SIAM J. NUMER. ANAL.
Vol. 35, No. 2, pp. 806–835, April 1998
© 1998 Society for Industrial and Applied Mathematics
022

LEAST-SQUARES FINITE-ELEMENT SOLUTION OF THE NEUTRON TRANSPORT EQUATION IN DIFFUSIVE REGIMES*

THOMAS A. MANTEUFFEL† AND KLAUS J. RESSEL‡

Abstract. A systematic solution approach for the neutron transport equation, based on a least-squares finite-element discretization, is presented. This approach includes the theory for the existence and uniqueness of the analytical as well as of the discrete solution, bounds for the discretization error, and guidance for the development of an efficient multigrid solver for the resulting discrete problem. To guarantee the accuracy of the discrete solution for diffusive regimes, a scaling transformation is applied to the transport operator prior to the discretization. The key result is the proof of the V-ellipticity and continuity of the scaled least-squares bilinear form with constants that are independent of the total cross section and the absorption cross section. For a variety of least-squares finite-element discretizations this leads to error bounds that remain valid in diffusive regimes. Moreover, for problems in slab geometry a full multigrid solver is presented with $V(1,1)$-cycle convergence factors approximately equal to 0.1 independent of the size of the total cross section and the absorption cross section.

Key words. Boltzmann equation, neutron transport equation, least-squares variational formulation, finite-element discretization, diffusion limit, multilevel methods

AMS subject classifications. 65N30, 65N55, 65N15

PII. S0036142996299708

1. Introduction.

The deterministic numerical solution of neutron transport problems becomes difficult in diffusive regimes, which are characterized by very large scattering cross sections and very small absorption cross sections. In these regimes the transport equation is nearly singular and its solution in the interior of the computational domain is close to the solution of a diffusion equation. In order to solve diffusive transport problems numerically, it is advantageous to use a discretization for the transport operator that resembles a good approximation of a diffusion operator in diffusive regimes. In the past, special discretizations for transport problems in slab geometry have been developed that have this property. Among them are the diamond difference scheme [18], the linear discontinuous scheme [2], and the modified linear discontinuous scheme [17]. However, these discretizations have the disadvantage that either the solution of the resulting discrete system [20], [21] or their extension to higher dimensions is difficult.

In this paper we present a general framework for constructing discretizations of transport problems that are accurate in diffusive regimes. This framework, which is based on a least-squares variational formulation in combination with a scaling transformation, represents a systematic solution approach, since it includes the theory for the existence and uniqueness of the analytical as well as the discrete solution, bounds for the discretization error, and guidance for the development of an efficient multigrid solver for the resulting discrete problem.

*Received by the editors February 26, 1996; accepted for publication (in revised form) February 14, 1997. This work was supported by DOE grant DE-FG03-93ER25165 and NSF grant DMS-8704169.

http://www.siam.org/journals/sinum/35-2/29970.html

†Program in Applied Mathematics, University of Colorado at Boulder, CB 526, Boulder, CO 80309-0526 (tmanteuf@boulder.colorado.edu).

‡German Remote Sensing Data Center (DFD), German Aerospace Research Establishment (DLR), P.O. Box 1116, 82234 Oberpfaffenhofen, Germany (kjr@dfd.dlr.de).

To introduce our notation we recall that the single group, steady state, isotropic form of the neutron transport equation is given by [18],

$$\left\{ \begin{array}{rcll} [\underline{\Omega}\cdot\underline{\nabla}+\sigma_t I-\sigma_s P]\,\psi(\underline{r},\underline{\Omega}) & = & \tilde{q}(\underline{r},\underline{\Omega}) & \text{for } (\underline{r},\underline{\Omega})\in\mathcal{R}\times\mathcal{S}^1 \\ \psi(\underline{r},\underline{\Omega}) & = & g(\underline{r},\underline{\Omega}) & \text{for } \underline{r}\in\partial\mathcal{R}\ \wedge\ \underline{n}(\underline{r})\cdot\underline{\Omega}<0 \end{array} \right\}, \tag{1.1}$$

where σ_t is the total cross section, σ_s is the scattering cross section, and $\psi(\underline{r},\underline{\Omega})$ is the *angular flux*, to be determined for all points $\underline{r}=(x,y,z)$ in a region $\mathcal{R}\subset\mathbb{R}^3$ with sufficiently smooth boundary (e.g., of class $C^{1,1}$; see [11, p. 5]) and all possible travel directions $\underline{\Omega}=:(\theta,\varphi)$ on the unit sphere $\mathcal{S}^1$. Using the normalization

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

for the incremental solid angle element, the operator P is defined by

$$P\psi(\underline{r},\underline{\Omega}):=\int\limits_{\mathcal{S}^1}\psi(\underline{r},\underline{\Omega}')\,d\Omega', \tag{1.2}$$

which is an L^2-projection onto the space of functions that are independent of direction angle $\underline{\Omega}$. The boundary conditions specify the inflow of particles into the region $\mathcal{R}$, since $\underline{n}(\underline{r})$ denotes the unit outgoing normal at $\underline{r}\in\partial\mathcal{R}$. Such problems arise as the inner loop of time-dependent, multi-energy group problems [18].

