A LEAST-SQUARES FINITE ELEMENT METHOD FOR THE LINEAR BOLTZMANN EQUATION WITH ANISOTROPIC SCATTERING*

TRAVIS M. AUSTIN † and THOMAS A. MANTEUFFEL ‡

Abstract. Least-squares methods have been applied to a wide range of differential equations and have been established to be competitive with other existing discretization strategies [P. B. Bochev and M. D. Gunzburger, SIAM Rev., 40 (1998), pp. 789–837]. In this article, we consider a leastsquares method for the linear Boltzmann equation with anisotropic scattering. A similar method has already been developed, and extensively examined, for the linear Boltzmann equation with isotropic scattering. The success of the least-squares method for isotropic scattering depends on scaling the linear Boltzmann equation so that minimization of the least-squares functional in a discrete space always yields accurate discrete solutions. A similar scaling of the linear Boltzmann equation is employed for anisotropic scattering. In the previous work for isotropic scattering, coercivity and continuity results were established for the scaled least-squares functional relative to a physically reasonable norm. In this paper, we extend the previous coercivity and continuity results so that they hold in this more general case of anisotropic scattering. Additionally, we extend the bounds for the discretization error for the thin regime and for the thick regime. For the thick regime, we establish optimal error estimates for the case of highly anisotropic scattering.

Key words. least-squares, neutron transport, anisotropic scattering, finite elements

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/040610519

1. Introduction. In this paper, we examine a least-squares method that is used to obtain discrete solutions to the single-group, steady-state linear Boltzmann equation with anisotropic scattering. The least-squares method for isotropic scattering was carefully analyzed for slab geometry in [12] and for xyz-geometry in [13]. Here, in the context of xyz-geometry, we extend the generality of the least-squares method by allowing for anisotropic scattering, whereby a particle has a preferential direction of scatter after collision.

For isotropic scattering, where particles have no preferential direction of scatter, it was proved in [12, 13] that the least-squares method yields discrete solutions that exhibit the correct asymptotic behavior in the diffusion limit. In this limit, the leadingorder asymptotic solution of the Boltzmann equation converges to the solution of a diffusion equation. In [13], ellipticity of the least-squares functional was proved and the existence of optimal error estimates for a P_N angular discretization and a finite element spatial discretization was established. In [14], the authors enhanced the leastsquares approach by adding a boundary functional to the least-squares functional, thus, weakly imposing the boundary conditions.

In [5], anisotropic scattering in the scaled least-squares approach was first considered in the context of multigroup transport. A scaling operator for the least-squares

^{*}Received by the editors June 24, 2004; accepted for publication (in revised form) November 17, 2005; published electronically March 15, 2006.

http://www.siam.org/journals/sinum/44-2/61051.html

[†]Bioengineering Institute, University of Auckland, Private Bag 92019, Auckland, New Zealand (t.austin@auckland.ac.nz). This work is also referred to as LAUR-04-2523. Parts of this work were sponsored by the National Science Foundation under grant number DMS-8704169 and the Department of Energy, Applied Math Program grant DE-FG03-94ER25217.

[‡]Dept. of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@colorado.edu).

approach with anisotropic scattering in the multigroup context was introduced. Additionally, convergence results for a multilevel solution algorithm for the multigroup version of anisotropic scattering were presented. However, ellipticity of the scaled least-squares functional or estimates of the least-squares error were not considered.

For the single group equation, we set a foundation for the scaled least-squares approach introduced in [5] by proving ellipticity of the least-squares functional. We employ a scaling operator that is a single-group form of the scaling operator used in [5]. The proof of ellipticity for the thick regime with large absorption is identical to the equivalent case for isotropic scattering. The remaining cases depend on new techniques that were not used in [14]. Moreover, the proofs for the thin regime and the thick regime with small absorption have as special cases the proofs for the isotropic scattering [14]. It is the opinion of the authors that the proofs presented here are simpler and clearer than the proofs of [14], albeit resulting in small coercivity bounds.

The ellipticity results imply that the least-squares variational problem is well posed in an appropriate norm with ellipticity constants that are independent of the problem parameters. This ellipticity allows us to use Céa's lemma in establishing error bounds. Thus, once we introduce the discretization scheme (drawn from [5, 13, 14]), we use Céa's lemma to illustrate optimal bounds on the discretization error in the context of anisotropic scattering for the thin and thick regimes. For the thin regime, the proof from [14] can be invoked. For the thick regime with mildly anisotropic scattering, we merely indicate that the results are of the same form as [14]. There will, however, be new results for the thick regime with highly anisotropic scattering. These results will depend on an asymptotic expansion from Larsen and Pomraning in [10].

Most of the research on numerical methods for the linear Boltzmann equation with anisotropic scattering has focused on devising a plan to speed up source iteration, which is the standard iterative solution method used to solve isotropic transport problems [11]. Research has not focused on tailoring the discretization schemes used for isotropic scattering problems to anisotropic scattering problems because, in general, the same discretization techniques may be used [2, 15, 16]. Here, we focus on the formulation and discretization using a least-squares approach. We will not address the issue of what is the appropriate method for solving the resulting system of equations. For now we refer the reader to [5]. Since the approach for anisotropic scattering first described in [5] has not been studied theoretically, we concentrate on placing the method on firm ground. To this end, we proceed in the following way.

In section 2, we present the necessary preliminaries. Previous results for isotropic scattering are described in section 3. The scattering operator is presented in section 4 along with ellipticity results. In section 5, we describe the spatial and angular discretization scheme and present error estimates. In the final section, we discuss future work and further extensions of the least-squares method.

2. Preliminaries. As discussed in [11], the single-group, steady-state linear Boltzmann equation with anisotropic scattering is given by

(2.1)
$$\begin{bmatrix} \mathbf{\Omega} \cdot \nabla + \sigma_t \, \mathcal{I} - \sigma_s \, \mathcal{K} \end{bmatrix} \psi(\mathbf{x}, \mathbf{\Omega}) = q \quad \text{for } (\mathbf{x}, \mathbf{\Omega}) \in R \times S^2, \\ \psi(\mathbf{x}, \mathbf{\Omega}) = g \quad \text{for } \mathbf{x} \in \partial R \text{ with } \mathbf{n} \cdot \mathbf{\Omega} < 0$$

where σ_t is the total cross section, σ_s is the scattering cross section, and ψ is the angular flux to be determined for all points $\mathbf{x} \in R \subset \mathbb{R}^3$ and all possible travel directions $\mathbf{\Omega} = (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta) \in S^2$. Spatial domain R is assumed to be an open connected set with diam(R) = 1 and to have a piecewise $C^{1,1}$ boundary denoted by $\Gamma := \partial R$.

To define anisotropic scattering operator \mathcal{K} , we must recall the normalized spherical harmonics from [3], given by

(2.2)
$$Y_{\ell m}(\mathbf{\Omega}) := Y_{\ell m}(\theta, \varphi) = (-1)^m \sqrt{\frac{(2\ell+1)(\ell-m)!}{(\ell+m)!}} P_{\ell}^m(\cos \theta) e^{im\varphi},$$

where $P_{\ell}^{m}(\cdot)$ corresponds to the (ℓm) th associated Legendre moment. Normalization

$$\mathrm{d}\mathbf{\Omega} = \frac{\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\varphi}{4\pi}$$

allows us to expand the scattering operator \mathcal{K} as

(2.3)
$$(\mathcal{K}v)(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \sigma_{\ell} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}') v(\mathbf{x},\mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}',$$

where $\sigma_{\ell} \in [0, 1]$ for all l > 0 (with $\sigma_0 \equiv 1$) and $Y_{\ell m}^*$ is the complex conjugate of $Y_{\ell m}$. This infinite sum is truncated, in practice, such that for some $N_S \ge 0$,

(2.4)
$$(\mathcal{K}v)(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{N_S} \sigma_\ell \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}') v(\mathbf{x}, \mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}'.$$

Note that N_S depends on the degree of anisotropy in the scattering, and that when $N_S = 0$ in (2.4), we have

(2.5)
$$(\mathcal{K}v)(\mathbf{x},\mathbf{\Omega}) = \int_{S^2} v(\mathbf{x},\mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}',$$

resulting in the isotropic transport operator. For the remainder, we refer to the operator in (2.5) as \mathcal{P} . Note that for subset $\Xi \subset \mathbb{N} \equiv \{0, 1, 2, 3, ...\}$ we can define the more general operator

(2.6)
$$(\mathcal{P}_{\Xi} v)(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell \in \Xi} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}') v(\mathbf{x}, \mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}'.$$

Next, to distinguish between the isotropic and anisotropic transport operators, we introduce the notation \mathcal{L}_I and \mathcal{L}_A such that, for $v: R \times S^2 \to \Re$,

(2.7)
$$\mathcal{L}_{I} v := \mathbf{\Omega} \cdot \nabla v + \sigma_{t} \left(\mathcal{I} - \mathcal{P} \right) v + \sigma_{a} \mathcal{P} v$$

and

(2.8)
$$\mathcal{L}_A v := \mathbf{\Omega} \cdot \nabla v + \sigma_t \left(\mathcal{I} - \mathcal{K} \right) v + \sigma_a \mathcal{K} v,$$

where $\sigma_a := \sigma_t - \sigma_s$ represents the *absorption cross section*. Also, at times, the scattering term of the anisotropic transport operator will be represented by

(2.9)
$$\mathcal{S} = \sigma_t (\mathcal{I} - \mathcal{K}) + \sigma_a \mathcal{K},$$

or by

(2.10)
$$\mathcal{S}v(\mathbf{x},\mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \mu_{\ell} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\mathbf{\Omega}) \int_{S^2} Y_{\ell m}^*(\mathbf{\Omega}') v(\mathbf{x},\mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}',$$

where

(2.11)
$$\mu_{\ell} = \sigma_t (1 - \sigma_{\ell}) + \sigma_a \sigma_{\ell}$$

Next, let the standard L^2 inner product and norm be denoted by

$$\langle u,v
angle := \int_{S^2} \int_R \, u\,v^*\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{\Omega}' \qquad \mathrm{and}\qquad \|u\|:=\sqrt{\langle u,u
angle},$$

where v^* again is the complex conjugate of v. Denote by $L^2(S^2 \times R)$ the set of functions that are L²-integrable on $S^2 \times R$. Any function in $L^2(S^2 \times R)$ has a unique expression in terms of the spherical harmonics since the spherical harmonics are an orthonormal basis for $L^2(S^2)$. Specifically, every $v \in L^2(S^2 \times R)$ has the expansion

