$. This has a unique solution, which we write as $u^h = \mathcal{S}^h_\mu g$. Analogously to (2.13) we also define

$$\mathcal{K}^{h,N} g \; = \; \frac{1}{2} \sum_{|k|=1}^{N} w_k \; \mathcal{S}^h_{\mu_k} g \; . \tag{2.18}$$

Identifying any fully discrete solution $\psi^{h,N}_{k,j}$ of (2.7) – (2.9) with the function $\psi^{h,N}_k \in V^h$ by interpolation at the nodes $\{x_j\}$, we can see that (2.7) – (2.9) is equivalent to seeking $\psi^{h,N}_k \in V^h$, $|k| = 1, ..., N$, that satisfy

$$\int_{I_j} \left(\mu_k \frac{\mathrm{d}\psi^{h,N}_k}{\mathrm{d}x} + \mathcal{P}^h \sigma \; \psi^{h,N}_k \right) \; = \; \int_{I_j} \mathcal{P}^h \left(\sigma_S \phi^{h,N} + f \right) \; , \quad j = 1, \ldots, M \; , \tag{2.19}$$

where

$$\phi^{h,N} \; = \; \frac{1}{2} \sum_{|k|=1}^{N} w_k \psi^{h,N}_k \tag{2.20}$$

and

$$\psi^{h,N}_k(0) = 0 \quad \text{when} \quad k > 0 \quad \text{and} \quad \psi^{h,N}_k(1) = 0 \quad \text{when} \quad k < 0. \tag{2.21}$$

We then have the discrete analogue of Proposition 2.3:

Proposition 2.4. *The system* (2.19) – (2.21) *is equivalent to* (2.7) – (2.9)*, and its solution can be written:*

$$\psi^{h,N}_k \; = \; \mathcal{S}^h_{\mu_k} \mathcal{P}^h (\sigma_S \phi^{h,N} + f) \; , \qquad |k| = 1, ..., N \; . \tag{2.22}$$

Moreover,

$$\phi^{h,N} \; = \; \mathcal{K}^{h,N} \mathcal{P}^h (\sigma_S \phi^{h,N} \; + \; f) \; . \tag{2.23}$$

Now, to estimate the error in our approximation to ϕ, we use (2.16) and (2.23) to obtain

$$\phi - \phi^{h,N} \; = \; (I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S)^{-1} (\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h)(\sigma_S \phi + f) \; . \tag{2.24}$$

We prove later in (4.15) that $(I - \mathcal{K}^{h,N} \mathbb{P}^h \sigma_S)^{-1}$ is bounded on C. Hence,

$$\| \phi - \phi^{h,N} \|_\infty \; \leq \; \| (I - \mathcal{K}^{h,N} \mathcal{P}^h \sigma_S)^{-1} \|_{\mathrm{C} \mapsto \mathrm{C}} \| (\mathcal{K} - \mathcal{K}^{h,N} \mathcal{P}^h)(\sigma_S \phi + f) \|_\infty \; . \tag{2.25}$$

In §4, we use (2.25) to obtain a data-explicit error estimate for $\phi - \phi^{h,N}$. First, we prove a number of data-explicit properties of the operator $\mathcal{K}$ which will be needed later.

3 Properties of the Operators

In this section we briefly present some properties of the operators $\mathcal{S}_\mu$ and $\mathcal{K}$. These could be viewed as technical extensions of the classical results in [37], generalised to the heterogeneous case. Proofs of Lemma 3.2 and Theorem 3.3 are available in [27] and [36].

Notation 3.1. To simplify presentation, for any $a \in \mathbb{R}$, we will use the notation $\overline{a} := \max\{1, a\}$ and $\underline{a} := \min\{a, 1\}$. 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.

We will make use of the following bounds, a consequence of Assumptions 2.2 and (2.12),

$$\sigma_{\min}|y-x| \le \mathrm{sgn}(y-x)\tau(x,y) \;=\; |\tau(x,y)| \;\le \sigma_{\max}|y-x| \;, \tag{3.1}$$

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

Lemma 3.2.

(i) *For all* $\mu \in [-1,1]$, $\|\mathcal{S}_\mu\|_{L_\infty \mapsto L_\infty} \le \sigma_{\min}^{-1}$, *and* $\|\mathcal{S}_\mu\|_{L_\infty \mapsto \mathrm{C}} \le \sigma_{\min}^{-1}$, *when* $\mu \neq 0$.

(ii) $\left\| \dfrac{\partial}{\partial x}\mathcal{S}_\mu \right\|_{L_\infty \mapsto L_\infty} \;\le\; 2\left(\dfrac{\sigma_{\max}}{\sigma_{\min}}\right)|\mu|^{-1}$, *for all* $\mu \in [-1,1]\backslash\{0\}$.

(iii) $\sup\limits_{x\in[0,1]} \displaystyle\int_{-1}^{1} |\mu|^\beta \left| \frac{\partial}{\partial \mu}(\mathcal{S}_\mu g)(x) \right| \, d\mu \;\le\; 2\beta^{-1}\sigma_{\min}^{-1}\|g\|_\infty$, *for all* $g \in L_\infty$ *and* $\beta > 0$.

Theorem 3.3. *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) $\|\mathcal{K}\|_{L_2 \mapsto L_\infty} \;\le\; \sqrt{\log(2)}\,\sigma_{\min}^{-1/2}$;

(ii) $\|\mathcal{K}\|_{L_\infty \mapsto \mathrm{C}^\xi} \;\le\; c_\xi\, \overline{\sigma_{\max}}/\underline{\sigma_{\min}}$,

where c_ξ *may depend on* ξ *and where* $\overline{a}$ *and* $\underline{a}$ *are defined in Notation 3.1.*

Lemma 3.4. *The operator* $(I - \mathcal{K}\sigma_S)$ *is invertible on* C *with the bound*

$$\|\,(I-\mathcal{K}\sigma_S)^{-1}\,\|_{\mathrm{C}\mapsto\mathrm{C}} \;\le\; 2\overline{\sigma_{\max}}^{1/2}\,\frac{\sigma_{\max}}{\sigma_{\min}}\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1} \;=:\; \mathcal{R}_1(\sigma,\sigma_S)\;. \tag{3.2}$$

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

Proof. Let $g \in \mathrm{C}$ and suppose that

$$(I - \mathcal{K}\sigma_S)\,v = g, \quad \text{or equivalently that} \;\; v = \mathcal{K}\sigma_S v + g. \tag{3.3}$$

This allows us to apply a bootstrapping argument. From [11, Theorem 1], it follows that $v \in L_2$ and

$$\|v\|_2 \le \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{1/2}\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1}\|g\|_2 \le \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{1/2}\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1}\|g\|_\infty \;. \tag{3.4}$$

Using (3.3) again, this time together with Theorem 3.3(i) we get $v \in L_\infty$. Finally, using (3.3) with Theorem 3.3(ii) we conclude that $v \in C$, and (using Theorem 3.3 (i)), that

$$\|v\|_\infty \le \|\mathcal{K}\sigma_S v\|_\infty + \|g\|_\infty \le \sigma_{\min}^{-1/2}(\sigma_S)_{\max}\|v\|_2 + \|g\|_\infty \le \sigma_{\max}^{1/2}\Big(\frac{\sigma_{\max}}{\sigma_{\min}}\Big)^{1/2}\|v\|_2 + \|g\|_\infty \;. \tag{3.5}$$

The bound in (3.2) follows on combining (3.4) and (3.5).

Now suppose $g \in \mathrm{C}_{pw}$. Then $g \in L_2$ and the argument above holds verbatim to show that then $v \in \mathrm{C}_{pw}$ and that the bounds in (3.4) and (3.5) hold again. □

The main result of this section now follows from Theorem 3.3 and Lemma 3.4.

Corollary 3.5. *Let $f \in \mathrm{C}_{pw}$, and let ϕ be the solution to (2.16). Then $\phi \in \mathrm{C}^{\xi}$, for any $\xi \in (0,1)$, and the following two bounds hold:*

$$\|\phi\|_\infty \;\le\; c\sigma_{\min}^{-1/2}\mathcal{R}_1(\sigma,\sigma_S)\|f\|_\infty \qquad \textit{and} \qquad \|\phi\|_\xi \;\le\; c_\xi \mathcal{R}_2(\sigma,\sigma_S)\|f\|_\infty \;, \tag{3.6}$$

where c_ξ may depend on ξ, $\mathcal{R}_1(\sigma,\sigma_S)$ is defined in (3.2) and

$$\mathcal{R}_2(\sigma,\sigma_S) \;:=\; \overline{\sigma_{\max}}^{1/2}\,(\overline{\sigma_{\max}}/\underline{\sigma_{\min}})^{3/2}\,\mathcal{R}_1(\sigma,\sigma_S)\;. \tag{3.7}$$

Proof. Recall (2.16), so we have $(I-\mathcal{K}\sigma_S)\phi = \mathcal{K}f$. From the proof of Theorem 3.3(ii), $\mathcal{K}f \in \mathrm{C}^{\xi}$ for any $\xi \in (0,1)$ and $\|\mathcal{K}f\|_\infty \le c\sigma_{\min}^{-1/2}\|f\|_\infty$. Then Lemma 3.4 implies that

$$\|\phi\|_\infty \;\le\; c\,\sigma_{\min}^{-1/2}\mathcal{R}_1(\sigma,\sigma_S)\|f\|_\infty.$$

To obtain the second bound, we use Theorem 3.3(ii) again to obtain

$$\|\phi\|_\xi \le \|\mathcal{K}\left(\sigma_S\phi + f\right)\|_\xi \;\le\; c\,(\overline{\sigma_{\max}}/\underline{\sigma_{\min}})\left((\sigma_S)_{\max}\|\phi\|_\infty + \|f\|_\infty\right)$$

and then combine this with the first bound in (3.6). (Again c may depend on ξ.) □

4 Deterministic Error Estimate

We now return to estimating the error $\phi - \phi^{h,N}$ using (2.24). Introducing the operator

$$\mathcal{K}^N g(x) \;:=\; \frac{1}{2}\sum_{|k|=1}^{N} w_k\left(\mathcal{S}_{\mu_k} g\right)(x) \;=\; \frac{1}{2}\int_0^1 E_1^N(|\tau(x,y)|)g(y)\,dy\;, \tag{4.1}$$

with $E_1^N(z) := \sum_{k=1}^N w_k\mu_k^{-1}\exp(-\mu_k^{-1}z)$ denoting the N-point quadrature approximation of the exponential integral (2.14), we can write (2.24) as

$$\phi - \phi^{h,N} \;=\; \left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}\left(e^N + e^{h,N}\right)\;, \tag{4.2}$$

where

$$e^N := \left(\mathcal{K}-\mathcal{K}^N\right)\left(\sigma_S\phi+f\right) \quad \text{and} \quad e^{h,N} := \left[\left(\mathcal{K}^N - \mathcal{K}^{h,N}\right) + \mathcal{K}^{h,N}\left(I-\mathcal{P}^h\right)\right]\left(\sigma_S\phi+f\right)\;. \tag{4.3}$$

Finally, to obtain an error estimate we apply the supremum norm to (4.2), and by trivial manipulation write

$$\|\phi-\phi^{h,N}\|_\infty \;\le\; \|\left(I-\mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}\|_{\mathrm{C}\mapsto\mathrm{C}}\left(\|e^N\|_\infty + \|e^{h,N}\|_\infty\right)\;. \tag{4.4}$$

The error analysis proceeds 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 is done in Sections 4.1 and 4.2, while the second is done in Section 4.3.