In the case of *slab geometry* (semi-infinite slab) it is assumed that $\frac{\partial\psi}{\partial x}=\frac{\partial\psi}{\partial y}\equiv 0$. Defining $\mu:=\cos(\theta)$, where θ denotes the angle between $\underline{\Omega}$ and the z-axis, the flux becomes $\psi(\underline{r},\underline{\Omega})=\psi(z,\mu)$ and equation (1.1) reduces to [18]

$$\left\{ \begin{array}{rcll} \left[\mu\dfrac{\partial}{\partial z}+\sigma_t I-\sigma_s P\right]\psi(z,\mu) & = & \tilde{q}(z,\mu) & \text{for } (z,\mu)\in[z_l,z_r]\times[-1,1] \\ \psi(z_l,\mu) & = & g_l(\mu) & \text{for } \mu>0 \\ \psi(z_r,\mu) & = & g_r(\mu) & \text{for } \mu<0 \end{array} \right\}. \tag{1.3}$$

Now, the operator P is given by

$$P\psi(z,\mu):=\frac{1}{2}\int\limits_{-1}^{1}\psi(z,\mu')\,d\mu', \tag{1.4}$$

which is an L^2-projection onto the space of functions that are independent of μ.

Without loss of generality, we assume in the following vacuum boundary conditions, which means that $g(\underline{r},\underline{\Omega})\equiv 0$ in (1.1) and $g_l(\mu)\equiv g_r(\mu)\equiv 0$ in (1.3), respectively. Furthermore, we suppose that $\operatorname{diam}(\mathcal{R})=1$ in (1.1) and $|z_r-z_l|=1$ in (1.3), respectively. All these assumptions can be established by simple transformations [29].

When $\sigma_t\to\infty$ and $\frac{\sigma_s}{\sigma_t}\to 1$, equations (1.1) become singular. Considering this limit after dividing (1.1) by σ_t results in the limit equation $(I-P)\psi=0$. Therefore, the limit solution is independent of direction $\underline{\Omega}$. The limit solution can be further characterized, if $\sigma_t\to\infty$ and $\frac{\sigma_s}{\sigma_t}\to 1$ in a certain way, which is called *diffusion limit*. Then, it can be shown [15] that the limit solution converges to a solution of a diffusion equation. To be more specific, we introduce the absorption cross section $\sigma_a:=\sigma_t-\sigma_s$ and a small parameter ε. The diffusion limit can then be defined as the limit $\varepsilon\to 0$ after scaling the cross sections and the source in the following way:

$$\tilde{q}(\underline{r},\underline{\Omega})\to\varepsilon q(\underline{r},\underline{\Omega}),\quad \sigma_t\to\frac{1}{\varepsilon},\quad \sigma_a\to\varepsilon\alpha, \tag{1.5}$$

where α and q are assumed to be $O(1)$. In this parameterization the transport equation becomes

$$\widetilde{\mathcal{L}}\psi(\underline{r},\underline{\Omega}) := \left[\underline{\Omega}\cdot\underline{\nabla} + \frac{1}{\varepsilon}(I-P) + \varepsilon\alpha P\right]\psi(\underline{r},\underline{\Omega}) = \varepsilon q(\underline{r},\underline{\Omega}). \tag{1.6}$$

If we assume that $q(\underline{r},\underline{\Omega}) = Pq(\underline{r},\underline{\Omega}) + O(\varepsilon)$, then by an asymptotic expansion in ε it can be proven [15], [28] that the solution of (1.6) has the *diffusion expansion*

$$\psi(\underline{r},\underline{\Omega}) = \phi_0(\underline{r}) + \varepsilon\phi_R(\underline{r},\underline{\Omega}), \tag{1.7}$$

where ϕ_0 is a solution of the diffusion equation

$$-\nabla\cdot\frac{1}{3\sigma_t}\nabla\phi_0(\underline{r}) + \sigma_a\phi_0(\underline{r}) = Pq(\underline{r},\underline{\Omega}). \tag{1.8}$$

Boundary conditions for $\phi_0(\underline{\Omega})$ can be obtained by matching a boundary layer analysis to the asymptotic expansion of the solution in the interior of the computational domain [28], [19] (for slab geometry, see [12]). Since we consider here vacuum boundary conditions for the transport equation (1.1) we have that [28, Eq. 146]:

$$\phi_0(\underline{r}) \equiv 0 \qquad \text{for all } \underline{r} \in \partial\mathcal{R}. \tag{1.9}$$

In the following analysis of a least-squares finite-element discretization of the transport equation (1.1) we use the form (1.6) of the transport operator.

This paper is organized as follows. In section 2 we briefly describe the least-squares finite-element discretization. Further, we introduce and motivate in this section a scaling transformation that is applied to the transport operator prior to the discretization in order to assure the accuracy of the discrete solution for diffusive regimes. In section 3 we prove that the scaled least-squares bilinear form is continuous and V-elliptic in a certain norm with constants independent of ε and α. The existence and uniqueness of the analytical as well as of the discrete problem then follows directly from the Lax–Milgram lemma [7].

Furthermore, the continuity and the V-ellipticity, in combination with Céa's lemma [6], are the basis for discretization error bounds that are established in section 4 for a variety of conforming finite-element spaces. Since the continuity and the V-ellipticity constants are independent of ε and α, these error bounds remain valid for diffusive regimes. Thus, the least-squares discretization of the scaled transport equation with simple conforming finite elements yields an accurate discrete solution even in diffusive regimes. In section 5 we summarize some numerical results, including the convergence rates of a full multigrid solver for slab geometry. Finally, in section 6 we draw some conclusions and discuss directions for future research.