(2.12)
$$v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

with moments $\phi_{\ell m}(\mathbf{x})$ given by

(2.13)
$$\phi_{\ell m}(\mathbf{x}) = \int_{S^2} Y^*_{\ell m}(\mathbf{\Omega}') v(\mathbf{x}, \mathbf{\Omega}') \,\mathrm{d}\mathbf{\Omega}'.$$

3. Previous results for isotropic scattering. In [13], a scaling operator of the form $\mathcal{R} := a(\mathcal{I} - \mathcal{P}) + b\mathcal{P}$ was defined, with a and b depending on σ_t and σ_a . The isotropic form of (2.1) was then restated as the minimization of the least-squares functional

(3.1)
$$G_0(\psi;q) := \left\| \mathcal{R}^{-1/2} (\mathcal{L}_I \psi - q) \right\|^2.$$

The main result presented in [13] showed coercivity and continuity of the bilinear form $\langle \mathcal{R}^{-1}\mathcal{L}_I u, \mathcal{L}_I v \rangle$ with respect to

(3.2)
$$\|v\|_{V}^{2} := \left\langle \mathcal{R}^{-1} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \left\langle \mathcal{R}v, v \right\rangle$$

To be more precise, defining V as the space of functions bounded in the V-norm and V_0 as the subspace of V with homogeneous inflow boundary conditions, the authors established V-ellipticity, i.e., constants C_e and C_c , independent of σ_t and σ_a , such that

$$(3.3) C_e \|v\|_V^2 \le \left\langle \mathcal{R}^{-1} \mathcal{L}_I v, \mathcal{L}_I v \right\rangle \le C_c \|v\|_V^2$$

for any $v \in V_0$. In [14], they extended this work by adding a boundary functional to the V-norm and the least-squares functional, and again proved ellipticity.

To describe the work in [14], it is necessary to describe the boundary functional. For each $\mathbf{x} \in \Gamma$, define $\mathbf{n}(\mathbf{x})$ to be the outward unit normal, define

(3.4)
$$\Gamma_I(\mathbf{\Omega}) := \{ \mathbf{x} \in \Gamma : \mathbf{n} \cdot \mathbf{\Omega} < 0 \},\$$

and define $\Gamma_O(\mathbf{\Omega}) := \Gamma / \Gamma_I(\mathbf{\Omega})$ to be the set of inflow and outflow particle travel directions. By defining $D := R \times S^2$, we then denote the inflow and outflow boundary of D by

(3.5)
$$\partial D_I := \{ (\mathbf{x}, \mathbf{\Omega}) \in D : \mathbf{x} \in \Gamma_I(\mathbf{\Omega}) \}$$

and

(3.6)
$$\partial D_O := \{ (\mathbf{x}, \mathbf{\Omega}) \in D : \mathbf{x} \in \Gamma_O(\mathbf{\Omega}) \}.$$

Corresponding to the inflow and outflow boundary of ${\cal D}$ are

(3.7)
$$b_I(u,v) := \int_{\partial R} \int_{\mathbf{n} \cdot \mathbf{\Omega} < 0} uv |\mathbf{n} \cdot \mathbf{\Omega}| \, \mathrm{d}\mathbf{\Omega} \, \mathrm{d}\sigma$$

and

(3.8)
$$b_O(u,v) := \int_{\partial R} \int_{\mathbf{n} \cdot \mathbf{\Omega} > 0} uv |\mathbf{n} \cdot \mathbf{\Omega}| \, \mathrm{d}\mathbf{\Omega} \, \mathrm{d}\sigma.$$

Associated with $b_I(\cdot, \cdot)$ is the inflow norm

(3.9)
$$||v||_{B_I}^2 := b_I(v, v)$$

and the corresponding Sobolev space

(3.10)
$$B_I := \overline{\left\{ v \in C^{\infty}(\overline{\partial R_I}) : \|v\|_{B_I}^2 < \infty \right\}}.$$

For $q \in L^2$ and $g \in B_I$, the least-squares functional studied in [14] is given by

(3.11)
$$G_I(\psi; q, g) := G_0(\psi; q) + 2b_I(\psi - g, \psi - g).$$

The authors obtained ellipticity results for (3.11) with respect to

(3.12)
$$\|v\|_{V_1}^2 := \|v\|_V^2 + b_I(v,v)$$

and the space V_1 consisting of functions bounded in the V_1 -norm. Since G_I offers a more robust approximation of boundary conditions than G_0 , we work exclusively in this paper with a least-squares functional that is similar in form to (3.12).

4. New results for anisotropic scattering. The scaling operator for anisotropic scattering is given by

(4.1)
$$\mathcal{R} := \begin{cases} \mathcal{I} & \text{in Region I}, \\ \sigma_t(\mathcal{I} - \mathcal{K}) + \sigma_a \mathcal{K} & \text{in Region II}, \\ \sigma_t(\mathcal{I} - \mathcal{K}) + \frac{1}{\sigma_t} \mathcal{K} & \text{in Region III}, \end{cases}$$

where Regions I, II, and III are defined in Figure 4.1. Note that \mathcal{R} is a continuous function in σ_t and σ_a for fixed \mathcal{K} and can be alternatively expressed as

(4.2)
$$\mathcal{R}v := \sum_{\ell=0}^{\infty} \nu_{\ell} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

using $\phi_{\ell m}$ in (2.13) and

(4.3)
$$\nu_{\ell} := \begin{cases} 1 & \text{for Region I}, \\ \sigma_t (1 - \sigma_{\ell}) + \sigma_a \sigma_{\ell} & \text{for Region II}, \\ \sigma_t (1 - \sigma_{\ell}) + \frac{1}{\sigma_t} \sigma_{\ell} & \text{for Region III}. \end{cases}$$



FIG. 4.1. Division of parameters into Regions I, II, and III.

In the remainder of this section, we develop V_1 -ellipticity proofs for the case of anisotropic scattering using the scaling operator (4.1). Firstly, though, we note that the V_1 -norm, defined by (3.12), does not change for anisotropic scattering once \mathcal{R} is defined in terms of (4.1). Secondly, for $q \in L^2$ and $g \in B_I$, we note that the solution of (2.1) can be expressed as

(4.4)
$$\psi = \arg\min_{v \in V_1} G_A(v; q, g),$$

where $G_A(v; q, g)$ is the anisotropic equivalent of $G_I(v; q, g)$. The corresponding variational form is: find $\psi \in V_1$ such that

(4.5)
$$a(\psi, v) := \left\langle \mathcal{R}^{-1} \mathcal{L}_A \psi, \mathcal{L}_A v \right\rangle + 2b_I(\psi, v) = \left\langle \mathcal{R}^{-1} q, \mathcal{L}_A v \right\rangle + 2b_I(g, v)$$

for every $v \in V_1$. Once we establish V_1 -ellipticity results for $a(\psi, v)$, we will have established that (4.5) is well posed. This well posedness will imply that, for each pair $q \in L^2$ and $g \in B_I$, there exists a unique $\psi \in V_1$ satisfying (2.1). Moreover, a standard stability result (cf. [4]) implies that we get the a priori estimate:

(4.6)
$$\|\psi\|_{V_1} \le C_e^{-1} \left(\left\| \mathcal{R}^{-1/2} q \right\| + b_I(g,g)^{1/2} \right),$$

where C_e is the coercivity bound.

4.1. Auxiliary lemmas. In this section, we present two lemmas. Most of the first lemma is a restatement of Lemma 3.1 from [14]. The second lemma is used in the thin and thick regime ellipticity proofs.

First, we define an operator that arises in Lemma 4.1, and the ellipticity proof for the thick regime with small absorption. To define the operator, we define $s = (\sigma_t - \frac{1}{\sigma_t})/(\sigma_t - \sigma_a)$ and split the moments into the two disjoint sets,

(4.7)
$$\Upsilon = \{ \ell \in \mathbb{N} : \sigma_l > s \} = \{ \ell \in \mathbb{N} : \mu_\ell < 1/\sigma_t \},$$

(4.8)
$$\Upsilon = \{ \ell \in \mathbb{N} : \sigma_l \leq s \} = \{ \ell \in \mathbb{N} : \mu_\ell \geq 1/\sigma_t \}.$$

We can then define the projection operator

$$\mathcal{P}_{\Upsilon} v(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell \in \Upsilon} \sum_{m = -\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

according to (2.6). A similar operator can be defined for $\widehat{\Upsilon}$. Furthermore, we introduce the operator \mathcal{D} given by

(4.9)
$$\mathcal{D}v(\mathbf{x},\mathbf{\Omega}) := \mathcal{P}_{\Upsilon}v(\mathbf{x},\mathbf{\Omega}) + \sum_{\ell \in \widehat{\Upsilon}} \zeta_l \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

where $\zeta_{\ell} := (1 - \sigma_{\ell}) + \frac{\sigma_a \sigma_{\ell}}{\sigma_t}$. Notice that \mathcal{D}^s is a meaningful operator for any $s \in \Re$ and that $\|\mathcal{D}\| \leq 1$ since $\zeta_{\ell} \leq 1$. In the following, we use the notation, $\sum_{\ell} = \sum_{\ell}^{\infty}$. LEMMA 4.1. For $v \in V_1$, we have

- (i) $2 \langle \mathbf{\Omega} \cdot \nabla v, v \rangle = b_O(v, v) b_I(v, v) \ge -b_I(v, v);$
- (ii) the Poincaré-Friedrichs inequality

(4.10)
$$||v||^2 \le 2 \operatorname{diam}(R)^2 ||\mathbf{\Omega} \cdot \nabla v||^2 + 2 \operatorname{diam}(R) b_I(v, v);$$

(iii) for diam(R) = 1 and $\sigma_t \ge 1$, we have

(4.11)
$$\left\|\mathcal{P}_{\Upsilon} v\right\|^{2} \leq 2 \left\|\mathcal{Q}^{-1/2} \mathbf{\Omega} \cdot \nabla v + \sigma_{t} \mathcal{P}_{\Upsilon} \mathbf{\Omega} \cdot \nabla v\right\|^{2} + 2b_{I}(v, v)$$

where $\mathcal{Q} := \mathcal{D} \left(\mathcal{I} - \mathcal{P}_{\Upsilon} \right)$ and $\mathcal{Q}^s := \mathcal{D}^s \left(\mathcal{I} - \mathcal{P}_{\Upsilon} \right)$ for $s \in \Re$.