4.1 Consistency under Angular Discretisation

Here we estimate e^N using the angular regularity (Lemma 3.2 (iii)) and the following result from De Vore and Scott [16] (see also [37, Prop. 3.2]):

Proposition 4.1. *Consider the N-point Gauss-Legendre rule on $[0,1]$ and let m be a positive integer with $m \le 2N-1$. Then we have*

$$\left| \int_0^1 g(\mu)\, d\mu - \sum_{k=1}^{N} w_k g(\mu_k) \right| \;\le\; cN^{-m} \int_0^1 \left[\mu\,(1-\mu)\right]^{m/2} |g^{(m)}(\mu)|\, d\mu \;,$$

whenever the integral on the right hand side exists.

Notation 4.2. In this paper the particular case of (2.5) where the N-point Gauss-Legendre rule is used on both $[-1,0]$ and on $[0,1]$, is called the *double Gauss rule.*

Theorem 4.3. *Let $\mathcal{K}^N$ be defined by (4.1) using the double Gauss rule. Then,*

$$\|\mathcal{K} - \mathcal{K}^N\|_{L_\infty \mapsto \mathrm{C}} \;\le\; c\,\sigma_{\min}^{-1} N^{-1} \;. \tag{4.5}$$

Proof. Using (2.13), (4.1), the (anti-)symmetry properties of the double Gauss rule, then Proposition 4.1 (with $m=1$) and finally Lemma 3.2 (iii) (with $\beta = 1/2$), we obtain for any $g \in L_\infty$,

$$\begin{aligned}
|\left(\mathcal{K} - \mathcal{K}^N\right) g(x)| \;&=\; \frac{1}{2}\left| \int_0^1 \left(\mathcal{S}_\mu + \mathcal{S}_{-\mu}\right) g(x)\, d\mu - \sum_{k=1}^{N} \left(w_k \mathcal{S}_{\mu_k} + w_{-k}\mathcal{S}_{\mu_{-k}}\right) g(x) \right| \\
&\le\; \frac{1}{2}\left| \int_0^1 \mathcal{S}_\mu g(x)\, d\mu - \sum_{k=1}^{N} w_k \mathcal{S}_{\mu_k} g(x) \right| + \frac{1}{2}\left| \int_0^1 \mathcal{S}_{-\mu} g(x)\, d\mu - \sum_{k=1}^{N} w_k \mathcal{S}_{-\mu_k} g(x) \right| \\
&\le\; cN^{-1} \left[\int_0^1 \mu^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x)\right) \right| d\mu + \int_{-1}^0 (-\mu)^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x)\right) \right| d\mu \right] \\
&\le\; cN^{-1} \int_{-1}^1 |\mu|^{1/2} \left| \frac{\partial}{\partial \mu} \left(\mathcal{S}_\mu g(x)\right) \right| d\mu \;\le\; cN^{-1}\sigma_{\min}^{-1} \|g\|_\infty \;.
\end{aligned}$$

Hence, $\left(\mathcal{K} - \mathcal{K}^N\right) : L_\infty \mapsto L_\infty$ satisfies the bound in (4.5). The extension to C holds because, by Lemma 3.2 (i), $\mathcal{S}_\mu$ maps from L_∞ to C. □

Corollary 4.4. *Under the conditions of Theorem 4.3, $e^N \in \mathrm{C}$, with the bound*

$$\|e^N\|_\infty \;\le\; c \frac{\overline{\sigma_{\max}}}{\sigma_{\min}} \sigma_{\min}^{-1/2} \mathcal{R}_1(\sigma, \sigma_S) N^{-1} \|f\|_\infty \;.$$

Proof. By (4.3), Theorem 4.3 and Corollary 3.5, we obtain

$$\|e^N\|_\infty \;\le\; cN^{-1}\sigma_{\min}^{-1}\left(\sigma_{\max}\|\phi\|_\infty + \|f\|_\infty\right) \le\; cN^{-1}\sigma_{\min}^{-1}\left(\sigma_{\max}\sigma_{\min}^{-1/2}\mathcal{R}_1(\sigma,\sigma_S) + 1\right)\|f\|_\infty,$$

from which the estimate follows. □

4.2 Consistency under Spatial Discretisation

From now on, for convenience we make the following quasi-uniformity assumption:

Assumption 4.5. For some constant $\rho \ge 1$, the local mesh diameters $h_j := x_j - x_{j-1}$ satisfy

$$\max_{j=1,\ldots,M} h_j \;=:\; h \;\le\; \rho \min_{j=1,\ldots,M} h_j \;. \tag{4.6}$$

Our main deterministic error estimate (Theorem 4.10) will contain an $h \log N$ term. The next result is the first indication of how this arises: the estimates below blow up as $|\mu| \to 0$ for fixed h. The following lemma is proved in [27] (and also in [36]) and can again be seen as a generalisation to the heterogeneous case of classical results in [37].

Lemma 4.6. *Let $\mu \in [-1,1]\backslash\{0\}$. For $\mathcal{S}_\mu^h$ defined by (2.17), $\mathcal{S}_\mu^h : L_\infty \mapsto V^h \subset \mathrm{C}$ and there is a constant $c > 0$, independent of all parameters, such that*

$$\|\mathcal{S}_\mu^h\|_{L_\infty \mapsto V^h} \;\le\; 2\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}\frac{h}{|\mu|}\right) \;, \tag{4.7}$$

$$\|\mathcal{S}_\mu - \mathcal{S}_\mu^h\|_{L_\infty \mapsto \mathrm{C}} \;\le\; c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2\frac{h}{|\mu|} + \|\sigma\|_{\eta,pw}h^\eta\right) \;. \tag{4.8}$$

The next lemma obtains some estimates needed to bound $e^{h,N}$.

Lemma 4.7. *Let $\mathcal{K}^N$ and $\mathcal{K}^{h,N}$ be defined by (4.1) and (2.18) respectively, with μ_k and w_k given by the double Gauss rule (Notation 4.2). Under Assumption (4.5), if $N \ge 2$, then*

$$\begin{aligned}
&(i) \qquad \|\mathcal{K}^N - \mathcal{K}^{h,N}\|_{L_\infty \mapsto \mathrm{C}} \;&&\le\; c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2 h\log N + \|\sigma\|_{\eta,pw}h^\eta\right) \;,\\
&(ii) \qquad \|\mathcal{K}^{h,N}\|_{L_\infty \mapsto \mathrm{C}} \;&&\le\; c\rho\sigma_{\min}^{-1}\left(1 + \sigma_{\max}h\log N\right) \;.
\end{aligned}$$

Proof. To prove part (i), let $g \in L_\infty$. Then, we have $(\mathcal{K}^N - \mathcal{K}^{h,N})g \in \mathrm{C}$ and

$$\begin{aligned}
\|\left(\mathcal{K}^N - \mathcal{K}^{h,N}\right)g\|_\infty &\le c\sum_{|k|=1}^N w_k\|\mathcal{S}_{\mu_k} - \mathcal{S}_{\mu_k}^h\|_{L_\infty\to\mathrm{C}}\|g\|_\infty\\
&\le c\rho\sigma_{\min}^{-2}\sum_{|k|=1}^N w_k\left(\sigma_{\max}^2\frac{h}{|\mu_k|} + \|\sigma\|_{\eta,pw}h^\eta\right)\|g\|_\infty \;.
\end{aligned} \tag{4.9}$$

Since the double Gauss rule integrates constants exactly, we have $\sum_{|k|=1}^N w_k = 2$. Also [37, Lemma 3.1] gives the estimate

$$\sum_{|k|=1}^N w_k|\mu_k|^{-1} \;\le\; c\,(1 + |\log\mu_1|) \;\le c\,(1+\log N)\,, \tag{4.10}$$

where the last inequality follows because $\mu_1 \sim N^{-2}$ for the Gauss rule on $[0,1]$. Substituting these estimates into (4.9) yields the result (i). To obtain (ii), we proceed similarly:

$$\|\mathcal{K}^{h,N}g\|_\infty \;\le\; c\sum_{|k|=1}^N w_k\|\mathcal{S}_{\mu_k}^h\|_{L_\infty\to\mathrm{C}}\,\|g\|_\infty \;\le\; c\rho\sigma_{\min}^{-1}\sum_{|k|=1}^N w_k\left(1 + \sigma_{\max}\frac{h}{|\mu_k|}\right)\|g\|_\infty \;,$$

from which the result follows analogously to (i). □

Theorem 4.8. *Suppose the assumptions of Lemma 4.7 hold. Then, for $e^{h,N}$ defined in (4.3),*

$$\|e^{h,N}\|_\infty \;\le\; c\rho\sigma_{\min}^{-2}\overline{\|\sigma_S\|_{\eta,pw}}\mathcal{R}_2(\sigma,\sigma_S)\left(\sigma_{\max}^2 h\log N + h^\eta\|\sigma\|_{\eta,pw}\right)\|f\|_{\eta,pw} \;.$$

Proof. It is easy to check that $\|\sigma_S\phi\|_{\eta,pw} \le 3\|\sigma_S\|_{\eta,pw}\|\phi\|_\eta$. Using this, Corollary 3.5, and recalling that $\mathcal{R}_2(\sigma,\sigma_S) \ge 1$, we obtain

$$\|\sigma_S\phi + f\|_{\eta,pw} \;\le\; (1 + 3\|\sigma_S\|_{\eta,pw}\mathcal{R}_2(\sigma,\sigma_S))\,\|f\|_{\eta,pw} \;\le 3\overline{\|\sigma_S\|_{\eta,pw}}\mathcal{R}_2(\sigma,\sigma_S)\|f\|_{\eta,pw} \;. \tag{4.11}$$