2. Least-squares discretization and scaling transformation. Let us denote the standard inner product and associated norm of $L^2(\mathcal{R}\times\mathcal{S}^1)$ by

$$\langle u,v\rangle := \int_{\mathcal{R}}\int_{\mathcal{S}^1} u\cdot v^*\, d\Omega dr, \qquad \|u\| := \sqrt{\langle u,u\rangle} \quad \forall\, u,v \in L^2(\mathcal{R}\times\mathcal{S}^1),$$

where v^* is the complex[1] conjugate of v. Further, let V be a Hilbert space with underlying norm $\|\cdot\|_V$, which we will specify later. Then, the least-squares variational

[1]We allow here complex-valued functions, since we use in section 4 the expansion of v into spherical harmonics.

formulation of (1.1) is given by

$$\min_{\psi\in V} \widetilde{F}(\psi), \qquad \text{with } \widetilde{F}(\psi) := \int_{\mathcal{R}}\int_{\mathcal{S}^1} \left|\widetilde{\mathcal{L}}\psi(\underline{r},\underline{\Omega}) - \tilde{q}(\underline{r},\underline{\Omega})\right|^2 d\Omega dr. \tag{2.1}$$

A necessary condition for $\psi \in V$ being a minimizer of the functional $\widetilde{F}$ in (2.1) is that the first variation of $\widetilde{F}$ vanishes at ψ for all admissible $v \in V$, which results in the problem: find $\psi \in V$ such that

$$\widetilde{a}(\psi, v) := \left\langle \widetilde{\mathcal{L}}\psi, \widetilde{\mathcal{L}}v \right\rangle \;=\; \left\langle \tilde{q}, \widetilde{\mathcal{L}}v \right\rangle \quad \forall v \in V. \tag{2.2}$$

For the least-squares finite-element discretization of (2.1) the Hilbert space V is replaced by a finite-dimensional subspace $V^h \subset V$. This leads to the discrete problem: find $\psi_h \in V^h$ such that

$$\widetilde{a}(\psi_h, v_h) = \left\langle \tilde{q}, \widetilde{\mathcal{L}}v_h \right\rangle \quad \forall v_h \in V^h. \tag{2.3}$$

By an asymptotic analysis it was shown in [22] and [29] for slab geometry and V^h formed by piecewise linear basis functions in space and a finite number of Legendre polynomials as basis functions in angle that this direct least-squares approach is not accurate in diffusive regimes.[2] This can also be explained by the following heuristic argument. Because of the diffusion expansion (1.7) the important component of the solution ψ in diffusive regimes is the part that is independent of direction angle $\underline{\Omega}$, which is given by $P\psi$. On the other hand, the component $(I-P)\psi$ of the solution is irrelevant in diffusive regimes. By Céa's lemma [7], the solution of the least-squares discretization can be viewed as the best approximation to the exact solution in the discrete space V^h with respect to the semi-norm $\sqrt{\widetilde{a}(\cdot,\cdot)} := \sqrt{\langle\widetilde{\mathcal{L}}\cdot, \widetilde{\mathcal{L}}\cdot\rangle}$. However, the different terms in the operator $\widetilde{\mathcal{L}}$, as defined in (1.6), are unbalanced (there are $O(\frac{1}{\varepsilon})$, $O(1)$, and $O(\varepsilon)$ terms), so that different components of the approximation error are weighted differently in $\sqrt{\widetilde{a}(\cdot,\cdot)}$. The leading term of $\widetilde{\mathcal{L}}$ is $\frac{1}{\varepsilon}(I-P)$, which means that the part of the error that is dependent on angle is weighted in this norm very strongly in diffusive regimes (very small ε), even though this part is irrelevant. On the contrary, the part of the error that is independent of angle, which is the important part in diffusive regimes, is hardly measured in the seminorm $\sqrt{\widetilde{a}(\cdot,\cdot)}$, since it is weighted by ε.

The idea is to scale equation (1.6), thus changing the weighting in the norm used in the least-squares discretization, which, in turn, alters the choice of the element of the discrete space as an approximation to the exact solution. Let us define the following *scaling transformation* and its inverse:

$$S := P + \varepsilon(I-P), \qquad S^{-1} = P + \frac{1}{\varepsilon}(I-P). \tag{2.4}$$

In real computations, where the equation is given in its original form (1.1), this scaling transformation converts according to (1.5) into $S = P + 1/\sigma_t(\underline{r})(I-P)$. Clearly, applying the scaling transformation S from the left to the transport equation prior to

[2] Piecewise quadratic functions in space are sufficient to yield an approximation that is correct in the diffusive limit, but using quadratic functions in combination with the scaling transformation (2.4) will increase the accuracy [29].

the least-squares discretization will increase the relative weight of the important error component and decrease the weight for the irrelevant component. After applying the scaling transformation S from the left and dividing by ε, equation (1.6) becomes

$$\mathcal{L}\psi := \frac{1}{\varepsilon} S\widetilde{\mathcal{L}}\psi = \frac{1}{\varepsilon} S\underline{\Omega} \cdot \underline{\nabla}\psi + \frac{1}{\varepsilon}(I-P)\psi + \alpha P\psi = q_s, \tag{2.5}$$

with $q_s := Sq$.

Equation (2.5) can be balanced further by applying the scaling transformation S also from the right. Let the domain of operator $\mathcal{L}$ in (2.5) be the Hilbert space V. Then, we define a space $\widehat{V}$ by

$$\widehat{V} := S^{-1}V, \tag{2.6}$$

so that

$$\widehat{v} = S^{-1}v \ \text{ and } \ S\widehat{v} = v \tag{2.7}$$

for all $v \in V$ and $\widehat{v} \in \widehat{V}$. Scaling (2.5) from the right results in

$$\mathcal{L}SS^{-1}\psi = \mathcal{L}S\widehat{\psi} = \underline{Q} \cdot \underline{\nabla}\widehat{\psi} + (I-P)\widehat{\psi} + \alpha P\widehat{\psi} = q_s, \tag{2.8}$$

where

$$\underline{Q} := \frac{1}{\varepsilon} S\underline{\Omega}S = (1-\varepsilon)\,(P\underline{\Omega} + \underline{\Omega}P) + \varepsilon\underline{\Omega}I.$$

In the double-scaled operator $\mathcal{L}S$ in (2.8) the derivative of zeroth moment ($P\underline{\nabla}\widehat{\psi}$), the derivative of the first moments ($P\underline{\Omega} \cdot \underline{\nabla}\widehat{\psi}$), and all components of $\widehat{\psi}$ themselves are weighted equally. Moreover, it is easily seen that the double-scaled operator $\mathcal{L}S$ goes to a bounded nonsingular limit operator as $\varepsilon \to 0$.