Proof. The proofs of (i) and (ii) are found in Lemma 3.1 of [14], while (iii) is proved by assuming (ii), and noting that $\|\mathcal{P}_{\Upsilon} v\|^2 \leq \|v\|^2$ and

$$\left\|\boldsymbol{\Omega}\cdot\nabla v\right\|^{2} \leq \left\|\mathcal{D}^{-1/2}\left(\mathcal{I}-\mathcal{P}_{\Upsilon}\right)\boldsymbol{\Omega}\cdot\nabla v+\sigma_{t}\,\mathcal{P}_{\Upsilon}\,\boldsymbol{\Omega}\cdot\nabla v\right\|^{2}.\qquad\Box$$

LEMMA 4.2. Given $\lambda_{\ell} > 0$ and $\omega_{\ell} > 0$, the minimum of

(4.12)
$$I(d) := \sum_{\ell} (\lambda_{\ell} - d)^2 \omega_{\ell}$$

for $d \in [0, 1]$ is achieved at

$$d_m = \frac{\sum_{\ell} \lambda_{\ell} \omega_{\ell}}{\sum_{\ell} \omega_{\ell}},$$

and furthermore,

$$I(d_m) = \sum_{\ell} \lambda_{\ell}^2 \omega_{\ell} - \frac{(\sum_{\ell} \lambda_{\ell} \omega_{\ell})^2}{\sum_{\ell} \omega_{\ell}}.$$

Proof. The result is established by differentiating I(d) with respect to d. The thin regime and thick regime with small absorption ellipticity proofs that follow make use of the projection operator

(4.13)
$$\mathcal{P}_{\ell} v := \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}),$$

implying that (2.10) can be expressed as

(4.14)
$$\mathcal{S}v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \mu_{\ell} \, \mathcal{P}_{\ell} \, v(\mathbf{x}, \mathbf{\Omega}).$$

Note the observation, $\sum_{\ell} \mathcal{P}_{\ell} = \mathcal{I}$, that we need in the following.

4.2. Thin regime $(0 \leq \sigma_a \leq \sigma_t \leq 1)$. For the thin regime, we have $\mathcal{R} = \mathcal{I}$ and

$$||u||_{V_1}^2 = ||\mathbf{\Omega} \cdot \nabla v||^2 + ||v||^2 + b_I(v, v).$$

THEOREM 4.3 (continuity and V₁-ellipticity for thin regime). Assume that $0 \leq$ $\sigma_a \leq \sigma_t \leq 1$. Then, for all $u, v \in V_1$, we have

(4.15)
$$|a(u,v)| = |\langle \mathcal{L}_A u, \mathcal{L}_A v \rangle + 2b_I(u,v)| \le C_c ||u||_{V_1} ||v||_{V_1},$$
$$a(v,v) = \langle \mathcal{L}_A v, \mathcal{L}_A v \rangle + 2b_I(v,v) \ge C_e ||v||_{V_1}^2,$$

with $C_c \leq 2$ and $C_e \geq 0.06574145$.

Proof. Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} a(u,v) &| \le \|\mathcal{L}_A \, u\| \, \|\mathcal{L}_A \, v\| + 2b_I(u,u)^{\frac{1}{2}} b_I(v,v)^{\frac{1}{2}} \\ &\le \left(\|\mathcal{L}_A \, u\|^2 + 2b_I(u,u)\right)^{\frac{1}{2}} \, \left(\|\mathcal{L}_A \, v\|^2 + 2b_I(v,v)\right)^{\frac{1}{2}}. \end{aligned}$$

Given that $\sigma_t(1 - \sigma_\ell) \leq 1$ and $\sigma_a \sigma_\ell \leq 1$, we have

$$\|\mathcal{L}_{A} u\|^{2} \leq 2\left(\|\mathbf{\Omega} \cdot \nabla u\|^{2} + \|(\mathcal{I} - \mathcal{P})u\|^{2} + \|\mathcal{P}u\|^{2}\right) = 2\left(\|\mathbf{\Omega} \cdot \nabla u\|^{2} + \|u\|^{2}\right)$$

so that $\|\mathcal{L}_A u\|^2 + 2b_I(u, u) \leq 2 \|u\|_{V_1}^2$. This proves continuity of $a(\cdot, \cdot)$. To prove ellipticity, we note that $0 \leq \mu_\ell \leq 1$ and refer the reader to (4.14). Using this definition of \mathcal{S} , we have

$$a(v,v) = \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v \rangle + 2 \langle \Omega \cdot \nabla v, \mathcal{S}v \rangle + \langle \mathcal{S}v, \mathcal{S}v \rangle + 2 b_I(v,v)$$

$$(4.16) = \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v \rangle + 2 \sum_{\ell} \mu_{\ell} \langle \Omega \cdot \nabla v, \mathcal{P}_{\ell}v \rangle + \sum_{\ell} \mu_{\ell}^2 \langle \mathcal{P}_{\ell}v, v \rangle + 2 b_I(v,v)$$

For any $d \in [0, 1]$, adding the identity

$$2d\left[\langle \Omega \cdot \nabla v, v \rangle - \sum_{\ell} \langle \Omega \cdot \nabla v, \mathcal{P}_{\ell} v \rangle\right] = 0$$

to (4.16) and using Lemma 4.1(i) yields

(4.17)
$$a(v,v) \ge \langle \Omega \cdot \nabla v, \Omega \cdot \nabla v \rangle + 2 \sum_{\ell} (\mu_{\ell} - d) \langle \Omega \cdot \nabla v, \mathcal{P}_{\ell} v \rangle + \sum_{\ell} \mu_{\ell}^2 \langle \mathcal{P}_{\ell} v, v \rangle + (2 - d) b_I(v,v)$$

For convenience, we put $A_{\ell} = \|\mathcal{P}_{\ell} \mathbf{\Omega} \cdot \nabla v\|^2$, $A = \|\mathbf{\Omega} \cdot \nabla v\|^2$, $B_{\ell} = \|\mathcal{P}_{\ell} v\|^2$, and $B = \|v\|^2$. We also define $\gamma_{\ell} = A_{\ell}/A$ and $\delta_{\ell} = B_{\ell}/B$ and note that

$$\sum_{\ell} \delta_{\ell} = \sum_{\ell} \gamma_{\ell} = 1.$$

Thus we seek a proof of

$$a(v,v) \ge C \left(A + B + b_I(v,v)\right)$$

for all $v \in V_1$. Applying the arithmetic-geometric inequality to the cross product term in (4.17) with $\eta > 0$ yields

$$a(v,v) \ge \left(A - \eta \sum_{\ell} A_{\ell}\right) + \left(\sum_{\ell} \mu_{\ell}^2 B_{\ell} - \frac{1}{\eta} \sum_{\ell} (\mu_{\ell} - d)^2 B_{\ell}\right) + (2 - d) b_I(v,v)$$

= $(1 - \eta) A + \left(\sum_{\ell} \mu_{\ell}^2 \delta_{\ell} - \frac{1}{\eta} \sum_{\ell} (\mu_{\ell} - d)^2 \delta_{\ell}\right) B + (2 - d) b_I(v,v)$

since $B_{\ell} = \delta_{\ell} B$. Making use of Lemma 4.2 we choose $d = \sum_{\ell} \mu_{\ell} \delta_{\ell} \leq \sum_{\ell} \delta_{\ell} = 1$, implying that

$$a(v,v) \ge [1-\eta] A + \left[\sum_{\ell} \mu_{\ell}^2 \delta_{\ell} - \frac{1}{\eta} \left(\sum_{\ell} \mu_{\ell}^2 \delta_{\ell} - \left(\sum_{\ell} \mu_{\ell} \delta_{\ell}\right)^2\right)\right] B + b_I(v,v)$$
$$= [1-\eta] A + \left[\left(1-\frac{1}{\eta}\right) \sum_{\ell} \mu_{\ell}^2 \delta_{\ell} + \frac{1}{\eta} \left(\sum_{\ell} \mu_{\ell} \delta_{\ell}\right)^2\right] B + b_I(v,v).$$

For convenience, define $\delta = \sum_{\ell} \mu_{\ell}^2 \delta_{\ell} \leq \sum_{\ell} \mu_{\ell} \delta_{\ell} \leq 1$ such that

$$a(v,v) \ge [1-\eta]A + \left[\left(1-\frac{1}{\eta}\right)\delta + \frac{1}{\eta}\delta^2\right]B + b_I(v,v)$$
$$= [1-\eta]A + \left[\delta - \frac{1}{\eta}\delta(1-\delta)\right]B + b_I(v,v).$$