Hence, using (4.3) and the results of Lemma 4.7,

$$\begin{aligned}
\|e^{h,N}\|_\infty \;&\le\; \|\mathcal{K}^N - \mathcal{K}^{h,N}\|_{L_\infty\mapsto\mathrm{C}}\|\sigma_S\phi + f\|_\infty + \|\mathcal{K}^{h,N}\|_{L_\infty\mapsto\mathrm{C}}\,h^\eta\|\sigma_S\phi + f\|_{\eta,pw}\\
&\le\; \left(\|\mathcal{K}^N - \mathcal{K}^{h,N}\|_{L_\infty\mapsto\mathrm{C}} + h^\eta\|\mathcal{K}^{h,N}\|_{L_\infty\mapsto\mathrm{C}}\right)\|\sigma_S\phi + f\|_{\eta,pw}\\
&\le c\rho\sigma_{\min}^{-2}\left(\sigma_{\max}^2 h\log N + h^\eta\|\sigma\|_{\eta,pw} + h^\eta\sigma_{\min}(1 + \sigma_{\max}h\log N)\right)\|\sigma_S\phi + f\|_{\eta,pw} \;.
\end{aligned} \tag{4.12}$$

The result is obtained by combining (4.11) and (4.12) and simplification. □

4.3 Stability and Convergence

So far we have shown that $\|e^N\|_\infty$ and $\|e^{h,N}\|_\infty$ approach zero as $h \log N \to 0$ and $N \to \infty$. See Corollary 4.4 and Theorem 4.8, respectively. To prove a final bound on (4.4) we need to show "stability", i.e. that $\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}$ exists and is bounded in the $\|\cdot\|_{\mathrm{C}\mapsto\mathrm{C}}$ norm, independently of h and N. To do this, a useful trick is to write

$$\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1} \;=\; I \;+\; \mathcal{K}^{h,N}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\mathcal{P}^h\sigma_S \;, \tag{4.13}$$

which holds when the inverses on each side exist. We obtain stability by proving that the inverse exists on the right-hand side and then estimating all the terms on the right-hand side.

Theorem 4.9. *Under the assumptions of Lemma 4.7, there is a constant $K > 0$ such that, if h and N^{-1} are sufficiently small so that $h \log N \le 1$ and*

$$(h^\eta + h\log N + N^{-1})^{-1} \;\ge\; K\left(\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\right)\left(\frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}}\right)^3 \overline{\|\sigma\|_{\eta,pw}}\mathcal{R}_1(\sigma,\sigma_S) =: \mathcal{R}_3(\sigma,\sigma_S) \;, \tag{4.14}$$

then $\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}$ is bounded on C*, with the bound*

$$\|\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}\|_{\mathrm{C}\mapsto\mathrm{C}} \;\le\; c\rho\left(\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\right)\left(\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\right)\overline{(\sigma_S)_{\max}}\mathcal{R}_1(\sigma,\sigma_S) \;=:\; \mathcal{R}_4(\sigma,\sigma_S) \;, \tag{4.15}$$

where $\mathcal{R}_1(\sigma,\sigma_S)$ is defined in Lemma 3.4.

Proof. Introduce the family of operators $\mathcal{A}^{h,N} := I - (I - \sigma_S\mathcal{K})^{-1}\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)$. Suppose that for some h and N, we can ensure

$$\|\mathcal{A}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;\le\; 1/2 \;. \tag{4.16}$$

Then it follows from the Banach Lemma that

$$\|\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}(I - \sigma_S\mathcal{K})\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;=\; \|(I - \mathcal{A}^{h,N})^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;\le\; 2 \;. \tag{4.17}$$

Therefore,

$$\begin{aligned}\|\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;&=\; \|\left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)^{-1}(I - \sigma_S\mathcal{K})(I - \sigma_S\mathcal{K})^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \\ &\le\; 2\,\|(I - \sigma_S\mathcal{K})^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;\le\; 2\,\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\mathcal{R}_1(\sigma,\sigma_S) \;,\end{aligned} \tag{4.18}$$

where we used the identity $(I - \sigma_S\mathcal{K})^{-1} = \sigma_S\,(I - \mathcal{K}\sigma_S)^{-1}\,\sigma_S^{-1}$ and Lemma 3.4. Thus, on the assumption that (4.16) holds, we have (on combining (4.13) with Lemma 4.7 (ii), and (4.18), and recalling $h\log N \le 1$),

$$\|\left(I - \mathcal{K}^{h,N}\mathcal{P}^h\sigma_S\right)^{-1}\|_{\mathrm{C}\mapsto\mathrm{C}} \;\le\; 1 \;+\; \left[c\rho\sigma_{\min}^{-1}(1 + \sigma_{\max})\right]\left[2\,\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\mathcal{R}_1(\sigma,\sigma_S)\right]\left[(\sigma_S)_{\max}\right] \;, \tag{4.19}$$

which yields (4.15).

It remains to find conditions which ensure (4.16). To do this, we write

$$\begin{aligned}\|\mathcal{A}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;&=\; \|(I - \sigma_S\mathcal{K})^{-1}\left[(I - \sigma_S\mathcal{K}) - \left(I - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\right)\right]\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \\ &\le\; \|(I - \sigma_S\mathcal{K})^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}}\|\mathcal{P}^h\sigma_S\mathcal{K}^{h,N} - \sigma_S\mathcal{K}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \\ &\le\; \frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\|(I - \mathcal{K}\sigma_S)^{-1}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}}\|\sigma_S\mathcal{K} - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \\ &\le\; \frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\mathcal{R}_1(\sigma,\sigma_S)\;\|\sigma_S\mathcal{K} - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;,\end{aligned} \tag{4.20}$$

where we again used Lemma 3.4. To estimate the right hand side of (4.20) we write

$$\begin{aligned}\|\sigma_S\mathcal{K} - \mathcal{P}^h\sigma_S\mathcal{K}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;&\leq\; \|\left(I-\mathcal{P}^h\right)\sigma_S\mathcal{K}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \\ &\quad+\; \|\mathcal{P}^h\sigma_S\|_{\mathrm{C}\mapsto\mathrm{C}_{pw}}\left(\|\mathcal{K}-\mathcal{K}^N\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}}+\|\mathcal{K}^N-\mathcal{K}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}}\right)\\ &=: T_1+T_2\;.\end{aligned}$$

We can bound T_2 using Theorem 4.3 and Lemma 4.7 to obtain

$$\begin{aligned}T_2 \;&\leq\; c\rho\,(\sigma_S)_{\max}\left(\sigma_{\min}^{-1}N^{-1}+\sigma_{\min}^{-2}\sigma_{\max}^2 h\log N+\sigma_{\min}^{-2}\|\sigma\|_{\eta,pw}h^\eta\right)\\ &\leq\; c\rho\left(\frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}}\right)^3\overline{\|\sigma\|_{\eta,pw}}(h^\eta+h\log N+N^{-1}).\end{aligned}$$

(The last inequality is an over-estimate, but we do this to reduce technicalities.) On the other hand, using Theorem 3.3, we have

$$T_1 \;\leq\; ch^\eta\|\sigma_S\|_{\eta,pw}\frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}}\;.$$

Combining the estimates for T_1 and T_2 we obtain

$$\|\mathcal{A}^{h,N}\|_{\mathrm{C}_{pw}\mapsto\mathrm{C}_{pw}} \;\leq\; c\rho\left(\frac{(\sigma_S)_{\max}}{(\sigma_S)_{\min}}\right)\left(\frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}}\right)^3\overline{\|\sigma\|_{\eta,pw}}\mathcal{R}_1(\sigma,\sigma_S)(h^\eta+h\log N+N^{-1})$$

and the result follows on recalling (4.16). □

We now have all the ingredients to prove the main result of this section.

Theorem 4.10. *Let $\phi^{h,N}$ be as defined in §2.1. Under the assumptions of Lemma 4.7, provided that $h\log N\leq 1$ and that (4.14) holds, we have*

$$\|\phi-\phi^{h,N}\|_\infty \;\leq\; c\mathcal{R}(\sigma,\sigma_S)\left(N^{-1}\;+\;h\log N\;+\;h^\eta\right)\|f\|_{\eta,pw}\;, \tag{4.21}$$

where

$$\mathcal{R}(\sigma,\sigma_S)=\rho\mathcal{R}_4(\sigma,\sigma_S)\mathcal{R}_2(\sigma,\sigma_S)\left(\frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}}\right)^2\overline{\|\sigma\|_{\eta,pw}}\;\overline{\|\sigma_S\|_{\eta,pw}}\;. \tag{4.22}$$

Proof. We employ (4.4), combined with Theorem 4.9, Corollary 4.4 and Theorem 4.8 to obtain:

$$\begin{aligned}\|\phi-\phi^{h,N}\|_\infty \;\leq\; c\rho\mathcal{R}_4(\sigma,\sigma_S)\Bigg[&\frac{\overline{\sigma_{\max}}}{\sigma_{\min}}\sigma_{\min}^{-1/2}\mathcal{R}_1(\sigma,\sigma_S)N^{-1}\\ &+\;\frac{\overline{\|\sigma_S\|_{\eta,pw}}}{\sigma_{\min}^2}\mathcal{R}_2(\sigma,\sigma_S)\left(\sigma_{\max}^2 h\log N+\|\sigma\|_{\eta,pw}h^\eta\right)\Bigg]\|f\|_{\eta,pw}\end{aligned}$$

and the result follows after some algebra and recalling $\mathcal{R}_1(\sigma,\sigma_S)\leq\mathcal{R}_2(\sigma,\sigma_S)$. □

Remark 4.11. In [36] a numerical example is given, providing evidence that the estimate in Theorem 4.10 is sharp in terms of its dependence on the spatial smoothness parameter η.

5 Application in Uncertainty Quantification

In this section we allow the cross-sections to be random fields. Our main result is Theorem 5.6, which is a probabilistic counterpart of Theorem 4.10. The chief technical difficulty in obtaining this is the coefficient-dependent stability condition (4.14), which in the random case becomes a path-dependent condition, and so simply integrating (4.21) in probability space is not possible. Instead we prove Theorem 5.6 for an *a priori* chosen deterministic stepsize h. This means that for some realisations the mesh might need to be further refined in order to obtain stability. However in our cost estimate (Lemma 5.8) we show that the expected value of the cost is unaffected by these (relatively rare) events.

5.1 Random Input Data and Probabilistic Error Estimates

To describe the random case, we let $\omega \in \Omega$ denote a random event from a sample space Ω, and let $\mathbb{P} : \Omega \mapsto [0,1]$ denote the associated probability measure. For any normed space $(X, \|\cdot\|_X)$, we define the Bochner space $L_p(\Omega; X) := \{g : \Omega \mapsto X \ : \ \|g\|^p_{L_p(\Omega;X)} := \int_\Omega \|g\|^p_X \, \mathrm{d}\mathbb{P}(\omega) < \infty\}$.