In the least-squares context, the additional scaling from the right can be avoided because

$$\min_{\widehat{\psi}\in\widehat{V}} \left\langle \mathcal{L}S\widehat{\psi} - q_s, \mathcal{L}S\widehat{\psi} - q_s \right\rangle \quad \Longleftrightarrow \quad \min_{\psi\in V} \left\langle \mathcal{L}\psi - q_s, \mathcal{L}\psi - q_s \right\rangle, \tag{2.9}$$

which will simplify the boundary conditions and thus the computations. However, for the theory we exploit the nice form of the double-scaled operator $\mathcal{L}S$ and use this form of the transport operator as a tool.

The least-squares variational formulation of the single-scaled equation (2.5) is given by the problem: find $\psi \in V$ such that

$$a(\psi, v) := \langle \mathcal{L}\psi, \mathcal{L}v \rangle = \langle q_s, \mathcal{L}v \rangle \quad \forall v \in V, \tag{2.10}$$

whereas, for the bilinear form that is associated to the least-squares variational formulation of the double-scaled equation (2.8) we have

$$\widehat{a}(\widehat{\psi}, \widehat{v}) := \left\langle \mathcal{L}S\widehat{\psi}, \mathcal{L}S\widehat{v} \right\rangle = \langle \mathcal{L}\psi, \mathcal{L}v \rangle = a(\psi, v). \tag{2.11}$$

For the sake of completeness we remark that for slab geometry the form of the scaling transformation S, as defined in (2.4), remains the same, except that for P the

definition (1.4) must be used. In the case of slab geometry, therefore, equation (2.5) reduces to

$$\mathcal{L}\psi := \frac{1}{\varepsilon} S\widetilde{\mathcal{L}}\psi = \frac{1}{\varepsilon} S\mu\frac{\partial\psi}{\partial z} + \frac{1}{\varepsilon}(I-P)\psi + \alpha P\psi = q_s, \tag{2.12}$$

and the double-scaled equation (2.8) becomes

$$\mathcal{L}SS^{-1}\psi = \mathcal{L}S\widehat{\psi} = \frac{1}{\varepsilon} S\mu S\frac{\partial\widehat{\psi}}{\partial z} + (I-P)\widehat{\psi} + \alpha P\widehat{\psi} = q_s, \tag{2.13}$$

where

$$\frac{1}{\varepsilon} S\mu S = (1-\varepsilon)(P\mu+\mu P) + \varepsilon\mu I =: Q.$$

3. Continuity and V-ellipticity. In this section we prove that the scaled least-squares bilinear form (2.10) is *continuous*; i.e., there exists a constant $C_c > 0$ such that for every $u, v \in V$

$$|a(u,v)| \le C_c \|u\|_V \|v\|_V, \tag{3.1}$$

and *V-elliptic*; i.e., there exists a constant $C_e > 0$ such that for all $v \in V$:

$$a(v,v) \ge C_e \|v\|_V^2. \tag{3.2}$$

The Hilbert space V and its norm $\|\cdot\|_V$ are specified below. It is crucial to prove these bounds with constants C_e and C_c that are independent of ε and α, since this makes it possible to establish discretization error bounds that remain valid in diffusive regimes. We first consider the slab geometry case and later generalize our results to three-dimensional geometry.

3.1. Slab geometry. Let $D := [z_l, z_r] \times [-1,1]$ denote the computational domain and let $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the standard inner product and the associated norm of $L^2(D)$, which are defined by

$$\langle u, v\rangle := \int_{z_l}^{z_r}\int_{-1}^{1} u\cdot v \, d\mu dx \quad \text{and} \quad \|u\| := \sqrt{\langle u,u\rangle}.$$

First, we have to choose an appropriate Hilbert space V with associated norm $\|\cdot\|_V$, for which we prove (3.1) and (3.2). From the view of standard elliptic boundary value problems, the choice $V = H^1([z_l,z_r]) \times L^2([-1,1])$ with the norm $\|v\|_{1,0}^2 := \left\|\frac{\partial v}{\partial z}\right\|^2 + \|v\|^2$ seems natural. However, by means of a counterexample it is possible to show [29] that the bilinear form defined in (2.10) cannot be bounded from below in this norm. Nevertheless, by looking at the double-scaled bilinear form (2.11) and (2.8) we observe that the norm

$$\|\widehat{v}\|_{\widehat{V}}^2 := \left\|Q\frac{\partial\widehat{v}}{\partial z}\right\|^2 + \|\widehat{v}\|^2 \tag{3.3}$$

would be a natural choice for bounding this bilinear from. Together with the relation (2.7) we can now derive a norm for bounding the single-scaled bilinear form (2.10) in

the following way:

$$\begin{aligned} \|\widehat{v}\|_{\widehat{V}}^2 &= \left\| \frac{1}{\varepsilon} S\mu S S^{-1} \frac{\partial v}{\partial z} \right\|^2 + \left\| S^{-1} v \right\|^2 \\ &= \left\| \frac{1}{\varepsilon} S\mu \frac{\partial v}{\partial z} \right\|^2 + \left\| \frac{1}{\varepsilon} (I-P) v \right\|^2 + \|Pv\|^2 =: \|v\|_V . \end{aligned} \tag{3.4}$$

Then the Hilbert space V can be defined by

$$V := \overline{\left\{ v \in C^\infty(\overline{D});\ v(z_l,\mu) = 0 \text{ for } \mu > 0;\ v(z_r,\mu) = 0 \text{ for } \mu < 0 \right\}}, \tag{3.5}$$

where the closure is taken with respect to the norm $\|\cdot\|_V$.