Using Lemma 4.1(ii) (assuming diam(R) = 1) we get

$$a(v,v) \ge \left[1 - \eta - \beta\right] A + \left[\delta - \frac{1}{\eta}\delta(1 - \delta) + \frac{\beta}{2}\right] B + \left[1 - \beta\right] b_I(v,v)$$

for any $\beta \geq 0$. Define

$$C_1 = 1 - \eta - \beta,$$

$$C_2 = \delta - \frac{1}{\eta}\delta(1 - \delta) + \frac{\beta}{2},$$

$$C_3 = 1 - \beta.$$

Next, set $\eta = \sqrt{\delta(1-\delta)}$ and choose β to make $C_1 = C_2$. This requires

$$1 - \sqrt{\delta(1-\delta)} - \beta = \delta - \sqrt{\delta(1-\delta)} + \frac{\beta}{2},$$

which implies

$$\beta = \frac{2}{3}(1-\delta) \ge 0.$$

Plugging back into C_1 yields

$$C_1 = C_2 = 1 - \sqrt{\delta(1-\delta)} - \frac{2}{3}(1-\delta),$$

$$C_3 = 1 - \frac{2}{3}(1-\delta) \ge \frac{1}{3}.$$

Numerically we find that the minimum value of C_1 occurs at $\delta = \frac{1-\sqrt{16/52}}{2} \approx 0.2226499$ and yields $C_1 = C_2 \approx 0.06574145$. Comparing to the bound on C_3 , we see that $C_e \geq 0.06574145$. \Box

4.3. Thick regime with "large" absorption $(1 \leq \sigma_t < \infty; \frac{1}{\sigma_t} \leq \sigma_a \leq \sigma_t)$. For the thick regime with "large" absorption, the scaling is given by

(4.18)
$$\mathcal{R} = \sigma_t \left(\mathcal{I} - \mathcal{K} \right) + \sigma_a \mathcal{K},$$

which implies that

$$\|v\|_{V}^{2} := \left\langle \mathcal{R}^{-1} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \left\langle \mathcal{R} v, v \right\rangle$$

and

$$||v||_{V_1}^2 = ||v||_V^2 + b_I(v,v).$$

THEOREM 4.4 (continuity and V₁-ellipticity for the thick regime with "large" absorption). Assume that $1 \leq \sigma_t < \infty$ and $\frac{1}{\sigma_t} \leq \sigma_a \leq \sigma_t$. Then, for all $u, v \in V_1$, we have

(4.19)
$$|a(u,v)| \le 2 ||u||_{V_1} ||v||_{V_1},$$
$$a(v,v) \ge ||v||_{V_1}^2.$$

Proof. See the proof establishing coercivity and continuity in thick regime with "large" absorption from [14]. Proof for coercivity is the same because, as in [14], scaling operator \mathcal{R} , in (4.18) is equal to scattering operator \mathcal{S} . \Box

4.4. Thick regime with "small" absorption $(1 \leq \sigma_t \leq \infty; \sigma_a \leq \frac{1}{\sigma_t})$. For the thick regime with "small" absorption, we must define the scaling operator so that it does not become singular as $\sigma_a \to 0$. Hence, the scaling operator defined in (4.18) is recast as

(4.20)
$$\mathcal{R} = \sigma_t (\mathcal{I} - \mathcal{K}) + \frac{1}{\sigma_t} \mathcal{K},$$

while the V-norm and the V_1 -norm have the same dependence on \mathcal{R} .

One of the key ingredients of the following proof is the intermediate scaling operator defined below. Recall Υ and $\widehat{\Upsilon}$ defined by (4.7) and (4.8), and Q defined by (4.11). The intermediate scaling operator is

$$\mathcal{T} := \sigma_t \mathcal{Q} + \frac{1}{\sigma_t} \mathcal{P}_{\Upsilon} = \mathcal{S}(\mathcal{I} - \mathcal{P}_{\Upsilon}) + \frac{1}{\sigma_t} \mathcal{P}_{\Upsilon}.$$

Note that $\mathcal{Q} \leq \mathcal{I}$, and additionally, that $\sigma_a \mathcal{P}_{\Upsilon} \leq \mathcal{SP}_{\Upsilon} \leq \frac{1}{\sigma_t} \mathcal{P}_{\Upsilon}$, which is equivalent to $\sigma_a \leq \mu_{\ell} \leq \frac{1}{\sigma_t}$ for $\ell \in \Upsilon$. The first inequality is true by definition of μ_{ℓ} , and the second is true by definition of Υ . We also introduce τ_{ℓ} according to

(4.21)
$$\tau_{\ell} := \begin{cases} \frac{1}{\sigma_t} & \text{for } \ell \in \Upsilon\\ \mu_{\ell} & \text{for } \ell \in \widehat{\Upsilon} \end{cases}$$

such that, from (4.13),

$$\mathcal{T}v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \tau_{\ell} \mathcal{P}_{\ell} v(\mathbf{x}, \mathbf{\Omega}).$$



FIG. 4.2. Graph of τ_{ℓ} and ν_{ℓ} , where the black region on $\sigma_{\ell} \in [0, 1]$ denotes $\widehat{\Upsilon}$ and the grey region denotes Υ . From the graph we can see that the largest ratio occurs at the interface between $\widehat{\Upsilon}$ and Υ , which is where $\sigma_{\ell} = s$.

For future reference, we note that

(4.22)
$$\mathcal{T}^{-1}\mathcal{S} = \mathcal{S}\mathcal{T}^{-1} = (\mathcal{I} - \mathcal{P}_{\Upsilon}) + \sigma_t \mathcal{S}\mathcal{P}_{\Upsilon},$$

(4.23)
$$\mathcal{T}^{-1}\mathcal{S}^2 = \mathcal{S}^2\mathcal{T}^{-1} = \sigma_t\mathcal{Q} + \sigma_t\mathcal{S}^2\mathcal{P}_{\Upsilon},$$

and \mathcal{S} can be expressed as

$$\mathcal{S} = \mathcal{S}(\mathcal{I} - \mathcal{P}_{\Upsilon}) + \mathcal{S}\mathcal{P}_{\Upsilon} = \sigma_t \mathcal{Q} + \mathcal{S}\mathcal{P}_{\Upsilon}.$$

Note the inequality

(4.24)
$$\langle Sv, v \rangle \le \langle Tv, v \rangle \le \langle \mathcal{R}v, v \rangle$$

and the related inequality

(4.25)
$$\langle \mathcal{T}^{-1}v, v \rangle \ge \langle \mathcal{R}^{-1}v, v \rangle.$$

We also need the following lemma, which further relates \mathcal{T} to \mathcal{R} .

LEMMA 4.5. For $\sigma_t \geq 1$ and $\sigma_a \leq 1/\sigma_t$, we have

$$2 \langle \mathcal{T}v, v \rangle \ge \langle \mathcal{R}v, v \rangle \qquad and \qquad \left\langle \mathcal{T}^{-1}v, v \right\rangle \le 2 \left\langle \mathcal{R}^{-1}v, v \right\rangle$$

Proof. To prove both results, we establish a bound relating τ_{ℓ} to ν_{ℓ} , where ν_{ℓ} is defined by (4.3). By observing the graph of τ_{ℓ} and ν_{ℓ} in Figure 4.2, we see that the ratio ν_{ℓ}/τ_{ℓ} is maximized at $s = (\sigma_t - \frac{1}{\sigma_t})/(\sigma_t - \sigma_a)$. Evaluating both ν_{ℓ} and τ_{ℓ} at s yields

$$\frac{\nu_\ell}{\tau_\ell} = \frac{\sigma_t(1-s) + s/\sigma_t}{1/\sigma_t} = \sigma_t^2(1-s) + s = \frac{2\sigma_t - \sigma_a\sigma_t^2 - 1/\sigma_t}{\sigma_t - \sigma_a}.$$

As $\sigma_a \to 0$, the value of s decreases to its minimum (dependent on σ_a) of $1 - 1/\sigma_t^2$. Since $\nu_\ell/\tau_\ell = \sigma_t^2(1-s) + s$, the maximum of the ratio is where s is at its minimum, i.e., where $\sigma_a = 0$. This implies

$$\frac{\nu_\ell}{\tau_\ell} \le 2 - 1/\sigma_t^2 \le 2.$$

Both results follow from this inequality. \Box

THEOREM 4.6 (continuity and V_1 -ellipticity for thick regime with "small" absorption). Assume that $1 \leq \sigma_t < \infty$, $0 \leq \sigma_a \leq \frac{1}{\sigma_t}$. Then, for all $u, v \in V_1$, we have

$$(4.26) |a(u,v)| = |\langle \mathcal{R}^{-1}\mathcal{L}_A u, \mathcal{L}_A v \rangle + 2b_I(u,v)| \le C_c ||u||_{V_1} ||v||_{V_1},$$

(4.27)
$$a(v,v) = \left\langle \mathcal{R}^{-1} \mathcal{L}_A v, \mathcal{L}_A v \right\rangle + 2b_I(v,v) \ge C_e \|v\|_{V_1}^2,$$

with $C_c \leq 2$ and $C_e \geq 0.01667$, independent of σ_t and σ_a .

Proof. The proof for continuity follows from the same reasoning as used in Theorem 4.4 since

$$|a(u,v)| \leq \left(\left\| \mathcal{R}^{-\frac{1}{2}} \mathcal{L}_A u \right\| + 2b_I(u,u) \right)^{\frac{1}{2}} \left(\left\| \mathcal{R}^{-\frac{1}{2}} \mathcal{L}_A v \right\| + 2b_I(v,v) \right)^{\frac{1}{2}}.$$

The observation that $\mu_{\ell} \leq \nu_{\ell}$ implies

$$\left\|\mathcal{R}^{-\frac{1}{2}}\mathcal{S}u\right\| \leq \left\|\mathcal{R}^{\frac{1}{2}}u\right\|.$$

One can easily show then that $\left\| \mathcal{R}^{-\frac{1}{2}} \mathcal{L}_A u \right\|^2 \leq 2 \left\| u \right\|_V^2$. To establish coercivity, we proceed as follows. We first define

(4.28)
$$\widetilde{a}(v,v) := \left\langle \mathcal{T}^{-1}\mathcal{L}_A v, \mathcal{L}_A v \right\rangle + 2 b_I(v,v),$$

and through Lemma 4.5, we get

(4.29)
$$a(v,v) \ge \frac{1}{2} \widetilde{a}(v,v).$$

Assume for now that we have

(4.30)
$$\widetilde{a}(v,v) \ge \widetilde{C}_e \left(\left\| \mathcal{T}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2 + \left\| \mathcal{T}^{1/2} v \right\|^2 + b_I(v,v) \right).$$

Using inequality (4.25) and Lemma 4.5, we can bound the first two terms on the right-hand side of (4.30) from below to get

(4.31)
$$\widetilde{a}(v,v) \ge \frac{\widetilde{C}_e}{2} \left(\left\| \mathcal{R}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2 + \left\| \mathcal{R}^{1/2} v \right\|^2 + b_I(v,v) \right).$$

With (4.29) we get as an ellipticity constant in (4.27) the value of $\tilde{C}_e/4$. Thus, we are only left to prove (4.30), and determine C_e .