Assumption 5.1. (Random Input Data) We assume $\sigma_S = \sigma_S(\omega, \cdot)$, $\sigma = \sigma(\omega, \cdot)$ and $f = f(\omega, \cdot)$ are now random fields. We set $\sigma_A(\omega, \cdot) = \sigma(\omega, \cdot) - \sigma_S(\omega, \cdot)$ and assume that σ_S, σ_A and hence σ are all positive-valued. Also, for an *a priori* specified partition (2.4), and for some $\eta \in (0,1)$ we assume:

(a) $\sigma,\ \sigma_S \in L_p(\Omega; \mathrm{C}^\eta_{pw})$, for all $p \in [1, \infty)$;

(b) $((\sigma_S)_{\min})^{-1}$, $((\sigma_A)_{\min})^{-1} \in L_p(\Omega)$, for all $p \in [1, \infty)$;

(c) $f \in L_{p_*}(\Omega; \mathrm{C}^\eta_{pw})$, for some $p_* \in (1, \infty]$.

We note that (a) (b), combined with the positivitiy of the cross-sections imply that, for all $p \in [1, \infty)$,

$$(\sigma_S)_{\max}(\omega) \le \sigma_{\max}(\omega) \ = \ \|\sigma(\omega, \cdot)\|_\infty \ \le \ \|\sigma(\omega, \cdot)\|_{\eta,pw} \in L_p(\Omega)\,, \tag{5.1}$$

$$\text{and} \qquad \sigma^{-1}_{\min} \le ((\sigma_S)_{\min})^{-1} \in L^p(\Omega)\,. \tag{5.2}$$

Example 5.2. A class of suitable random fields that can be shown to satisfy Assumption 5.1(a), (b) is got by choosing $\sigma_S(\omega, \cdot) = \exp(Z(\omega, \cdot))$ where Z is a centered Gaussian random field with Matérn covariance function, and choosing $\sigma_A(\omega, \cdot) = \sigma_A(\cdot)$ (a deterministic spatial function). Then, Assumption 5.1 (a), (b) hold true with $\eta < \nu$ and ν is the Matérn smoothness parameter [12]. This example will be used in the numerical experiments in Section 5.3.

Note that with these assumptions, the quantity $\mathcal{R}(\sigma, \sigma_S)$ appearing on the right-hand side of Theorem 4.10 is now a scalar-valued random variable. Our next result, Lemma 5.3, estimates the norm of this quantity in probability space.

Lemma 5.3. $\mathcal{R}(\sigma, \sigma_S) \in \ L_p(\Omega)\,, \quad$ *for all* $\quad p \in \ [1, \infty)$.

Proof. Using equations (4.22), (4.15) and (3.7), we see that there exists an integer $r > 0$ such that the random variable $\mathcal{R}(\sigma, \sigma_S)$ has a pathwise bound of the form

$$\mathcal{R}(\sigma, \sigma_S) \ \le \ c \ \left[\mathcal{R}'(\sigma, \sigma_S)\right]^r \ \left[\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1}\right]^2 \quad \text{with} \ \ \mathcal{R}'(\sigma, \sigma_S) := \frac{\overline{\sigma_{\max}}}{\underline{\sigma_{\min}}} \ \overline{\|\sigma\|_{\eta,pw}} \ \overline{\|\sigma_S\|_{\eta,pw}}\,. \tag{5.3}$$

Each of the terms in $\mathcal{R}'(\sigma, \sigma_S)$ can be shown to be in $L_p(\Omega)$ for all $p \in [1, \infty)$. We justify this only for $\overline{\sigma_{\max}}/\underline{\sigma_{\min}}$. The other terms are similar. Recall from Notation 3.1: If $a \in L_p(\Omega)$ is any scalar random variable, then $\overline{a} \in L_p(\Omega)$. Also if $a^{-1} \in L_p(\Omega)$, then also $(\underline{a})^{-1} = \overline{(a^{-1})} \in L_p(\Omega)$.

Assumption 5.1 ensures that $\sigma_{\max} \le \|\sigma\|_{\eta,pw} \in L_p(\Omega)$, and thus $\overline{\sigma_{\max}} \in L_p(\Omega)$, for all $p \in [1, \infty)$. Similarly, Assumption 5.1(b) ensures $\underline{\sigma_{\min}}^{-1} \in L_p(\Omega)$, for all $p \in [1, \infty)$. Using the generalised Hölder inequality, it follows that $\overline{\sigma_{\max}}/\underline{\sigma_{\min}} \in L_p(\Omega)$, for all $p \in [1, \infty)$. Proceeding similarly for the other terms, it follows that $\mathcal{R}'(\sigma, \sigma_S) \in L_p(\Omega)$ for all $p \in [1, \infty)$.

To finish the proof we show that $(1 - \|\sigma_S/\sigma\|_\infty)^{-1} \in L_p(\Omega)$, for all $p \in [0, \infty)$, from which the result follows. First note that

$$0 \ < \ \frac{\sigma_S}{\sigma} \ = \ \frac{\sigma_S}{\sigma_S + \sigma_A} \ = \ 1 - \frac{\sigma_A}{\sigma_S + \sigma_A} \ \le \ 1 - \frac{(\sigma_A)_{\min}}{(\sigma_S)_{\max} + (\sigma_A)_{\max}} \ < \ 1\,.$$

Then it follows that

$$\left(1 - \left\|\frac{\sigma_S}{\sigma}\right\|_\infty\right)^{-1} \le \ \left(1 - \left(1 - \frac{(\sigma_A)_{\min}}{(\sigma_S)_{\max} + (\sigma_A)_{\max}}\right)\right)^{-1} = \ \frac{(\sigma_S)_{\max} + (\sigma_A)_{\max}}{(\sigma_A)_{\min}}. \tag{5.4}$$

Now, by the Assumption 5.1, $\sigma_A = \sigma - \sigma_S \in L_p(\Omega, \mathrm{C}_{\eta,pw})$ and so $(\sigma_A)_{\max} \in L_p(\Omega)$, and so the numerator in (5.4) is $L_p(\Omega)$ for all $p \in [1,\infty)$. Assumption 5.1(b) allows us to estimate the denominator, and it follows that the second term in (5.3) is also in $L_p(\Omega)$ for all $p \in [1,\infty)$. The result follows again via the generalised Hölder inequality. □

We now establish a probabilistic counterpart of Theorem 4.10. To simplify the presentation, we assume the following relationship of the angular and spatial discretization parameters.

Assumption 5.4. For each mesh diameter h, we assume the number of angular quadrature points is $2N$, with

$$N \;=\; N(h) \;=\; \lceil c_0 h^{-\eta} \rceil \,, \tag{5.5}$$

for some constant $c_0 > 0$ independent of h and ω, where $\eta \in (0,1)$ is given in Assumption 5.1. We assume also that c_0 is chosen large enough so that $\log N(h) \geq 1$.

As a result of this assumption it is easily seen that $h \log N(h) \leq ch \log h^{-1}$, and hence

$$h \;\leq\; N(h)^{-1} + h \log N(h) + h^{\eta} \;\leq\; c'\, h^{\eta}, \tag{5.6}$$

for some (different) constant $c' > 1$ independent of h and ω.

Now recall the mesh-dependent stability condition (4.14) and note that Assumption 5.1 does not ensure that $\mathcal{R}_3(\sigma,\sigma_S) \in L_\infty(\Omega)$. Hence, for any fixed mesh size $h > 0$, it is impossible to guarantee that (4.14) is satisfied uniformly for all samples $\omega \in \Omega$. To deal with this problem we have to consider sample-dependent mesh sizes.

Definition 5.5. Let $h > 0$ be a deterministically chosen mesh diameter. For each $\omega \in \Omega$, let $h_\omega^{\max}$ be the largest possible value of h such that the stability condition (4.14) is satisfied with $\sigma_S = \sigma_S(\omega, x)$, $\sigma = \sigma(\omega, x)$ and $N = N(h_\omega^{\max})$ and let

$$h_\omega \;:=\; \min\{h, h_\omega^{\max}\}. \tag{5.7}$$

Solving the problem with mesh diameter h_ω is guaranteed to be stable and to have an accuracy determined by h. The resulting numerical solution (defined as in §2.1) is denoted:

$$\Phi^h(\omega, x) := \phi^{h_\omega, N(h_\omega)}(\omega, x) \,. \tag{5.8}$$

In Theorem 5.6, we quantify the accuracy of Φ^h as an approximation to ϕ in terms of the deterministic mesh diameter h. The question then arises: how significant is the set "Ω_{bad}", containing the samples ω which have to be computed using a finer mesh (i.e. where $h_\omega^{\max} < h$). We shall see in Lemma 5.8 that this set is small and decreases as $h \to 0$ in such a way as to ensure that the expected cost is not affected by the path-dependent stability criterion.

Theorem 5.6. *Under Assumptions 5.1 and 5.4, $\Phi^h(\omega,\cdot)$ exists for all $\omega \in \Omega$ and for any $1 \leq p < r \leq p_*$, there exists a positive constant $C_{p,r} > 0$ such that*

$$\|\phi - \Phi^h\|_{L_p(\Omega; L_\infty)} \;\leq\; C_{p,r}\, \|f\|_{L_r(\Omega; \mathrm{C}^\eta_{pw})}\, h^\eta \,. \tag{5.9}$$

Proof. Existence of Φ^h follows from the definition of h_ω. Then, using Theorem 4.10 together with (5.6) and Hölder's inequality, we obtain

$$\begin{aligned} \|\phi - \Phi^h\|^p_{L_p(\Omega;L_\infty)} &= \mathbb{E}\left[\|\phi - \phi^{h_\omega, N(h_\omega)}\|^p_\infty\right] \;\leq\; c^p \mathbb{E}\left[|\mathcal{R}(\sigma,\sigma_S)(\omega)|^p\, \|f(\omega,\cdot)\|^p_{\eta,pw} h_\omega^{p\eta}\right] \\ &\leq\; c^p \left(\mathbb{E}\left[|\mathcal{R}(\sigma,\sigma_S)(\omega)|^q\right]\right)^{p/q} \left(\mathbb{E}\left[\|f(\omega,\cdot)\|^r_{\eta,pw}\right]\right)^{p/r} h^{p\eta} \,, \end{aligned}$$

where $q^{-1} + r^{-1} = p^{-1}$, i.e., $q = pr/(r-p)$. The result follows with $C_{p,r} = c\|\mathcal{R}(\sigma,\sigma_S)\|_{L_q(\Omega)}$. □

Remark 5.7 (Uniform Random Input Data)**.** *If we strengthen Assumption 5.1 by requiring parts (a), (b) to hold for $p = \infty$ (i.e. uniformly bounded random fields), then $\mathcal{R}_3(\sigma, \sigma_S) \in L_\infty(\Omega)$. In this case there exists a deterministic $h_{\max}$ so that, when $h \le h_{\max}$, stability condition* (4.14) *holds uniformly over all samples, and we can choose $p = r = p_*$ in* (5.9)*. If, in addition, $p_* = \infty$ in Assumption 5.1(c), then $\|\phi - \Phi^h\|_{L_\infty(\Omega;L_\infty)} = \mathcal{O}(h^\eta)$. However, such a strengthening would rule out important cases: Assumption 5.1(a), (b) does not hold with $p = \infty$ for the lognormal fields considered in §5.3.*

For the general case of sample-dependent discretisations it is important to discuss the *expected computational cost* per sample. The following lemma shows that if the cost for computing each sample is of order $h_\omega^{-\gamma}$ with a sample dependent constant of proportionality which is sufficiently well-behaved and h_ω is given in (5.7) then the expected cost per sample is $\mathcal{O}(h^{-\gamma})$. Essentially this shows that the samples which need over-refinement to achieve stability (i.e. $h_\omega^{\max} \ll h$ in (5.7)) are small in measure. In Example 5.10 below, we give examples of solvers for (2.7)-(2.8) to which the lemma can be applied.