From the Cauchy–Schwarz inequality and discrete Hölder inequality we obtain that for all $u, v \in V$

$$|a(u,v)| = |\langle \mathcal{L}u, \mathcal{L}v \rangle| \leq \|\mathcal{L}u\| \, \|\mathcal{L}v\| \leq 3 \|u\|_V \, \|v\|_V . \tag{3.6}$$

Thus, the bilinear form (2.10) is continuous with respect to the norm $\|\cdot\|_V$ with $C_c = 3$.

To prove the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ we exploit the convenient form of the double-scaled transport operator and prove first that the double-scaled bilinear form $\widehat{a}(\cdot,\cdot)$ in (2.11) is $\widehat{V}$-elliptic with respect to $\|\cdot\|_{\widehat{V}}$. For this we need the following lemmas.

LEMMA 3.1. *For $\varepsilon \leq 1$ and for all $u, v \in V, \widehat{v} \in \widehat{V}$ we have*

(i) $\frac{1}{\sqrt{3}} \|P\widehat{v}\| - \varepsilon \|(I-P)\widehat{v}\| \leq \|\mu v\|$,
(ii) $\left\langle \mu \frac{\partial v}{\partial z}, v \right\rangle \geq 0$,
(iii) $\left\langle Q \frac{\partial \widehat{v}}{\partial z}, \widehat{v} \right\rangle \geq 0$,
(iv) $\|v\| \leq \left\| \frac{1}{\varepsilon} S v \right\|$.

Proof.

(i) Since P and $(I-P)$ are orthogonal projections it follows directly that

$$\|\mu v\| = \|\mu S \widehat{v}\| \geq \|\mu P \widehat{v}\| - \varepsilon \|\mu (I-P)\widehat{v}\| \geq \frac{1}{\sqrt{3}} \|P\widehat{v}\| - \varepsilon \|(I-P)\widehat{v}\| .$$

(ii) Applying integration by parts with respect to z and taking the boundary conditions into account, it follows that

$$\left\langle \mu \frac{\partial v}{\partial z}, v \right\rangle = \frac{1}{2} \left\{ \int_0^1 \mu v^2(z_r,\mu)\, d\mu - \int_{-1}^0 \mu v^2(z_l,\mu)\, d\mu \right\} \geq 0.$$

(iii) Since P and $(I-P)$ are orthogonal projections we have $\langle Su, v \rangle = \langle u, Sv \rangle$. Therefore, using relation (2.7) and (2.13) leads to

$$\left\langle Q \frac{\partial \widehat{v}}{\partial z}, \widehat{v} \right\rangle = \left\langle \frac{1}{\varepsilon} S\mu S \frac{\partial \widehat{v}}{\partial z}, \widehat{v} \right\rangle = \frac{1}{\varepsilon} \left\langle \mu S \frac{\partial \widehat{v}}{\partial z}, S\widehat{v} \right\rangle = \frac{1}{\varepsilon} \left\langle \mu \frac{\partial v}{\partial z}, v \right\rangle \geq 0,$$

where the last inequality follows from (ii).

(iv) Since by assumption $\varepsilon \leq 1$ it follows that

$$\|v\|^2 = \|Pv\|^2 + \|(I-P)v\|^2 \leq \frac{1}{\varepsilon^2} \|Pv\|^2 + \|(I-P)v\|^2 = \left\| \frac{1}{\varepsilon} S v \right\|^2 . \qquad \square$$

A standard tool for a V-ellipticity proof is a Poincaré–Friedrichs inequality, which is established in the next lemma.

LEMMA 3.2 (Poincaré–Friedrichs inequality). *For $\varepsilon \leq 1$ and for all $v \in V$ we have*

$$\|\mu v\| \;\leq\; \left\|\mu \frac{\partial v}{\partial z}\right\| \;\leq\; \left\|\frac{1}{\varepsilon} S\mu S \frac{\partial \hat{v}}{\partial z}\right\| \;=\; \left\|Q\frac{\partial \widehat{v}}{\partial z}\right\|. \tag{3.7}$$

Proof. Because of the boundary conditions we notice that

$$\mu v(z,\mu) = \begin{cases} \displaystyle\int_{z_l}^{z} \mu \frac{\partial v(s,\mu)}{\partial s} ds & \text{for } \mu \geq 0, \\ \\ \displaystyle -\int_{z}^{z_r} \mu \frac{\partial v(s,\mu)}{\partial s} ds & \text{for } \mu \leq 0, \end{cases}$$

which implies

$$|\mu v(z,\mu)| \leq \int_{z_l}^{z_r} \left|\mu \frac{\partial v}{\partial z}\right| \, dz \;\leq\; \left(\int_{z_l}^{z_r} \left|\mu \frac{\partial v}{\partial z}\right|^2 \, dz\right)^{1/2}.$$

Since by assumption $(z_r - z_l) = 1$, squaring and integrating implies that $\|\mu v\|^2 \leq \left\|\mu \frac{\partial v}{\partial z}\right\|^2$. Applying Lemma 3.1 (iv) and (2.7) leads then directly to (3.7). ∎

The proof of our major result, Theorem 3.4, can be simplified by the following lemma. For a proof of this lemma, which involves tedious but straightforward calculus, we refer to [29].