Noting (4.22) and (4.23) we write

$$\begin{split} \widetilde{a}(v,v) &= \left\langle \mathcal{T}^{-1} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \left\langle \mathcal{S}^2 \mathcal{T}^{-1} v, v \right\rangle + 2 \left\langle \mathcal{S} \mathcal{T}^{-1} \mathbf{\Omega} \cdot \nabla v, v \right\rangle + 2 b_I(v,v) \\ &= \frac{1}{\sigma_t} \left\langle \mathcal{Q}^{-1} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \sigma_t \left\langle \mathcal{P}_{\Upsilon} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \sigma_t \left\langle \mathcal{Q} v, v \right\rangle \\ &+ \sigma_t \sum_{\ell \in \Upsilon} \mu_\ell^2 \left\langle \mathcal{P}_\ell v, v \right\rangle + 2 \left\langle \mathbf{\Omega} \cdot \nabla v, (\mathcal{I} - \mathcal{P}_{\Upsilon}) v \right\rangle + 2 \sigma_t \sum_{\ell \in \Upsilon} \mu_\ell \left\langle \mathbf{\Omega} \cdot \nabla v, \mathcal{P}_\ell v \right\rangle \\ &+ 2 b_I(v,v). \end{split}$$

For any $d \in [0, 1]$, adding the identity

$$2d \left[\langle \mathbf{\Omega} \cdot \nabla v, v \rangle - \langle \mathbf{\Omega} \cdot \nabla v, (\mathcal{I} - \mathcal{P}_{\Upsilon})v \rangle - \langle \mathbf{\Omega} \cdot \nabla v, \mathcal{P}_{\Upsilon}v \rangle \right] = 0$$

to the last line above, and using the inequality from Lemma 4.1(ii), yields

$$\begin{split} \widetilde{a}(v,v) &\geq \frac{1}{\sigma_t} \left\langle \mathcal{Q}^{-1} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \sigma_t \left\langle \mathcal{P}_{\mathbf{\Upsilon}} \mathbf{\Omega} \cdot \nabla v, \mathbf{\Omega} \cdot \nabla v \right\rangle + \sigma_t \left\langle \mathcal{Q}v, v \right\rangle \\ &+ \sigma_t \sum_{\ell \in \mathbf{\Upsilon}} \mu_\ell^2 \left\langle \mathcal{P}_\ell v, v \right\rangle - 2|1 - d| \left| \left\langle \mathbf{\Omega} \cdot \nabla v, (\mathcal{I} - \mathcal{P}_{\mathbf{\Upsilon}})v \right\rangle \right| \\ &- 2 \sum_{\ell \in \mathbf{\Upsilon}} |\sigma_t \mu_\ell - d| \left| \left\langle \mathbf{\Omega} \cdot \nabla v, \mathcal{P}_\ell v \right\rangle \right| + (2 - d) b_I(v, v). \end{split}$$

It is convenient to note that we may write

$$|\langle \mathbf{\Omega} \cdot \nabla v, (\mathcal{I} - \mathcal{P}_{\Upsilon})v \rangle| \leq \left\| \frac{1}{\sqrt{\sigma_t}} Q^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\| \left\| \sqrt{\sigma_t} Q^{1/2}v \right\|,$$

and for $\ell \in \Upsilon$,

$$|\langle \mathbf{\Omega} \cdot \nabla v, \mathcal{P}_{\ell} v \rangle| \leq ||\sqrt{\sigma_t} \mathcal{P}_{\ell} \mathbf{\Omega} \cdot \nabla v|| \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{P}_{\ell} v \right\|.$$

Let's now define

$$\delta_{\ell} := \frac{\left\|\frac{1}{\sqrt{\sigma_{t}}}\mathcal{P}_{\ell}v\right\|^{2}}{\left\|\sqrt{\sigma_{t}}\mathcal{Q}^{1/2}v + \frac{1}{\sqrt{\sigma_{t}}}\mathcal{P}_{\Upsilon}v\right\|^{2}} \quad \text{and} \quad \gamma_{\ell} := \frac{\left\|\sqrt{\sigma_{t}}\mathcal{P}_{\ell}\mathbf{\Omega}\cdot\nabla v\right\|^{2}}{\left\|\frac{1}{\sqrt{\sigma_{t}}}\mathcal{Q}^{-1/2}\mathbf{\Omega}\cdot\nabla v + \sqrt{\sigma_{t}}\mathcal{P}_{\Upsilon}\mathbf{\Omega}\cdot\nabla v\right\|^{2}}$$

for $\ell \in \Upsilon$. Additionally, let $\delta_0 = 1 - \sum_{\ell \in \Upsilon} \delta_\ell$ and $\gamma_0 = 1 - \sum_{\ell \in \Upsilon} \gamma_\ell$. For convenience, set

$$A = \left\| \mathcal{T}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2 = \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{Q}^{-1/2} \mathbf{\Omega} \cdot \nabla v + \sqrt{\sigma_t} \mathcal{P}_{\Upsilon} \mathbf{\Omega} \cdot \nabla v \right\|^2$$

and

$$B = \left\| \mathcal{T}^{1/2} v \right\|^2 = \left\| \sqrt{\sigma_t} \mathcal{Q}^{1/2} v + \frac{1}{\sqrt{\sigma_t}} \mathcal{P}_{\Upsilon} v \right\|^2.$$

We can now use the arithmetic-geometric inequality to write, for any $\eta > 0$,

$$\widetilde{a}(v,v) \geq \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{Q}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2 + \sum_{\ell \in \Upsilon} \left\| \sqrt{\sigma_t} \mathcal{P}_{\ell} \mathbf{\Omega} \cdot \nabla v \right\|^2 + \left\| \sqrt{\sigma_t} \mathcal{Q}^{1/2} v \right\|^2 \\ + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell)^2 \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{P}_{\ell} v \right\|^2 - \eta \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{Q}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2 \\ - \frac{(1-d)^2}{\eta} \left\| \sqrt{\sigma_t} \mathcal{Q}^{1/2} v \right\|^2 - \sum_{\ell \in \Upsilon} \eta \left\| \sqrt{\sigma_t} \mathcal{P}_{\ell} \mathbf{\Omega} \cdot \nabla v \right\|^2 \\ - \sum_{\ell \in \Upsilon} \frac{(\sigma_t \mu_\ell - d)^2}{\eta} \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{P}_{\ell} v \right\|^2 + (2-d) b_I(v,v) \\ (4.32) \qquad = \left[(1-\eta)\gamma_0 + \sum_{\ell \in \Upsilon} (1-\eta)\gamma_\ell \right] A + \left[\left(1 - \frac{(1-d)^2}{\eta} \right) \delta_0 \right]$$

(4.33)
$$+\sum_{\ell\in\Upsilon} \left((\sigma_t\mu_\ell)^2 - \frac{(\sigma_t\mu_\ell - d)^2}{\eta} \right) \delta_\ell \right] B + (2-d)b_I(v,v).$$

We choose d to make the coefficient of B as large as possible. First we simplify this expression by temporarily setting $\mu_0 = 1/\sigma_t$ so that the coefficient on B can be expressed as

(4.34)
$$\sum_{\ell \in 0 \cup \Upsilon} (\sigma_t \mu_\ell)^2 \delta_\ell - \frac{1}{\eta} \sum_{\ell \in 0 \cup \Upsilon} (\sigma_t \mu_\ell - d)^2 \delta_\ell.$$

Then we see that d should be chosen to minimize the second sum, which by Lemma 4.2, yields

(4.35)
$$d = \sum_{\ell \in 0 \cup \Upsilon} (\sigma_t \mu_\ell) \, \delta_\ell,$$

making use of the fact that $\sum_{\ell \in 0 \cup \Upsilon} \delta_{\ell} = 1$. Note that $d \in [0, 1]$ because $(\sigma_t \mu_{\ell}) \in [0, 1]$. Substituting d defined by (4.35) into (4.34) yields

$$(4.36) \quad \delta_0 + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell)^2 \delta_\ell - \frac{1}{\eta} \left(\left(\delta_0 + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell)^2 \delta_\ell \right) - \left(\delta_0 + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell) \delta_\ell \right)^2 \right).$$

For convenience, we rewrite (4.36) as

$$\Delta_0 - \frac{1}{\eta} \left(\Delta_0 - \Delta_1^2 \right) = \Delta_0 \left(1 - \frac{1}{\eta} \right) + \frac{\Delta_1^2}{\eta},$$

where

$$\Delta_0 = \delta_0 + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell)^2 \delta_\ell \quad \text{and} \quad \Delta_1 = \delta_0 + \sum_{\ell \in \Upsilon} (\sigma_t \mu_\ell) \delta_\ell.$$

Note that Lemma 4.2 implies that $\Delta_0 \geq \Delta_1^2$. Also, $(\sigma_t \mu_\ell)^2 \leq (\sigma_t \mu_\ell) \leq 1$ for $\ell \in \Upsilon$ implies that $\Delta_0 \leq \Delta_1 \leq 1$. Hence, we get

(4.37)
$$\Delta_0 - \frac{1}{\eta} \left(\Delta_0 - \Delta_1^2 \right) = \Delta_0 \left(1 - \frac{1}{\eta} \right) + \frac{\Delta_1^2}{\eta} \ge \Delta_0 \left(1 - \frac{1}{\eta} \right) + \frac{\Delta_0^2}{\eta}.$$

Lastly, notice that $\Delta_0 \geq \delta_0$.