Lemma 5.8. *Let Assumptions 5.1 and 5.4 hold and let $\Phi^h(\omega, \cdot)$ be defined in* (5.8)*. Assume also that the cost $\mathcal{C}(\Phi^h)$ to compute one sample of Φ^h is bounded by*

$$\mathcal{C}(\Phi^h) \;\le\; C'(\omega) h_\omega^{-\gamma} \;, \tag{5.10}$$

with $C' \in L_p(\Omega)$, for some $p > 1$. Then $\mathbb{E}[\mathcal{C}(\Phi^h)] \le c h^{-\gamma}$. If $C' \in L_p(\Omega)$, for some $p > 2$, then we also have $\mathbb{V}[\mathcal{C}(\Phi^h)] \le c h^{-2\gamma}$.

Proof. For any sample ω, $h_\omega^{\max}$ is defined as the largest stepsize which satisfies (4.14). Hence inequality (4.14) must fail if we replace $h_\omega^{\max}$ by $2h_\omega^{\max}$, i.e.

$$\left((2h_\omega^{\max})^\eta + 2h_\omega^{\max} \log N(2h_\omega^{\max}) + N(2h_\omega^{\max})^{-1}\right)^{-1} \;< \mathcal{R}_3(\sigma, \sigma_S)(\omega).$$

Using (5.6), this ensures that $(h_\omega^{\max})^{-\eta} < 2^\eta c' \mathcal{R}_3(\sigma, \sigma_S)$. Then, using (5.10) and (5.7), we get

$$\begin{aligned} \mathcal{C}(\Phi^h) \;\le\; C'(\omega) h_\omega^{-\gamma} &= C'(\omega) \max\{h^{-\gamma},\, (h_\omega^{\max})^{-\gamma}\} \\ &\le C'(\omega) \left(h^{-\gamma} + 2^\gamma \left(c' \mathcal{R}_3(\sigma(\omega,\cdot), \sigma_S(\omega,\cdot))\right)^{\gamma/\eta}\right) \;. \end{aligned} \tag{5.11}$$

Now, taking the expectation of (5.11) and applying Hölder's inequality

$$\begin{aligned} \mathbb{E}\left[\mathcal{C}(\Phi^h)\right] \;&\le\; \mathbb{E}\left[C'(\omega)\right] h^{-\gamma} + c\,\mathbb{E}\left[C'(\omega) \mathcal{R}_3(\sigma(\omega,\cdot), \sigma_S(\omega,\cdot))^{\gamma/\eta}\right] \\ &\le\; \mathbb{E}\left[C'(\omega)\right] h^{-\gamma} + c\Big(\mathbb{E}\left[C'(\omega)^p\right]\Big)^{1/p} \Big(\mathbb{E}\left[\mathcal{R}_3(\sigma(\omega,\cdot), \sigma_S(\omega,\cdot))^{\gamma q/\eta}\right]\Big)^{1/q} \;, \end{aligned}$$

where $1 \le q < \infty$ is such that $p^{-1} + q^{-1} = 1$ and $c = 2^\gamma (c')^{\gamma/\eta}$. By a similar argument to that in Lemma 5.3, $\mathcal{R}_3(\sigma, \sigma_S)^{\gamma/\eta} \in L_q(\Omega)$ and so the result follows, since $h \le 1$.

The variance result follows in the same way upon noting that $\mathbb{V}[\mathcal{C}(\Phi^h)] \le \mathbb{E}\left[\mathcal{C}(\Phi^h)^2\right]$. □

Remark 5.9. The second term on the right-hand side of (5.11) is the contribution to the expected cost from samples for which the deterministically chosen mesh size h is not stable, and for which further mesh refinement is necessary (at least in theory). Since $h_\omega^{\max}$ is chosen to be the *largest* allowable mesh diameter which is stable, we can bound it both above and below in terms of the L_p integrable function $\mathcal{R}_3(\sigma, \sigma_S)$. Hence this second term remains bounded as $h \to 0$, giving the favourable complexity estimate proved in the lemma.

Example 5.10. In [26], two methods for computing the solution to (2.7)-(2.8) are presented. The corresponding system matrix is of dimension $\mathcal{O}(N(h_\omega) h_\omega^{-1}) = \mathcal{O}(h_\omega^{-1-\eta})$.

The first method is a direct solver where first ψ is eliminated from the coupled system (2.7)-(2.8) and then LU factorisation is applied to the resulting Schur complement system. The cost for this method is of order $h_\omega^{-2}\left(N(h_\omega)+h_\omega^{-1}\right)$ with a constant independent of ω [26]. Using (5.5) this implies that (5.10) holds with C' independent of ω and $\gamma = 3$.

The second method is a type of Richardson iteration known as *source iteration*. In that case, the cost is of order $h_\omega^{-1}N(h_\omega)$ with a constant proportional to $-\left[\log\left(\|\sigma_S(\omega,\cdot)/\sigma(\omega,\cdot)\|_\infty\right)\right]^{-1}$ [10, 11, 26]. Using again (5.5), this implies that (5.10) holds with $\gamma = 1+\eta$.

Corollary 5.11. *Suppose the assumptions of Lemma 5.8 hold and system* (2.7)-(2.8) *is solved with either of the Methods 1 or 2 in Example 5.10. Then, condition* (5.10) *holds with* $C' \in L_\infty(\Omega)$ *(Method 1) and* $C' \in L_p(\Omega)$*, for all* $1 \le p < \infty$ *(Method 2). Hence,*

$$\begin{aligned}
&\mathbb{E}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-3}\right) \quad \text{and} \quad \mathbb{V}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-6}\right) \quad \text{for Method 1, and}\\
&\mathbb{E}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-(1+\eta)}\right) \quad \text{and} \quad \mathbb{V}[\mathcal{C}(\Phi^h)] = \mathcal{O}\left(h^{-2(1+\eta)}\right) \quad \text{for Method 2,}
\end{aligned}$$

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, $\|\sigma_S(\omega,\cdot)/\sigma(\omega,\cdot)\|_\infty \in (0,1)$, for almost all $\omega \in \Omega$, and since $-\log(y) > 1-y$, for all $y \in (0,1)$, we have

$$C'(\omega) \;\le\; -c\left[\log\left(\left\|\frac{\sigma_S(\omega,\cdot)}{\sigma(\omega,\cdot)}\right\|_\infty\right)\right]^{-1} \;<\; c\left(1-\left\|\frac{\sigma_S(\omega,\cdot)}{\sigma(\omega,\cdot)}\right\|_\infty\right)^{-1}.$$

As we have seen in (5.4), this implies that $C' \in L_p(\Omega)$, for all $p \in (0,\infty]$. The bounds on the expected values and on the variances follow from Lemma 5.8. □

For solving the linear systems arising from the transport equation in this paper we use standard source iteration, and this corollary indicates the efficiency of this method for the problems considered here. More efficient and flexible solvers such as those in [31] could be used in more general situations.

5.2 Multilevel Monte Carlo Methods

In this subsection we will consider the application of MLMC techniques to compute functionals of the scalar flux ϕ. First we recall some general results on MLMC. Suppose $Q = Q(\omega)$ is a random variable, whose expected value we wish to compute, and suppose $Q_h(\omega)$ is an approximation of $Q(\omega)$ which becomes more accurate as the spatial mesh size $h \to 0$. With $\widehat{Q}_h$ denoting an unbiased estimator for $\mathbb{E}[Q_h]$ (i.e. $\mathbb{E}[\widehat{Q}_h] = \mathbb{E}[Q_h]$), the mean square error $e(\widehat{Q}_h)$ in approximating $\mathbb{E}[Q]$ with $\widehat{Q}_h$ is given by

$$e(\widehat{Q}_h)^2 \;:=\; \mathbb{E}\left[(\widehat{Q}_h - \mathbb{E}[Q])^2\right] \;=\; \left(\mathbb{E}\left[Q - Q_h\right]\right)^2 \;+\; \mathbb{V}[\widehat{Q}_h]\,, \tag{5.12}$$

the first term being the square of the *bias* due to discretization, and the second being the *sampling error* $\mathbb{V}[\widehat{Q}_h] = \mathbb{E}[(\widehat{Q}_h - \mathbb{E}[\widehat{Q}_h])^2]$.

In order to compare various estimators $\widehat{Q}_h$, we define, for any $\epsilon \in (0,1)$, the ϵ-cost $\mathcal{C}_\epsilon$ to be the number of (floating point) operations to achieve $e(\widehat{Q}_h)^2 \le \epsilon^2$, a sufficient condition for this being that each of the terms on the right-hand side of (5.12) should be bounded by $\epsilon^2/2$.

The standard Monte Carlo (MC) estimator for $\mathbb{E}[Q]$ with N_{MC} samples is

$$\widehat{Q}_h^{MC} \;:=\; \frac{1}{N_{MC}}\sum_{n=1}^{N_{MC}} Q_h(\omega^{(n)})\,. \tag{5.13}$$

The sampling error is $\mathbb{V}[\widehat{Q}_h^{MC}] = \mathbb{V}[Q_h]/N_{MC}$ and since $\mathbb{V}[Q_h]$ is bounded as $h \to 0$, bounding this by $\epsilon^2/2$ requires $N_{MC} \sim \epsilon^{-2}$.