LEMMA 3.3. *Suppose $0 \leq \varepsilon < \frac{1}{\sqrt{3}}$. Then, for any $\delta \in [0,1]$, there exists $b \geq 0$ such that*

$$H(b,\delta) := \frac{1}{2}\left\{\sqrt{\left(\delta - \left(1+\frac{s}{3}\right)b\right)^2 + 4\delta(1-\delta)} + \left(1-\frac{s}{3}\right)b + \delta\right\} < 0.988,$$

where $s := \left[\delta^{1/2} - \varepsilon\sqrt{3}(1-\delta)^{1/2}\right]^2$. In particular, for $\delta < 0.875$, we can choose $b = 0$. ∎

We are now in a position to prove the central result of this section.

THEOREM 3.4 ($\widehat{V}$-ellipticity of $\widehat{a}(\cdot,\cdot)$). *Suppose that $0 \leq \alpha \leq 1$ and $0 \leq \varepsilon < \frac{1}{\sqrt{3}}$. Let $\widehat{a}(\cdot,\cdot)$ be given as in (2.11), where for $\mathcal{LS}$ the definition (2.13) is used. Further, let $\|\cdot\|_{\widehat{V}}$ be given as in (3.3). Then there exists a constant $C_e \geq 0.012$ independent of ε and α, such that for all $\widehat{v} \in \widehat{V}$,*

$$\begin{aligned} \widehat{a}(\widehat{v},\widehat{v}) \;&=\; \left\|Q\frac{\partial \widehat{v}}{\partial z} + (I-P)\widehat{v} + \alpha P\widehat{v}\right\|^2 \\ &\geq C_e\left(\left\|Q\frac{\partial \widehat{v}}{\partial z}\right\|^2 + \|\widehat{v}\|^2\right) = C_e\,\|\widehat{v}\|_{\widehat{V}}^2. \end{aligned} \tag{3.8}$$

Proof. We have

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) &= \left\| Q\frac{\partial \widehat{v}}{\partial z} + (I-P)\widehat{v} + \alpha P\widehat{v} \right\|^2 \\ &= \left\| Q\frac{\partial \widehat{v}}{\partial z} \right\|^2 + \|(I-P)\widehat{v}\|^2 + \alpha^2 \|P\widehat{v}\|^2 \\ &\quad +2\alpha \left\langle Q\frac{\partial \widehat{v}}{\partial z}, \widehat{v} \right\rangle + 2(1-\alpha)\left\langle Q\frac{\partial \widehat{v}}{\partial z}, (I-P)\widehat{v} \right\rangle . \end{aligned} \tag{3.9}$$

For the last term, we may write for any $d \in [0,1]$, by using that $\langle Pu, v\rangle = \langle Pu, Pv\rangle$ and (iii) of Lemma 3.1:

$$\begin{aligned}\left\langle Q\frac{\partial \hat{v}}{\partial z}, (I-P)\widehat{v} \right\rangle &= d\left\langle Q\frac{\partial \hat{v}}{\partial z}, \widehat{v} \right\rangle - d\left\langle Q\frac{\partial \hat{v}}{\partial z}, P\widehat{v} \right\rangle + (1-d)\left\langle Q\frac{\partial \hat{v}}{\partial z}, (I-P)\widehat{v} \right\rangle \\ &\geq -d\left\langle PQ\frac{\partial \hat{v}}{\partial z}, P\widehat{v} \right\rangle + (1-d)\left\langle (I-P)Q\frac{\partial \hat{v}}{\partial z}, (I-P)\widehat{v} \right\rangle . \end{aligned}$$

Substituting this into (3.9) and bounding the fourth term in (3.9) by (iii) of Lemma 3.1 leads to

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) &\geq \left\| Q\tfrac{\partial \hat{v}}{\partial z} \right\|^2 + \|(I-P)\widehat{v}\|^2 + \alpha^2\|P\widehat{v}\|^2 \\ &\quad -2(1-\alpha)d\left|\left\langle PQ\tfrac{\partial \hat{v}}{\partial z}, P\widehat{v}\right\rangle\right| - 2(1-\alpha)(1-d)\left|\left\langle (I-P)Q\tfrac{\partial \hat{v}}{\partial z}, (I-P)\widehat{v}\right\rangle\right|, \end{aligned}$$

which can be reduced by setting $\alpha = 0$ to

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) &\geq \left\| Q\tfrac{\partial \hat{v}}{\partial z} \right\|^2 + \|(I-P)\widehat{v}\|^2 \\ &\quad -2d\left|\left\langle PQ\tfrac{\partial \hat{v}}{\partial z}, P\widehat{v}\right\rangle\right| - 2(1-d)\left|\left\langle (I-P)Q\tfrac{\partial \hat{v}}{\partial z}, (I-P)\widehat{v}\right\rangle\right|. \end{aligned} \tag{3.10}$$

Defining

$$\delta := \frac{\|P\widehat{v}\|^2}{\|\widehat{v}\|^2}, \qquad \text{so that} \quad (1-\delta) = \frac{\|(I-P)\widehat{v}\|^2}{\|\widehat{v}\|^2},$$

$$\gamma := \frac{\left\|PQ\frac{\partial \hat{v}}{\partial z}\right\|^2}{\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2}, \quad \text{so that} \quad (1-\gamma) = \frac{\left\|(I-P)Q\frac{\partial \hat{v}}{\partial z}\right\|^2}{\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2},$$

and using the Cauchy–Schwarz inequality, we conclude from (3.10) that

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) &\geq \left\| Q\frac{\partial \hat{v}}{\partial z} \right\|^2 + (1-\delta)\|\widehat{v}\|^2 - 2d\sqrt{\delta}\sqrt{\gamma}\left\| Q\frac{\partial \hat{v}}{\partial z} \right\| \cdot \|\widehat{v}\| \\ &\quad -2(1-d)\sqrt{(1-\delta)}\sqrt{(1-\gamma)}\left\| Q\frac{\partial \hat{v}}{\partial z} \right\| \cdot \|\widehat{v}\|. \end{aligned} \tag{3.11}$$