Next, using (4.37) and the fact that $\gamma_0 = 1 - \sum_{\ell \in \Upsilon} \gamma_\ell$, we obtain from (4.32)–(4.33)

$$\widetilde{a}(v,v) \ge [1-\eta]A + \left[\Delta_0\left(1-\frac{1}{\eta}\right) + \frac{\Delta_0^2}{\eta}\right]B + b_I(v,v)$$
$$\ge [1-\eta]A + \left[\Delta_0 - \frac{\Delta_0(1-\Delta_0)}{\eta}\right]B + b_I(v,v).$$

If we set $\delta = (1 - \Delta_0) \le (1 - \delta_0)$, then using Lemma 4.1(iii), we can say

$$\frac{\beta\delta}{2}B \leq \frac{\beta(1-\delta_0)}{2}B = \frac{\beta}{2} \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{P}_{\Upsilon} v \right\|^2$$
$$\leq \beta \left\| \frac{1}{\sqrt{\sigma_t}} \mathcal{Q}^{-1/2} \mathbf{\Omega} \cdot \nabla v + \sqrt{\sigma_t} \mathcal{P}_{\Upsilon} \mathbf{\Omega} \cdot \nabla v \right\|^2 + \frac{\beta}{\sigma_t} b_I(v,v) = \beta A + \frac{\beta}{\sigma_t} b_I(v,v).$$

This implies

$$(4.38) \quad \widetilde{a}(v,v) \ge [1-\eta-\beta]A + \left[(1-\delta) - \frac{\delta(1-\delta)}{\eta} + \frac{\beta\delta}{2}\right]B + \left[1-\frac{\beta}{\sigma_t}\right]b_I(v,v),$$

where $\beta \ge 0$ and $\eta \ge 0$ are to be determined.

Our choice of β is given by

(4.39)
$$\beta = \frac{2}{2+\delta} \left(\delta + \frac{\delta(1-\delta)}{\eta} - \eta \right),$$

which is found by setting $C_1 = C_2$, that is, setting

$$1 - \eta - \beta = (1 - \delta) - \frac{\delta(1 - \delta)}{\eta} + \frac{\beta\delta}{2}.$$

Plugging (4.39) into the coefficient on A generates the lower bound

(4.40)
$$C = 1 - \eta - \beta = \left(\frac{(2-\delta)}{(2+\delta)} - \frac{\delta}{2+\delta}\left(\eta + \frac{2(1-\delta)}{\eta}\right)\right).$$

This is maximized for $\eta = \sqrt{2(1-\delta)}$, which yields

(4.41)
$$C = \frac{(2-\delta) - 2\delta\sqrt{2(1-\delta)}}{2+\delta}.$$

Note that this is only valid when the corresponding $\beta \geq 0$, that is, (4.41) is valid only for $\delta \geq \delta_c$, where δ_c is the root of

$$\beta = \frac{2}{2+\delta} \left(\delta + \frac{\delta(1-\delta)}{\sqrt{2(1-\delta)}} - \sqrt{2(1-\delta)} \right) = \frac{2\delta - (2-\delta)\sqrt{2(1-\delta)}}{2+\delta} = 0$$

This root is the only real root of the polynomial

$$\delta^3 - 3\delta^2 + 8\delta - 4 = 0.$$

Numerically we find that $\delta_c \approx 0.6117$, implying $C_c \approx 0.1188$. We also find numerically the minimum of C in (4.41) on $[\delta_c, 1]$ to be

(4.42)
$$C_m = \min_{\delta \in [\delta_c, 1]} C \approx 0.06667,$$

which occurs at $\delta_m \approx 0.7836$.

Now, for $\delta \leq \delta_c$, we set $\beta = 0$ and choose η such that

$$1 - \eta = (1 - \delta) - \frac{\delta(1 - \delta)}{\eta},$$

resulting in

$$\eta = \frac{\delta + \sqrt{4\delta - 3\delta^2}}{2},$$

and, subsequently,

$$C = \frac{2 - \delta - \sqrt{4\delta - 3\delta^2}}{2}$$

Clearly, C is a decreasing function of $\delta \in [0, \delta_c]$ implying it takes on its smallest value at δ_c . As mentioned previously $C \approx 0.1188$ for this value of δ .

We finally complete the proof with a bound on the coefficient of $b_I(v, v)$ in (4.38). The minimization of this term is only considered on $[\delta_c, 1]$ since $\beta = 0$ on $[0, \delta_c]$. Plotting β on $[\delta_c, 1]$ shows that β is an increasing function of δ on this interval implying $\beta \leq 2/3$. Thus, we have that $[1 - \beta/\sigma_t] \geq 1/3$. We finally conclude that

$$\widetilde{a}(v,v) \ge C_e(A+B+b_I(v,v)),$$

where $\widetilde{C}_e = C_m \ge 0.06666$ coming from (4.42). Since $A \ge \left\| \mathcal{R}^{-1/2} \mathbf{\Omega} \cdot \nabla v \right\|^2$ and $2B \ge \left\| \mathcal{R}^{1/2} v \right\|^2$, we get

$$\widetilde{a}(v,v) \ge 0.03333 \|v\|_{V_1}^2$$

implying that $C_e = 0.03333/2 = 0.01667$ from (4.31).

5. Discretization and error bounds. Any finite dimensional subspace of V_1 may be used to construct an approximation to the solution of ψ . One approach, which is a subject of future research, is using a tessellation of the sphere to represent angular dependence and nonconforming finite elements to describe spatial variability. However, in this paper, we develop error bounds associated with a P_N approximation in angle and standard H^1 conforming finite elements in space. The angular approximation is represented by a truncated expansion of (2.12), which must be of greater order than the finite sum that represents the scattering kernel (i.e., $N \geq N_S$). A finite element approximation of the moments is defined on a triangulation \mathcal{T}_h of R into hexahedrals or tetrahedrons.

Let $\mathbb{P}_k(\mathcal{T}_h)$ denote the space of piecewise polynomials of degree $\leq k$ on \mathcal{T}_h , let Π_h be the corresponding interpolation operator on $\mathbb{P}_k(\mathcal{T}_h)$, and let the truncation operator Π_N be defined by

(5.1)
$$\Pi_N v(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell=0}^N \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}).$$

Then the discrete space V^h is defined by

(5.2)
$$V^{h} := \left\{ v_{h} \in V : v_{h} = \sum_{\ell=0}^{N} \sum_{m=-\ell}^{\ell} \phi_{\ell m}^{h}(\mathbf{x}), Y_{\ell m}(\mathbf{\Omega}); \phi_{\ell m}^{h}(\mathbf{x}) \in \mathbb{P}_{k}(\mathcal{T}_{h}) \right\}.$$

The definition of V^h yields the discrete problem: find $\psi^h \in V^h$ such that

(5.3)
$$a(\psi^h, v^h) = \left\langle \mathcal{R}^{-1} q, \mathcal{L}_A v^h \right\rangle + 2b_I(g, v^h)$$

for all $v^h \in V^h$.

Bounds for the discretization error are obtained by following the procedure outlined in [14]. Thus, let the components of $\Omega \in S^2$ and $\mathbf{x} \in R$ be denoted by $\Omega = (\Omega_1, \Omega_2, \Omega_3), \mathbf{x} = (x_1, x_2, x_3)$, respectively, and let β, γ be a multi-index such that $D_{\mathbf{x}}^{\beta} := \frac{\partial^{|\beta|}}{\partial_{x_1}^{\beta_1} \partial_{x_2}^{\beta_2} \partial_{x_3}^{\beta_3}}$, and $D_{\Omega}^{\gamma} := \frac{\partial^{|\gamma|}}{\partial_{\Omega_1}^{\gamma_1} \partial_{\Omega_2}^{\gamma_2} \partial_{\Omega_3}^{\gamma_3}}$. Recall that the standard norms [1] of $H^k(R) \times H^l(S^2)$ and $H^k(\partial R) \times H^l(S^2)$ are given by

$$\|v\|_{k,l}^{2} := \sum_{|\beta| \le k} \sum_{|\gamma| \le l} \left\| D_{\mathbf{\Omega}}^{\gamma} D_{\mathbf{x}}^{\beta} v \right\|,$$
$$\|v\|_{k,l,\partial R}^{2} := \sum_{|\beta| \le k} \sum_{|\gamma| \le l} \int_{\partial R} \int_{S^{2}} |D_{\mathbf{x}}^{\beta} v|^{2} \,\mathrm{d}\mathbf{\Omega} \,\mathrm{d}\sigma.$$

Note also the following bounds for the interpolation error (see [4]):

(5.4)
$$\begin{aligned} \|v - \Pi_h v\|_{p,0} &\leq Ch^{k+1-p} \|v\|_{k+1,0} \ \forall v \in H^{k+1}(R) \times H^l(S^2), \\ \|v - \Pi_h v\|_{p,0,\partial R} &\leq Ch^{k+1-p} \|v\|_{k+1,0,\partial R} \ \forall v \in H^{k+1}(\partial R) \times H^l(S^2). \end{aligned}$$

for $p \in 0, 1$. We also define

(5.5)
$$\mathcal{E}_h(v) := v - \Pi_h v$$

and

(5.6)
$$\mathcal{E}_N(v) := v - \Pi_N v$$

for all $v \in L^2(S^2 \times R)$.

To bound the error of the truncated expansion (5.1), we recall that the spherical harmonics are eigenfunctions of the Laplacian operator on the unit sphere, which implies

(5.7)
$$\Delta_{\mathbf{\Omega}} Y_{\ell m}(\Omega) = \left[\frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta}\right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2}\right] Y_{\ell m}(\mathbf{\Omega}) \\ = -l(l+1)Y_{\ell m}(\mathbf{\Omega})$$

for $l \geq 0$ and $m = -l, \ldots, 0, \ldots, l$. Next, for the reader's convenience, we include Lemma 4.1 from [14], as it is used throughout the remaining proofs.