To go further with the analysis one must make assumptions about the accuracy of the approximation $Q_h \approx Q$ and the cost $\mathcal{C}(Q_h(\omega))$ of computing a single sample of Q_h. Following [13] we assume that there exist two constants $\alpha, \gamma > 0$ such that

$$\left|\mathbb{E}\left[Q - Q_h\right]\right| \;=\; \mathcal{O}\left(h^{\alpha}\right) \;, \tag{5.14}$$

$$\mathbb{E}\left[\mathcal{C}(Q_h)\right] \;=\; \mathcal{O}\left(h^{-\gamma}\right) \;. \tag{5.15}$$

Then, to achieve an error of order ϵ in the "bias" (5.14), we need to take $h \sim \epsilon^{1/\alpha}$, leading, via (5.15), to a mean cost per sample of order $\epsilon^{-\gamma/\alpha}$. It follows immediately that there exists a constant $c_\mu > 0$ independent of ϵ such that the mean ϵ-cost of the standard Monte Carlo estimator satisfies

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\right] \;=\; \mathbb{E}\left[\sum_{n=1}^{N_{MC}} \mathcal{C}(Q_h(\omega^{(n)}))\right] \;=\; N_{MC}\,\mathbb{E}\left[\mathcal{C}(Q_h)\right] \;\leq\; c_\mu\,\epsilon^{-2-\frac{\gamma}{\alpha}} \;. \tag{5.16}$$

In fact, this result can be strengthened, as we now show.

Theorem 5.12. *Let $\delta \in (0,1)$. In addition to* (5.14) *and* (5.15)*, we assume that*

$$\mathbb{V}\left[\mathcal{C}(Q_h)\right] \;=\; \mathcal{O}\left(h^{-2\gamma}\right) \;. \tag{5.17}$$

Then there exist constants c_μ and c_σ such that, for any $\epsilon \leq \epsilon_{\max} \leq 1$,

$$\mathbb{P}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC}) < \left(c_\mu + c_\sigma \epsilon_{\max}\delta^{-1}\right)\epsilon^{-2-\frac{\gamma}{\alpha}}\right] \;>\; 1-\delta^2.$$

Proof. Let $X \in \mathbb{R}^+$ be any positive random variable with $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$. For any $\delta \in (0,1)$, the Chebyshev Inequality implies that

$$1-\delta^2 \;<\; \mathbb{P}\left[|X-\mu| < \frac{\sigma}{\delta}\right] \;\leq\; \mathbb{P}\left[X < \mu + \frac{\sigma}{\delta}\right] \;. \tag{5.18}$$

Now, choosing $X = \mathcal{C}_\epsilon(\widehat{Q}_h^{MC})$ and using the choices $N_{\rm MC} \sim \epsilon^{-2}$, $h \sim \epsilon^{1/\alpha}$, an estimate for $\mu = \mathbb{E}[X]$ is given in (5.16). Moreover, using in addition, (5.17), we have

$$\sigma^2 \;=\; \mathbb{V}\left[X\right] \;=\; \mathbb{V}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\right] = \mathbb{V}\left[\sum_{n=1}^{N_{MC}} \mathcal{C}(Q_h(\omega^{(n)}))\right] \;=\; N_{MC}\,\mathbb{V}\left[\mathcal{C}(Q_h)\right] \;\leq\; \left(c_\sigma\,\epsilon^{-1-\frac{\gamma}{\alpha}}\right)^2 \;.$$

The result then follows by inserting these estimates in (5.18). □

Thus, the ϵ-cost of a particular realisation of the standard Monte Carlo estimator $\widehat{Q}_h^{MC}$ is $\mathcal{O}(\epsilon^{-2-\frac{\gamma}{\alpha}})$ with probability $1-\delta^2$, for any $\delta \in (0,1)$, i.e., arbitrarily close to 1 and not just in mean. In general, the asymptotic constant $c_\delta := c_\mu + c_\sigma\epsilon_{\max}\delta^{-1}$ blows up, as $\delta \to 0$, but for $\epsilon_{\max} = \mathcal{O}(\delta)$, c_δ can be bounded independently of δ.

To reduce the high cost of the MC method, 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 $h_0 > h_1 > ... > h_L$. By choosing $h_L = h \sim \epsilon^{1/\alpha}$ as above, the most accurate model on level L is designed to provide full bias accuracy of $\mathcal{O}(\epsilon)$. However, the samples on the coarser grids can be used as control variates. Writing

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

each of the expected values on the right hand side is then estimated separately. In particular, using a standard MC estimator with N_ℓ samples for the ℓth term, we obtain the MLMC estimator

$$\widehat{Q}_h^{MLMC} \;:=\; \sum_{\ell=0}^{L} \widehat{Y}_\ell^{MC} \;=\; \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} Y_\ell(\omega^{(\ell,n)}) \;. \tag{5.19}$$

Here, the notation $\{\omega^{(\ell,n)}\}_{n=1}^{N_\ell}$ means that N_ℓ i.i.d. samples are chosen on level ℓ, independently from the samples on the other levels.

Since Q_{h_ℓ} and $Q_{h_{\ell-1}}$ were both assumed to converge in mean to Q as $h_{\ell-1} \to 0$, it follows that $\mathbb{E}[Y_\ell] \to 0$. To achieve a reduced cost for the MLMC estimator we need the additional assumption that there exists a $\beta > 0$ such that

$$\mathbb{V}[Y_\ell] \;=\; \mathcal{O}\left(h_\ell^\beta\right). \tag{5.20}$$

For this reason, MLMC is often referred to as a *variance reduction method*. In Theorem 5.14 below, we shall give a simple sufficient condition for (5.20) to hold in our context.

The following theorem is a simple extension of [13, Thm. 1] to the random cost case (see also [19]). As in [13], we assume for simplicity that there exists a $q \in (0,1)$ such that

$$h_\ell = q h_{\ell-1}, \qquad \text{for all } \; \ell = 1, \dots, L.$$

Theorem 5.13. *Assume that* (5.14), (5.15), (5.20) *hold with* $\alpha, \beta, \gamma > 0$ *and* $\alpha \geq \frac{1}{2}\min\{\beta, \gamma\}$. *Then, for any* $\epsilon < e^{-1}$, *there exist choices* $L \sim \log(\epsilon^{-1})$ *and* $\{N_\ell\}_{\ell=0}^L$ *such that* $e(\widehat{Q}_h^{MLMC})^2 \leq \epsilon^2$ *and*

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\right] \;\leq\; c_\mu\, \epsilon^{-2-\max\{0,\ (\gamma-\beta)/\alpha\}}\,, \quad \textit{for } \beta \neq \gamma\,, \tag{5.21}$$

with $c_\mu > 0$ *independent of* ϵ. *For* $\beta = \gamma$, *we can achieve* $\mathbb{E}\big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\big] \leq c_\mu\, \epsilon^{-2}(\log \epsilon)^2$.

Let $\delta \in (0,1)$ *and let us assume in addition that* (5.17) *holds. Then there exists a constant* $c_\sigma > 0$ *independent of* ϵ *and* δ *such that*

$$\mathbb{P}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC}) \leq (c_\mu + c_\sigma \delta^{-1})\, \epsilon^{-2-\max\{0,\ (\gamma-\beta)/\alpha\}}\right] \;>\; 1-\delta^2\,, \quad \textit{for } \beta \neq \gamma\,, \tag{5.22}$$

with $c_\mu > 0$ *as above. For* $\beta = \gamma$, *we have* $\mathbb{P}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC}) \leq (c_\mu + c_\sigma \delta^{-1})\, \epsilon^{-2}(\log \epsilon)^2\right] > 1 - \delta^2$.

Proof. As in [19], the proof of (5.21) follows easily from [13, Append. A]. Due to the independence of the samples $\omega^{(\ell,n)}$, there exists $c'_\mu > 0$ independent of ϵ such that

$$\mathbb{E}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\right] \;=\; \mathbb{E}\left[\sum_{\ell=0}^{L}\sum_{n=1}^{N_\ell} \mathcal{C}\left(Q_{h_\ell}(\omega^{(\ell,n)})\right) + \mathcal{C}\left(Q_{h_{\ell-1}}(\omega^{(\ell,n)})\right)\right] \;\leq\; c'_\mu \sum_{\ell=0}^{L} N_\ell h_\ell^{-\gamma}\,, \tag{5.23}$$

i.e., the same asymptotic bound as in the deterministic case, and the result follows as in [13] with identical choices for L and $\{N_\ell\}_{\ell=0}^L$.

To prove (5.22), we exploit again the independence of the samples and show as in (5.23), with the same values for L and $\{N_\ell\}_{\ell=0}^L$, that there exist $c'_\sigma, c_\sigma > 0$ independent of ϵ such that

$$\mathbb{V}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\right] \;\leq\; (c'_\sigma)^2 \sum_{\ell=0}^{L} N_\ell h_\ell^{-2\gamma} \;\leq (c_\sigma)^2 \begin{cases} \epsilon^{-2-\max\{0,\ (2\gamma-\beta)/\alpha\}} & \text{for } \beta \neq 2\gamma\,, \\ \epsilon^{-2}(\log\epsilon)^2 & \text{for } \beta = 2\gamma\,. \end{cases} \tag{5.24}$$

The second estimate in (5.24) follows as in [13, Theorem 1] after replacing γ with 2γ.

The result (5.22) then follows as in the proof of Theorem 5.12 via Chebyshev's Inequality. To see this, consider (5.18) with $X = \mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})$ and assume first that $\beta < \gamma$. Using the bounds on the expected value and variance of X in (5.21) and (5.24) we get

$$\mathbb{P}\left[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC}) \le c_\delta\, \epsilon^{-2-(\gamma-\beta)/\alpha}\right] \;>\; 1-\delta^2$$

with

$$c_\delta \;:=\; c_\mu + c_\sigma \epsilon^{1-\frac{\beta}{2\alpha}}\, \delta^{-1} \;\le\; c_\mu + c_\sigma \delta^{-1}\,,$$

since $\epsilon < e^{-1}$ and $\alpha \ge \frac{1}{2}\beta$.

The cases $\beta = \gamma$, $\beta \in (\gamma, 2\gamma)$, $\beta = 2\gamma$ and $\beta > 2\gamma$ can all be shown similarly. □

Theorem 5.13 states that provided (5.20) holds for some $\beta > 0$, the MLMC always achieves a gain of $\mathcal{O}\left(\epsilon^{-\min\{\beta,\gamma\}/\alpha}\right)$ in the asymptotic cost over standard Monte Carlo, even in the case when the cost per sample is random and with probability arbitrarily close to 1. For sufficiently large β, the cost of the MLMC method is $\mathcal{O}\left(\epsilon^{-2}\right)$, independent of α. This fact can be exploited to design unbiased multilevel estimators of $\mathbb{E}[Q]$ with cost $\mathcal{O}\left(\epsilon^{-2}\right)$ [40]. On the other hand, if $\gamma > \beta = 2\alpha$, the cost of the MLMC method is $\mathcal{O}\left(\epsilon^{-\gamma/\alpha}\right)$ which 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 accuracy $\mathcal{O}(\epsilon)$.