To maximize the lower bound in (3.11), we divide the region $(\delta,\gamma) \in [0,1]\times[0,1]$ into two triangles and choose d as follows:

$$d = \begin{cases} 1 & \text{for} \quad \delta+\gamma \leq 1, \\ 0 & \text{for} \quad \delta+\gamma \geq 1. \end{cases}$$

Next, we consider these two cases separately.

Case 1. $\delta + \gamma \leq 1, d = 1$: for any $\eta > 0$ and any u, v and any norm $\|\cdot\|$, the arithmetic-geometric mean inequality

$$2\,\|u\| \cdot \|v\| \;\leq\; \eta\|u\|^2 + \frac{1}{\eta}\|v\|^2$$

holds. We thus have

$$\widehat{a}(\widehat{v},\widehat{v}) \geq (1-\gamma\eta)\;\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2 + \left(1-\delta-\frac{\delta}{\eta}\right)\;\|\widehat{v}\|^2. \tag{3.12}$$

It remains to choose η such that the terms $(1-\gamma\eta)$ and $(1-\delta-\frac{\delta}{\eta})$ in (3.12) can be bounded by a positive constant from below for all possible γ, δ with $\gamma + \delta \leq 1$. For $\delta < 0.5$, we choose η so that

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

which yields

$$\eta = \frac{\delta + \sqrt{\delta^2 + 4\delta\gamma}}{2\gamma}.$$

Since $\gamma + \delta \leq 1$, thus $\gamma \leq 1-\delta$, we have by Lemma 3.3 that

$$\gamma\eta = \frac{1}{2}\left\{\delta + \sqrt{\delta^2+4\delta\gamma}\right\} \leq \frac{1}{2}\left\{\delta + \sqrt{\delta^2+4\delta(1-\delta)}\right\} = H(0,\delta) \leq 0.988.$$

From (3.12), the $\widehat{V}$-ellipticity of $\widehat{a}(\cdot,\cdot)$ follows directly in this case with $C_e \geq 0.012$.

On the other hand, as $\delta \to 1.0$, we see that $\gamma \to 0$ and $\eta\gamma \to 1.0$ and the coefficients on the right-hand side of (3.12) are not bounded away from zero. To keep them positive, we then rewrite (3.12) for any $b \in [0,1]$ as follows:

$$\widehat{a}(\widehat{v},\widehat{v}) \geq (1-b-\gamma\eta)\;\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2 + b\;\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2 + \left(1-\delta-\frac{\delta}{\eta}\right)\;\|\widehat{v}\|^2.$$

Considering now the case $\delta \geq 0.5$ we have $0 \leq \left(\frac{1}{\sqrt{3}}\sqrt{\delta} - \varepsilon\sqrt{1-\delta}\right)$, since $\varepsilon < \frac{1}{\sqrt{3}}$ by assumption. Therefore, we can square (i) of Lemma 3.1 and use the Poincare–Friedrichs inequality (3.7) of Lemma 3.2 to bound the second term by

$$\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2 \geq \|\mu v\|^2 \geq \left[\frac{1}{\sqrt{3}}\sqrt{\delta} - \varepsilon\sqrt{(1-\delta)}\right]^2 \|\widehat{v}\|^2,$$

which results in

$$\widehat{a}(\widehat{v},\widehat{v}) \;\geq\; (1-b-\gamma\eta)\;\left\|Q\frac{\partial \hat{v}}{\partial z}\right\|^2 \;+\; \left(1-\delta+\frac{b}{3}s-\frac{\delta}{\eta}\right)\;\|\widehat{v}\|^2, \tag{3.13}$$

where

$$s := \left(\sqrt{\delta} - \varepsilon\sqrt{3}\sqrt{(1-\delta)}\right)^2. \tag{3.14}$$

Again, we choose η so that

$$(1-b-\gamma\eta) = \left(1-\delta+\frac{b}{3}s-\frac{\delta}{\eta}\right),$$

which yields

$$\eta = \frac{-\left(b+\frac{b}{3}s-\delta\right)+\sqrt{\left(b+\frac{b}{3}s-\delta\right)^2+4\delta\gamma}}{2\gamma}.$$

Next, to attain a positive constant in the lower bound in (3.13), we need to show that for all possible δ,γ with $\delta+\gamma\le 1$, $\delta\ge 0.5$, a positive b can be selected so that

$$G_1(b,\delta,\gamma) := b+\gamma\eta = \frac{1}{2}\left\{\sqrt{\left(b+\frac{b}{3}s-\delta\right)^2+4\delta\gamma}\ +\ \left(b+\delta-\frac{b}{3}s\right)\right\}\ \le\ C,$$

for some $C<1.0$. Since $G_1(b,\delta,\gamma)\le G_1(b,\delta,1-\delta)$ for $\delta+\gamma\le 1$, it then is sufficient to prove

$$\forall\,\delta\in[0.5,1]\ \ \exists\, b>0\ :\ G_1(b,\delta,1-\delta)<C.$$

But this follows immediately from Lemma 3.3 with $C=0.988$ since $G_1(b,\delta,1-\delta)=H(b,\delta)$. The $\widehat{V}$-ellipticity of $\widehat{a}(\cdot,\cdot)$ for $\gamma+\delta\le 1$ then follows directly from (3.13) with $C_e\ge 1-0.988=0.012$.