LEMMA 5.1. For $N \ge 1$, $|\beta| \le k+1$, and $v \in V \cap (H^{k+1}(R) \times H^2(S^2))$ with $v(\mathbf{x}, \mathbf{\Omega}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega}), \text{ we have the following:}$

- (i) $\|\mathbf{\Omega} \cdot \nabla v\| \le \sqrt{3} \sum_{i=1}^{3} \left\| \frac{\partial v}{\partial x_i} \right\|.$ (ii) $\left\| D_{\mathbf{x}}^{\beta} \phi_{\ell m}(\mathbf{x}) \right\|^{2} \leq \frac{1}{[l(l+1)]^{2}} \int_{S^{2}} |\Delta_{\mathbf{\Omega}} D_{\mathbf{x}}^{\beta} v(\mathbf{x}, \mathbf{\Omega})|^{2} d\mathbf{\Omega}.$ (iii) $\left\| D_{\mathbf{x}}^{\beta} \mathcal{E}_{N}(v) \right\| \leq \frac{2}{N+1} \left\| \Delta_{\mathbf{\Omega}} D_{\mathbf{x}}^{\beta} v \right\|.$ (iv) $b_{I}(v, v) \leq \|v\|_{0,0,\partial R}^{2}.$

- (v) $\|\mathcal{E}_N(v)\|_{0,0,\partial R} \leq \frac{2}{N+1} \|\Delta_{\mathbf{\Omega}}v\|_{0,0,\partial R}$. (vi) If, in addition, v satisfies the asymptotic expansion (5.9), then

$$\left\|\Pi_N(\mathcal{E}_h(v))\right\|_{0,0,\partial R} \le C \frac{1}{\sigma_t} h^k \left\|\Delta_{\mathbf{\Omega}} \phi_R\right\|_{k+1,\partial R}.$$

Proof. See the proof in [14].

In the following, we present theorems for the thin regime (without proof) and for the thick regime with highly anisotropic scattering.

THEOREM 5.2 (thin regime). Suppose that $N > N_S \ge 1, 0 \le \sigma_a \le \sigma_t \le 1$ and that $\|\cdot\|_{V_1}$ is defined as in (3.12). Let $\psi \in V_1 \cap (H^{k+1}(R) \times H^2(S^2))$ be the solution of (4.5), and let ψ^h be the solution of (5.3) with V^h defined by (5.2). Then we have

$$\left\|\psi - \psi^{h}\right\|_{V_{1}} \leq \frac{C_{1}}{N+1} \left(\left\|\Delta_{\Omega}\psi\right\|_{1,0} + \left\|\Delta_{\Omega}\psi\right\|_{0,0,\partial R}\right) + C_{2}h^{k} \left(\left\|\psi\right\|_{k+1,0} + \left\|\psi\right\|_{0,\partial R}\right)$$

with C_1 and C_2 independent of σ_t and σ_a .

Proof. See the proof of the isotropic case in [14].

The error bounds for the thick regime require considering the asymptotic limit defined by $\sigma_t \to \infty$. These bounds depend on the asymptotic form of ψ , which itself depends on assumptions regarding material parameters. For the case of isotropic

scattering, this limit has been extensively examined, where it is assumed that $\sigma_a = \zeta/\sigma_t$, where ζ is bounded independently of σ_t as $\sigma_t \to \infty$ (see [7, 6, 9, 8, 17]). In this limit, referred to as the diffusion limit, the solution to (2.1) with isotropic scattering can then be expressed as

$$\psi(\mathbf{x}, \mathbf{\Omega}) = \phi_D(\mathbf{x}) + \frac{1}{\sigma_t} \phi_R(\mathbf{x}, \mathbf{\Omega}),$$

with ϕ_R bounded independently of σ_t and the leading-order term ϕ_D satisfying a diffusion equation.

In [10], for anisotropic scattering Larsen and Pomraning presented two different asymptotic limits for $\sigma_t \to \infty$. These two different limits rely on different assumptions on the degree of anisotropy in the scattering. One is for the case of mildly anisotropic scattering, and the other is for highly anisotropic scattering. In the following, we only examine the case of highly anisotropic scattering because the case of mildly anisotropic scattering yields results identical to Theorem 4.3 of [14].

To define the asymptotic limit, we let ζ_l and ω_l be O(1) constants. Then we define μ_ℓ in terms of these constants as

(5.8)
$$\mu_{\ell} = \begin{cases} \zeta_l, & 0 \le l \le N_S, \\ \frac{\sigma_t}{\omega_l + 1}, & l > N_S. \end{cases}$$

Note that with these assumptions, we get that $\sigma_a = O(1)$ and

$$\sigma_l = \begin{cases} \frac{\sigma_t - \zeta_l}{\sigma_t - \sigma_a}, & 0 \le l \le N_S, \\ \frac{\sigma_t \omega_l}{\sigma_s(\omega_l + 1)}, & l > N_S. \end{cases}$$

Under these assumptions, Larsen and Pomraning in [10] illustrated that ψ can be expressed as

(5.9)
$$\psi(\mathbf{x}, \mathbf{\Omega}) = \hat{\phi}_D(\mathbf{x}, \mathbf{\Omega}) + \frac{1}{\sigma_t} \hat{\phi}_R(\mathbf{x}, \mathbf{\Omega}),$$

where

$$\hat{\phi}_D(\mathbf{x}, \mathbf{\Omega}) := \sum_{\ell=0}^{N_S} \sum_{m=-\ell}^{\ell} \phi_{\ell m}(\mathbf{x}) Y_{\ell m}(\mathbf{\Omega})$$

satisfies the first-order P_{N_S} equations and $\hat{\phi}_R(\mathbf{x}, \mathbf{\Omega})$ can be bounded independently of σ_t .

Remark. The P_{N_S} equations are a set of $(N_S + 1)^2$ differential equations for $(N_S + 1)^2$ unknowns, which are obtained by substituting $\psi_{N_S} := \prod_{N_S} \psi$ for ψ in (2.1) and setting the resulting equation orthogonal to all spherical harmonics up to order N_S . Furthermore, as was illustrated in [14], the least-squares formulation described here is nearly identical to a least-squares minimization of the P_{N_S+1} equations.

The two components, $\hat{\phi}_D$ and $\hat{\phi}_R$, of (5.9) are not orthogonal in $L^2(S^2)$. But we can rewrite (5.9) such that this condition holds. Note that this condition is employed in the proof of Theorem 5.3. This new expression is

(5.10)
$$\psi(\mathbf{x}, \mathbf{\Omega}) = \phi_D(\mathbf{x}, \mathbf{\Omega}) + \frac{1}{\sigma_t} \phi_R(\mathbf{x}, \mathbf{\Omega}),$$

where $\mathcal{P}_{\Sigma} \phi_R = 0$ for $\Sigma = \{l \in \mathbb{N} : l \leq N_S\}$. Before proving Theorem 5.3, we introduce the notation $||a|| \leq ||b||$ meaning $||a|| \leq C ||b||$, where C denotes an arbitrary, parameter-independent, positive constant.

THEOREM 5.3. (diffusive regime with highly anisotropic scattering) Suppose that $N > N_S \ge 1, 1 \le \sigma_t < \infty$ and that $\|\cdot\|_{V_1}$ is defined as in (3.12). Let $\psi \in V_1 \cap$ $(H^{k+1}(R) \times H^2(S^2))$ be the solution of (4.5), and let ψ^h be the solution of (5.3) with V^h as defined in (5.2) using the scaling operator (4.18). Assuming that $\zeta_{\ell} \in (\zeta_m, \zeta_M)$ and $\omega_{\ell} \in (\omega_m, \omega_M) \forall l$, where the minimum and maximum terms are O(1) and independent of σ_t , and assuming that ψ satisfies the expansion (5.10), then

$$\left\| \psi - \psi^h \right\|_{V_1} \leq \frac{C_1 D_1(\sigma_t, \phi_R)}{\sigma_t^{1/2}(N+1)} + C_2 D_2(\sigma_t, \phi_D, \phi_R) h^k$$

with C_1 and C_2 independent of σ_t and σ_a , and

$$D_1(\sigma_t, \phi_R) := \frac{1}{\sigma_t} \sum_{i=1}^3 \left\| \Delta_{\mathbf{\Omega}} \frac{\partial \phi_R}{\partial x_i} \right\| + \left\| \Delta_{\mathbf{\Omega}} \phi_R \right\| + \frac{1}{\sigma_t^{1/2}} \left\| \Delta_{\mathbf{\Omega}} \phi_R \right\|_{0,0,\partial R},$$

and

$$D_2(\sigma_t, \phi_D, \phi_R) := \|\phi_D\|_{k+1,0} + \|\Delta_{\Omega}\phi_D\|_{k+1,0,\partial R} + \frac{1}{\sigma_t^{3/2}} \|\phi_R\|_{k+1,0} + \frac{1}{\sigma_t} \|\Delta_{\Omega}\phi_R\|_{k+1,0,\partial R}.$$

Proof. Combining Céa's lemma with Theorem 4.4 yields

$$\|\psi - \psi_h\|_{V_1} \lesssim (\|\mathcal{E}_N(\psi)\|_{V_1} + \|\Pi_N(\psi - \Pi_h\psi)\|_{V_1}).$$

Using

$$\|v\|_{V_1} \le \left(\|v\|_V^2 + \|v\|_{0,0,\partial R}^2\right)^{1/2} \le \|v\|_V + \|v\|_{0,0,\partial R},$$

which is obtained from

$$b_I(v,v) \le \int_{\partial R} \int_{S^2} |v|^2 = ||v||_{0,0,\partial R}^2,$$

we have

(5.11)
$$\|\psi - \psi_h\|_{V_1} \lesssim \Big(\|\mathcal{E}_N(\psi)\|_V + \|\mathcal{E}_N(\psi)\|_{0,0,\partial R} + \|\Pi_N(\mathcal{E}_h(\psi))\|_V + \|\Pi_N(\mathcal{E}_h(\psi))\|_{0,0,\partial R} \Big).$$