Application to Neutron Transport. Suppose now that $\mathcal{Q} : \mathrm{C} \to \mathbb{R}$ is a (linear or nonlinear) functional (operating with respect to the spatial variable x), and we are interested in computing the expected value of $Q(\omega) := \mathcal{Q}(\phi(\omega,\cdot))$ where ϕ is the scalar flux satisfying (2.1) and (2.3). This will be approximated by $Q_h(\omega) := \mathcal{Q}(\Phi^h(\omega,\cdot))$, with Φ^h as defined in Theorem 5.6.

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 5.11, taking into account the sample-dependent mesh size. The following result gives estimates for α and β under a quite general assumption on $\mathcal{Q}$.

Theorem 5.14. *Make the same assumptions as in Theorem 5.6 but assume $p_* > 2$ in Assumption 5.1. Let $2 < p < p_*$ and let $q = 2p/(p-2)$. Suppose, in addition, $\mathcal{Q}$ satisfies the Lipschitz condition:*

$$|\mathcal{Q}(\phi(\omega,\cdot)) - \mathcal{Q}(\tilde{\phi}(\omega,\cdot))| \;\le C'(\omega)\|\phi(\omega,\cdot) - \tilde{\phi}(\omega,\cdot)\|_\infty, \quad \textit{for all} \quad \phi,\ \tilde{\phi} \in L^p(\Omega, L_\infty)\,,$$

where $C' \in L^q(\Omega)$. Then, (5.14) *and* (5.20) *hold with*

$$\alpha \;=\; \eta, \quad \textit{and} \quad \beta \;=\; 2\eta\,.$$

Proof. From the given hypothesis and Hölder's inequality

$$\mathbb{E}[|Q(\omega) - Q_h(\omega)|] \;=\; \mathbb{E}[|\mathcal{Q}(\phi(\omega,\cdot)) - \mathcal{Q}(\Phi^h(\omega,\cdot))|] \;\le\; \|C'\|_{L^{q'}(\Omega)}\, \|\phi - \Phi^h\|_{L^p(\Omega,L_\infty)},$$

where $q' = p/(p-1) < q$ and (5.14) with $\alpha = \eta$ follows from Theorem 5.6.

Also, with $Y_\ell = Q_{h_\ell} - Q_{h_{\ell-1}}$, we have

$$\mathbb{V}[Y_\ell] \;\le\; \mathbb{E}\big[Y_\ell^2\big] \;\le\; 2\left(\mathbb{E}\left[|Q - Q_{h_\ell}|^2\right] + \mathbb{E}\big[\,\big|Q - Q_{h_{\ell-1}}\big|^2\,\big]\right).$$

Arguing as before,

$$\mathbb{E}\left[|Q(\omega) - Q_{h_\ell}(\omega)|^2\right] = \mathbb{E}\left[\left|\mathcal{Q}(\phi(\omega,\cdot) - \mathcal{Q}(\Phi^{h_\ell}(\omega,\cdot))\right|^2\right] \le\; \|C'\|^2_{L^q(\Omega)}\, \|\phi - \Phi^{h_\ell}\|^2_{L^p(\Omega,L_\infty)},$$

where we used Hölder's inequality with conjugate indices $p/2$ and $q/2$. Then (5.20) follows with $\beta = 2\eta$. □

Example 5.15. Consider the qth moment of the spatial average of ϕ (over $[0,1]$):

$$Q(\omega) \;=\; \|\phi\|_1^q \;:=\; \left(\int_0^1 |\phi(\omega,x)|\;dx\right)^q\;, \quad \text{for some integer } q\geq 1\;. \tag{5.25}$$

This satisfies the assumptions of Theorem 5.14. The details are given in [36].

Corollary 5.16. *Suppose the assumptions of Theorem 5.14 hold and system* (2.7)-(2.8) *is solved with Method 2 in Example 5.10. Then, the ϵ-costs of the Monte Carlo method and of the multilevel Monte Carlo method satisfy, respectively,*

$$\mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MC})\Big] \;=\; \mathcal{O}\big(\epsilon^{-4-\chi}\big) \quad \textit{and} \quad \mathbb{E}\Big[\mathcal{C}_\epsilon(\widehat{Q}_h^{MLMC})\Big] \;=\; \mathcal{O}\big(\epsilon^{-2-\chi}\big)\;,$$

where $\chi := \dfrac{1-\eta}{\eta} > 0$.

This corollary shows that indeed in theory, for the neutron transport problem, a theoretical gain in the asymptotic computational cost of up to two orders of magnitude in ϵ^{-1} is possible on average. In fact, since we have also established a bound on the variance of the cost of Methods 1 and 2 in Corollary 5.11, we could even deduce such a result with probability arbitrarily close to 1 from Theorems 5.12 and 5.13. However, in the numerical section we will see that this theoretical result is overly optimistic, since in particular the bound on α in Theorem 5.14 is not sharp. Nevertheless, we do observe gains of (at least) one order of magnitude.

Similar results can also be shown for other functionals of ϕ that are bounded in $L^p(\Omega, L_\infty)$.

5.3 Numerical Results

In this section, we give some numerical experiments for the case when $f(x) = e$, for all $x \in (0,1)$, and when $\sigma = \sigma_S + \sigma_A$ with fixed absorption cross-section $\sigma_A = \exp(0.5)$ and scattering cross-section σ_S chosen to be a lognormal random field with Matérn covariance, i.e., $\log\sigma_S$ is a centred Gaussian random field, with covariance function:

$$C_\nu(x,y) \;=\; \sigma_{var}^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(2\sqrt{\nu}\frac{|x-y|}{\lambda_C}\right)^\nu K_\nu\left(2\sqrt{\nu}\frac{|x-y|}{\lambda_C}\right)\;. \tag{5.26}$$

Here, Γ is the gamma function, K_ν is the modified Bessel function of the second kind, λ_C is the correlation length and σ_{var}^2 is the variance. By increasing the positive parameter ν we can increase the smoothness of realizations (see, e.g. [25]).

To sample from σ_S we use the Karhunen-Loève (KL) expansion of $\log\sigma_S$, i.e.,

$$\log\sigma_S(x,\omega) \;=\; \sum_{i=1}^{\infty}\sqrt{\xi_i}\;\eta_i(x)\;Z_i(\omega)\;, \tag{5.27}$$

where $Z_i \sim \mathcal{N}(0,1)$ i.i.d., and ξ_i and η_i are the eigenvalues and the $L_2(0,1)$-orthogonal eigenfunctions of the integral operator induced by the kernel (5.26). In practice, the KL expansion needs to be truncated after a finite number of terms, and the accuracy of the truncated expansion depends on the rate of decay of the eigenvalues – this rate gets faster as ν increases – see, e.g. [35, 25].

We will give experiments for the cases $\nu = 0.5$ (when the ξ_i and η_i are known analytically [35]), and $\nu = 1.5$ (where ξ_i and η_i are computed using the Nyström method - see, for example, [18]). When $\nu = 0.5$, it is known that Assumption 5.1 holds with $\eta < 0.5$. When $\nu = 1.5$ then realizations of σ_S have Hölder continuous first derivative, and hence Assumption 5.1 holds for all $\eta < 1$ – see, e.g., [12, 25]. In our experiments we set $\lambda_C = 1$ and $\sigma_{var}^2 = 1$.

We discretise using the method described in §2.1. Following (5.5), we set the angular discretisation level at $N = 2\lceil 2h^{-1/2}\rceil$ when $\nu = 0.5$ and $N = (2h)^{-1}$ when $\nu = 1.5$. We truncate

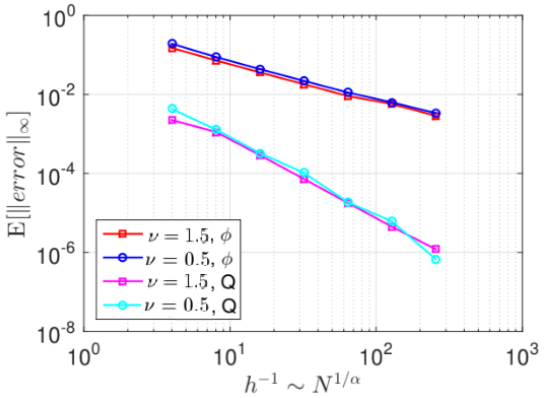

Figure 1: Convergence of the mean error(s) $\mathbb{E}[\|\phi - \phi^{h,N(h)}\|_\infty]$ and $\mathbb{E}[|Q - Q_h|]$

the KL expansion after $225\lceil h^{-1/2}\rceil$ terms when $\nu = 0.5$ and after $8\lceil h^{-1}\rceil$ terms when $\nu = 1.5$, which ensures in both cases that the truncation error is negligible compared to the discretisation error.

We compute two measurements of error in the mean,

$$\mathbb{E}[\|\phi - \phi^{h,N(h)}\|_\infty] \quad \text{and} \quad \mathbb{E}[|Q - Q_h|] \tag{5.28}$$

(where Q is defined in (5.25) and we take $q = 1$). To compute these, we estimate ϕ by a reference solution with $h^{-1} = 512$, $N = 256$, and we choose 3600/ 2048 KL modes for the cases $\nu = 0.5, 1.5$ respectively. The expectations in (5.28) are estimated using a standard Monte Carlo estimator (cf. (5.13)) with $32,768$ samples. Note that when computing $\phi^{h,N(h)}(\omega, \cdot)$ we ignore the theoretical path dependent stability criterion (4.14), which led to the construction (5.7), and simply compute solutions with mesh parameters h and $N(h)$ for each sample.

Numerical computations of (5.28) are presented in Figure 1. For $\mathbb{E}[\|\phi - \phi^{h,N(h)}\|_\infty]$ we observe $\mathcal{O}(h)$ convergence, for both $\nu = 0.5, 1.5$, even though when $\nu = 0.5$ we are only able to prove convergence of order $\mathcal{O}(h^\eta)$ with $\eta < 0.5$. We also observe smaller errors and a faster convergence rate (close to $\mathcal{O}(h^2)$) for the error in the functional $\mathbb{E}[|Q - Q_h|]$.

Our final set of results concern the ϵ-cost of the standard (MC) and multilevel (MLMC) Monte Carlo methods for computing $\mathbb{E}[Q]$ where $Q(\omega) = Q(\phi(\omega, \cdot))$ and Q is given by (5.25) with $q = 1$. We use Method 2 of Example 5.10 as the linear system solver for each realisation. Then Corollary 5.16 gives theoretical projections for the ϵ-costs for each of these methods in terms of η. The relevant values of η are $\eta < 0.5$ when $\nu = 0.5$ and $\eta < 1$ when $\nu = 1.5$, in which case $\chi > 1$ (for $\nu = 0.5$) and $\chi > 0$ (for $\nu = 1.5$). Hence the theoretical ϵ-costs given by Corollary 5.11 are

$$\mathcal{O}(\epsilon^{-s}) \tag{5.29}$$

with s as given in the column "s, `theory`" in Table 1. To compare these to the observed ϵ-costs, we give in the column "s, `observed`" the corresponding observed rates of growth of ϵ-cost when estimating $\mathbb{E}[Q]$, using the data which went into the construction of Figure 2.

The graphs in Figure 2 depict the growth in ϵ-cost for each of the methods and each ν in the case when $h = 1/512$ and $N = N(h)$ and show the superiority of the multilevel method. We observe from Table 1 and Figure 2 that for both values of ν, the MLMC method gives us excellent gains over the MC method in practice, of at least one order of magnitude. The discrepancy between the theory and numerics here arises because, for the two specific cases considered, the observed value of α in (5.14) is somewhat higher than the theoretically predicted value of $\alpha = \eta$ (with $\eta < 0.5$ when $\nu = 0.5$ and $\eta < 1$ when $\nu = 1.5$).

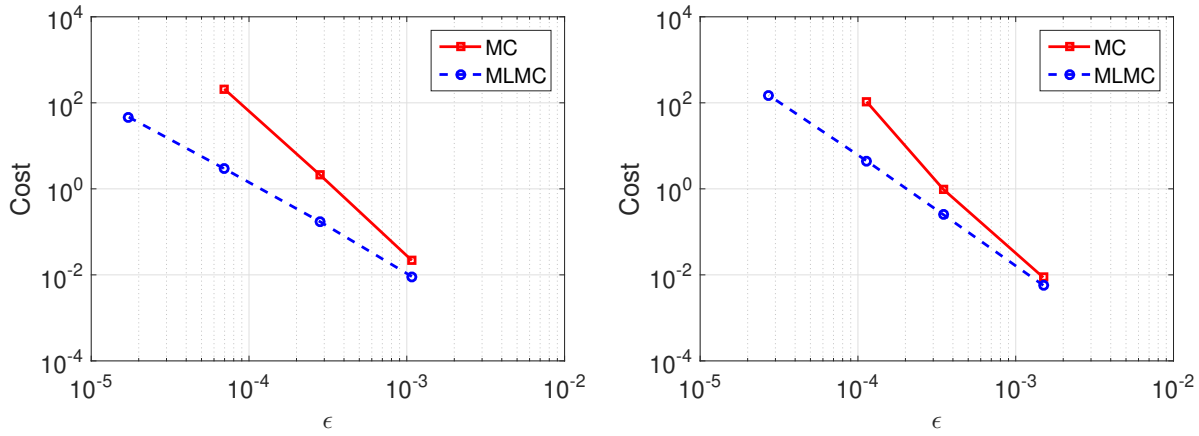

Figure 2: Cost (in seconds) plotted against ϵ for the error (computed with respect to a reference solution) for standard and multilevel Monte Carlo. (Left) $\nu = 1.5$ and (Right) $\nu = 0.5$.

ν	Method	s, `theory`	s, `observed`
0.5	MC	> 5	3.4
0.5	MLMC	> 3	2.4
1.5	MC	> 4	3.3
1.5	MLMC	> 2	2.1

Table 1: Summary of computational ϵ-cost rates with s as in (5.29).

6 Conclusion

We have given a novel error analysis for the discretised heterogeneous transport equation, demonstrating how the error depends on the heterogeneity. Although this is done for the 1d space and 1d angle case (slab geometry) and a classical discrete ordinates discretisation, the analysis could be extended to higher dimensional cases, although with considerably more technicalities. This analysis is based on a careful analytical treatment of a certain underlying integral equation. We then applied this analysis to the case when the cross-sections are given by a random fields and presented error estimates in suitable Bochner norms for both the scalar flux and for functionals of it. We assumed the input data could be piecewise continuous Hölder fields with low regularity. We outlined the Monte Carlo and multilevel Monte Carlo methods for quantifying the propagation of uncertainty in this model problem. Using our probabilistic error estimates we then rigorously proved estimates for the cost of these methods. These predict the superiority of the multilevel methods, even in the case of very rough input data. Finally we presented numerical results to support the theory. Further numerical investigation of uncertainty quantification for the transport equation, including some 2d in space and 1d in angle model problems is given in the PhD thesis [36].

Acknowledgement. Matthew Parkinson was supported by the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath, under project EP/L015684/1. We thank EPSRC and Wood plc. for financial support for this project and we particularly thank Professor Paul Smith of the Answers Software Team for many helpful discussions. This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath.

References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*, Dover, NY, 1964.

[2] M.L. Adams and E.W. Larsen. Fast iterative methods for discrete-ordinates particle transport calculations. *Prog. Nucl. Energy*, 40(1):3–159, 2002.

[3] E.J. Allen, H.D. Victory Jr, and K. Ganguly. On the convergence of finite-differenced multigroup, discrete-ordinates methods for anisotropically scattered slab media. *SIAM J. Numer. Anal.*, 26(1):88–106, 1989.

[4] M. Asadzadeh, P. Kumlin, and S. Larsson. The discrete ordinates method for the neutron transport equation in an infinite cylindrical domain. *Math. Models Meth. Appl. Sci.* 2(3):317-338, 1992.

[5] 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.

[6] M. Asadzadeh and L. Thevenot. On discontinuous Galerkin and discrete ordinates approximations for neutron transport equation and the critical eigenvalue. *Il Nuovo Cimento*, 33(1):21-29, 2010.

[7] 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.

[8] 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.

[9] G.I. Bell and S. Glasstone. *Nuclear Reactor Theory*, Van Nostrand Reinhold, NY, 1970.

[10] J.C.H. Blake. *Domain Decomposition Methods for Nuclear Reactor Modelling with Diffusion Acceleration.* PhD thesis, University of Bath, 2016.

[11] J.C.H. Blake, I.G. Graham, F. Scheben and A. Spence The radiative transport equation with heterogeneous cross-sections, In *On the Frontiers of High-Dimensional Computation, F. Kuo Guest Editor*, 2018 Matrix Annals, D.R. Wood, J. de Gier, C. Praeger, and T. Tao (Eds.), Springer, 2019. `https://arxiv.org/abs/1903.08623`

[12] 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.

[13] 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. Visual. Sci.*, 14(1):3–15, 2011.

[14] W. Dahmen, F. Gruber, and O. Mula, An adaptive nested source term iteration for radiative transfer equations, *Math. Comput.*, 89(324):1605–1646, 2020.

[15] R. Dautray and J.L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 6: Evolution Problems II.* Springer-Verlag, 2012.

[16] 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.

[17] 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.

[18] M. Eiermann, O.G. Ernst, and E. Ullmann. Computational aspects of the stochastic finite element method. *Comput. Visual. Sci.*, 10(1):3–15, 2007.

[19] A. Ferreiro-Castilla, A.E. Kyprianou, R. Scheichl, and G. Suryanarayana. Multilevel Monte Carlo simulation for Lévy processes based on the Wiener–Hopf factorisation. *Stoch. Proc. Appl.*, 124(2):985–1010, 2014.

[20] 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.

[21] C. Führer and R. Rannacher, Error analysis for the finite element approx- imation of a radiative transfer model. *RAIRO Model. Math. Anal. Numr.* 30(6):743–762, 1996.

[22] M.B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[23] 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.

[24] L. Giret. *Numerical Analysis of a Non-Conforming Domain Decomposition for the Multi-group SPN Equations.* PhD Thesis, Université Paris-Saclay, 2018.

[25] 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.

[26] I.G. Graham, M.J. Parkinson, and R. Scheichl. Modern Monte Carlo variants for uncertainty quantification in neutron transport. In "*Festschrift for the 80th Birthday of Ian Sloan*", J. Dick, F.Y. Kuo, and H. Wozniakowski (Eds.), Springer-Verlag, 2018.

[27] I.G. Graham, M.J. Parkinson, and R. Scheichl. Full error analysis and uncertainty quantification for the heterogeneous transport equation in slab geometry, Preprint arXiv:1903.11838, 2019 (extended version of this paper with additional detailed proofs).

[28] S. Heinrich. Multilevel Monte Carlo methods. In: S. Margenov, J. Waśniewski, P. Yalamov (Eds.), Large-Scale Scientific Computing, *Lecture Notes in Computer Science* 2179:58–67, Springer-Verlag, Berlin, 2001.

[29] S. Jin and D. Levermore, The discrete-ordinate method in diffusive regime, *Transport Theor. Stat.* 20:413–439, 1991.

[30] S. Jin, H. Lu and L. Pareschi. Efficient stochastic asymptotic-preserving IMEX methods for transport equations with diffusive scalings and random inputs. *SIAM J. Sci. Comput.*, 40:A671–A696, 2018.

[31] G. Kanschat and J.C. Ragusa. A robust multigrid preconditioner for SNDG approximation of monochromatic, isotropic radiation transport problems. *SIAM J. Sci. Comput.*, 36(5): 2326–2345, 2014.

[32] H.B. Keller. On the pointwise convergence of the discrete-ordinate method. *SIAM J. Appl. Math.*, 8(4):560–567, 1960.

[33] 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.

[34] E.E. Lewis and W.F. Miller. *Computational Methods of Neutron Transport.* American Nuclear Society, 1993.

[35] G.J. Lord, C.E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs.* Cambridge University Press, 2014.

[36] M.J. Parkinson, *Uncertainty Quantification in Radiative Transport*, PhD Thesis, University of Bath, 2018.

[37] J. Pitkäranta and L.R. Scott. Error estimates for the combined spatial and angular approximations of the transport equation in slab geometry. *SIAM J. Numer. Anal.*, 20(5):922–950, 1983.

[38] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Los Alamos Report LA-UR-73-479, 1973.

[39] K. Ren. Recent developments in numerical techniques for transport-based medical imaging methods. *Commun. Comput. Phys.*, 8(1):1–50, 2010.

[40] C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.

[41] R. Sanchez and N.J. McCormick. Review of neutron transport approximations. *Nucl. Sci. Eng.*, 80(4):481-535, 1982.

[42] 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.

[43] 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.

[44] B. Zhang, H. Liu and S. Jin. An asymptotic preserving Monte Carlo method for the multispecies Boltzmann equation , *J. Comput. Phys.* 305: 575-588, 2016.

[45] X. Zhong and Q. Li. Galerkin Methods for stationary radiative transfer equations with uncertain coefficients, *J. Sci. Comput.*, 76: 1105–1126, 2018.