Next, we note that $\sigma_t \mathcal{E}_N(\psi) = \mathcal{E}_N(\phi_R)$ because of the fact that ψ satisfies (5.10). $\Pi_{N_S}(\mathbf{\Omega} \cdot \nabla(\mathcal{E}_N(\phi_R))) = 0$ and $\Pi_{N_S}(\mathcal{E}_N(\phi_R)) = 0$ because of our our assumptions on N. Now, we bound the first term of (5.11) as

$$\begin{split} \|\mathcal{E}_{N}(\psi)\|_{V} &= \frac{1}{\sigma_{t}} \|\mathcal{E}_{N}(\phi_{R})\|_{V} \\ &\lesssim \frac{1}{\sigma_{t}^{3/2}} \|\mathcal{E}_{N_{S}}((\mathbf{\Omega} \cdot \nabla) \mathcal{E}_{N}(\phi_{R}))\| + \frac{1}{\sigma_{t}^{1/2}} \|\mathcal{E}_{N_{S}}(\mathcal{E}_{N}(\phi_{R}))\| \\ &\lesssim \frac{1}{\sigma_{t}^{1/2}} \left(\frac{1}{\sigma_{t}} \sum_{i=1}^{3} \left| \left| \frac{\partial}{\partial x_{i}} \mathcal{E}_{N}(\phi_{R}) \right| \right| + \|\mathcal{E}_{N}(\phi_{R})\| \right) \\ &\lesssim \frac{1}{\sigma_{t}^{1/2}(N+1)} \left(\frac{1}{\sigma_{t}} \sum_{i=1}^{3} \left| \left| \Delta_{\mathbf{\Omega}} \frac{\partial \phi_{R}}{\partial x_{i}} \right| \right| + \|\Delta_{\mathbf{\Omega}} \phi_{R}\| \right), \end{split}$$

where we used (i) and (iii) of Lemma 5.1 and the fact that $(\mathcal{I} - \Pi_{N_S})$ is an $L^2(S^2)$ orthogonal projection. We bound the second term of (5.11) according to

$$\left\|\mathcal{E}_{N}(\psi)\right\|_{0,0,\partial R} = \frac{1}{\sigma_{t}} \left\|\mathcal{E}_{N}(\phi_{R})\right\|_{0,0,\partial R} \leq \frac{2}{\sigma_{t}(N+1)} \left\|\Delta_{\Omega}\phi_{R}\right\|_{0,0,\partial R},$$

given (v) of Lemma 5.1.

For the third term of (5.11), we first need to introduce $\mathcal{O}_{N_S} := \prod_{N_S} (\mathbf{\Omega} \cdot \nabla)$. Since $\sigma_a > 1/\sigma_t$, we have

$$||\Pi_N(\mathcal{E}_h(\psi))||_V = ||\mathcal{E}_h(\phi_D)||_V + \frac{1}{\sigma_t} ||\Pi_N(\mathcal{E}_h(\phi_R))||_V.$$

We then say

$$\begin{split} ||\mathcal{E}_{h}(\phi_{D})||_{V} \lesssim \frac{1}{\sigma_{t}^{1/2}} ||\mathcal{E}_{N_{S}}((\mathbf{\Omega} \cdot \nabla) \mathcal{E}_{h}(\phi_{D}))|| + ||\mathcal{O}_{N_{S}} \mathcal{E}_{h}(\phi_{D})|| + ||\mathcal{E}_{h}(\phi_{D})|| \\ \lesssim ||\mathbf{\Omega} \cdot \nabla(\mathcal{E}_{h}(\phi_{D}))|| + ||\mathcal{E}_{h}(\phi_{D})|| \\ \lesssim \sum_{i=1}^{3} \left| \left| \frac{\partial}{\partial x_{i}} \mathcal{E}_{h}(\phi_{D}) \right| \right| + ||\mathcal{E}_{h}(\phi_{D})|| \\ \lesssim h^{k} \left(||\phi_{D}||_{k+1,0} + h ||\phi_{D}||_{k+1,0} \right) \\ \lesssim h^{k} ||\phi_{D}||_{k+1,0} \end{split}$$

and

$$\begin{split} ||\Pi_{N}(\mathcal{E}_{h}(\phi_{R}))||_{V} \lesssim \frac{1}{\sigma_{t}} ||\mathcal{O}_{N_{S}}\mathcal{P}_{N_{S}+1}(\mathcal{E}_{h}(\phi_{R}))|| + \frac{1}{\sigma_{t}^{3/2}} ||\mathcal{E}_{N_{S}}((\mathbf{\Omega}\cdot\nabla)\Pi_{N}\mathcal{E}_{h}(\phi_{R}))|| \\ + \frac{1}{\sigma_{t}^{1/2}} ||\Pi_{N}\mathcal{E}_{h}(\phi_{R})|| \\ \lesssim \frac{1}{\sigma_{t}} ||\mathbf{\Omega}\cdot\nabla(\mathcal{E}_{h}(\phi_{R}))|| + \frac{1}{\sigma_{t}^{1/2}} ||\Pi_{N}\mathcal{E}_{h}(\phi_{R})|| \\ \lesssim \sum_{i=1}^{3} \frac{1}{\sigma_{t}} \left| \left| \frac{\partial}{\partial x_{i}}\mathcal{E}_{h}(\phi_{R}) \right| \right| + \frac{1}{\sigma_{t}^{1/2}} ||\mathcal{E}_{h}(\phi_{R})|| \\ \lesssim \frac{h^{k}}{\sigma_{t}^{1/2}} \left(\frac{1}{\sigma_{t}^{1/2}} ||\phi_{R}||_{k+1,0} + h ||\phi_{R}||_{k+1,0} \right) \\ \lesssim \frac{h^{k}}{\sigma_{t}^{1/2}} ||\phi_{R}||_{k+1,0} \,. \end{split}$$

Subsequently,

$$\|\Pi_N(\mathcal{E}_h(\psi))\|_V \lesssim h^k \left(\|\phi_D\|_{k+1,0} + \frac{1}{\sigma_t^{3/2}} \|\phi_R\|_{k+1,0} \right).$$

Last, we can bound the fourth term of (5.11) according to

$$\begin{split} \|\Pi_N(\mathcal{E}_h(\psi))\|_{0,0,\partial R} \lesssim h^k \, \|\Delta_{\Omega}\psi\|_{k+1,0,\partial R} \lesssim \|\Delta_{\Omega}\phi_D\|_{k+1,0,\partial R} + \frac{1}{\sigma_t} \, \|\Delta_{\Omega}\phi_R\|_{k+1,0,\partial R} \,, \end{split}$$
 where we have used (iv) of Lemma 5.1.

6. Final remarks. In this paper, we have extended the least-squares method for the linear Boltzmann equation to the case of anisotropic scattering by establishing uniqueness and existence of the minimization problem (4.5). Furthermore, the ellipticity is with respect to a physically meaningful norm, and the ellipticity constants are independent of the problem parameters. Using the ellipticity constants, we have also established error bounds in all three parameter regimes.

Future work consists of examining the least-squares approach with respect to more complex discretization approaches. Besides spherical harmonics approximations, one can use tessellations of the sphere as a finite element representation of the angular dependency. One of the main advantages to this approach is that there is a reduced coupling among moments as compared to the spherical harmonics approach used here. For the spatial domain, we plan to investigate nonconforming finite elements so as to better approximate problems having discontinuous solutions. Lastly, since this work provides a partial foundation for [5], we hope to provide a complete foundation by extending the results introduced here to the case of multiple energy groups.

REFERENCES

- [1] R. A. ADAMS, Sobolev Spaces, Academic Press, New York, 1975.
- [2] B. AKHERRAZ, C. FEDON-MAGNAUD, J. LAUTARD, AND R. SANCHEZ, Anisotropic scattering treatment for the neutron-transport equation with primal finite-elements, Nuclear Sci. Engrg., 120 (1995), pp. 187–198.
- [3] G. ARFKEN, Mathematical Methods for Physicists, 3rd ed., Academic Press, San Diego, 1985.
- [4] D. BRAESS, *Finite Elements*, Cambridge University Press, Cambridge, 1997.
- B. CHANG AND B. LEE, A multigrid algorithm for solving the multigroup, anisotropic scattering Boltzmann equation using first-order system least-squares methodology, Electron. Trans. Numer. Anal., 15 (2001), pp. 132–151.
- [6] E. LARSEN, Diffusion theory as an asymptotic limit of transport theory for nearly critical systems with small mean free path, Ann. of Nuclear Energy, 7 (1980), pp. 249–255.
- [7] E. LARSEN, On numerical solutions of transport problems in the diffusion limit, Nuclear Sci. Engrg., 83 (1983), pp. 90–99.
- [8] E. LARSEN AND J. MOREL, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II, J. of Comput. Phys., 83 (1989), p. 212.
- E. LARSEN, J. MOREL, AND W. MILLER, Asymptotic solutions of numerical transport problems in optically thick diffusive regimes, J. of Comput. Phys., 69 (1987), pp. 283–324.
- [10] E. LARSEN AND G. POMRANING, The P_N theory as an asymptotic limit of transport theory in planar geometry I: Analysis, Nuclear Sci. Engrg., 109 (1991), pp. 49–75.
- [11] E. LEWIS AND W. MILLER, Computational Methods of Neutron Transport, American Nuclear Society, La Grange Park, IL, 1993.
- [12] T. MANTEUFFEL AND K. RESSEL, Multigrid methods for transport equations in diffusive regimes, in Proceedings on the Copper Mountain Conference on Multigrid Methods, 1995.
- [13] T. MANTEUFFEL AND K. RESSEL, Least-squares finite element solution of the neutron transport equation in diffusive regimes, SIAM J. Numer. Anal., 35 (1998), pp. 806–835.
- [14] T. MANTEUFFEL, K. RESSEL, AND G. STARKE, A boundary functional for the least-squares finite element solution of neutron transport problems, SIAM J. Numer. Anal., 37 (2000), pp. 556–586.
- [15] J. E. MOREL, Fokker-Planck calculations using standard discrete ordinates transport codes, Nuclear Sci. Engrg., 79 (1981), pp. 340–356.
- [16] G. PALMIOTTI, C. B. CARRICO, AND E. E. LEWIS, Variational nodal transport methods with anisotropic scattering, Nuclear Sci. Engrg., 115 (1993), pp. 233–243.
- [17] G. POMRANING, Diffusive limits for linear transport equations, Nuclear Sci. Engrg., 112 (1992), pp. 239–255.