Case 2. $\delta+\gamma\ge 1$: setting $d=0$ in (3.11) and proceeding as in Case 1 results in

$$\widehat{a}(\widehat{v},\widehat{v})\ge(1-(1-\gamma)\eta)\ \left\|Q\frac{\partial\hat{v}}{\partial z}\right\|^2+\left(1-\delta-\frac{1-\delta}{\eta}\right)\ \|\widehat{v}\|. \tag{3.15}$$

For $\delta\le 0.5$, we choose

$$\eta=\frac{\delta+\sqrt{\delta^2+4(1-\delta)(1-\gamma)}}{2(1-\gamma)},$$

so that

$$(1-(1-\gamma)\eta)=\left(1-\delta-\frac{1-\delta}{\eta}\right).$$

Since $\delta+\gamma\ge 1$, thus $\delta\ge 1-\gamma$, we then have by Lemma 3.3

$$\begin{aligned}(1-\gamma)\eta\ &=\frac{1}{2}\left\{\sqrt{\delta^2+4(1-\delta)(1-\gamma)}+\delta\right\}\\ &\le\frac{1}{2}\left\{\sqrt{\delta^2+4(1-\delta)\delta}+\delta\right\}=H(0,\delta)\le 0.988.\end{aligned}$$

On the other hand, for $\delta\ge 0.5$, we introduce, as in Case 1, a parameter $b\in[0,1]$ and use the Poincarè–Friedrichs inequality (3.7) to conclude from (3.15) that

$$\widehat{a}(\widehat{v},\widehat{v})\ \ge\ (1-b-(1-\gamma)\eta)\ \left\|Q\frac{\partial\hat{v}}{\partial z}\right\|^2\ +\ \left(1-\delta+\frac{b}{3}s-\frac{1-\delta}{\eta}\right)\ \|\widehat{v}\|\,, \tag{3.16}$$

with s as defined in (3.14). Again, we first choose η, so that

$$(1-b-(1-\gamma)\eta) = \left(1-\delta+\frac{b}{3}s-\frac{1-\delta}{\eta}\right),$$

which yields

$$\eta = \frac{-\left(b+\frac{b}{3}s-\delta\right)+\sqrt{\left(b+\frac{b}{3}s-\delta\right)^2+4(1-\delta)(1-\gamma)}}{2(1-\gamma)}.$$

In order to attain a positive constant in the lower bound (3.16) we need to show that for all possible δ, γ with $\delta+\gamma \geq 1$, $\delta \geq 0.5$, a positive b can be selected so that

$$\begin{aligned} G_2(b,\delta,\gamma) \quad &:= b+(1-\gamma)\eta \\ &= \frac{1}{2}\left\{\sqrt{\left(b+\frac{b}{3}s-\delta\right)^2+4(1-\delta)(1-\gamma)}+\left(b+\delta-\frac{b}{3}s\right)\right\} \;\leq\; C, \end{aligned}$$

for some $C < 1.0$. Since $G_2(b,\delta,\gamma) \leq G_2(b,\delta,1-\delta) = H(b,\delta)$ for $\delta+\gamma \geq 1$, this follows directly from Lemma 3.3 with $C = 0.988$. Finally, from (3.16) with $C_e := 1-C = 0.012$, the $\widehat{V}$-ellipticity of $\widehat{a}(\cdot,\cdot)$ follows for the case $\gamma+\delta \geq 1$, which proves the theorem. □

From (2.11) and (3.4) the V-ellipticity of the single-scaled bilinear form $a(\cdot,\cdot)$ follows directly from the $\widehat{V}$-ellipticity of the double-scaled bilinear form, which is summarized in the following corollary.

COROLLARY 3.5 (V-ellipticity of $a(\cdot,\cdot)$). *Let $a(\cdot,\cdot)$ and $\|\cdot\|_V$ be given as in* (2.10) *and* (3.4). *Assume that $0 \leq \alpha \leq 1$, $0 \leq \varepsilon < \frac{1}{\sqrt{3}}$. Then there exists a constant $C_e \geq 0.012$, independent of α and ε, such that for all $v \in V$,*

$$a(v,v) \;\geq\; C_e\,\|v\|_V^2\,. \qquad \square \tag{3.17}$$

3.2. Three-dimensional geometry. In this section we prove the continuity and the V-ellipticity result in three-dimensional geometry, obtaining a much larger ellipticity, constant, C_e, due to a stronger Poincaré–Friedrichs inequality. Taking into consideration the form of the double-scaled operator (2.8) we see that

$$\|\widehat{v}\|_{\widehat{V}}^2 := \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2 + \|\widehat{v}\|^2 \tag{3.18}$$

would be a natural norm to bound the double-scaled bilinear form (2.11). Similar to the slab geometry case, we use now the relation (2.7) to define a norm for bounding the single-scaled bilinear form $a(\cdot,\cdot)$ in (2.10) by

$$\begin{aligned} \|\widehat{v}\|_{\widehat{V}} \quad &= \left\|S^{-1}v\right\|_{\widehat{V}}^2 \\ &= \left\|\frac{1}{\varepsilon}S\underline{\Omega}\cdot\underline{\nabla}v\right\|^2 + \left\|\frac{1}{\varepsilon}(I-P)v\right\|^2 + \|Pv\|^2 =: \|v\|_V^2. \end{aligned} \tag{3.19}$$

Denoting the computational domain by $D := \mathcal{R}\times\mathcal{S}^1$, we define the following Hilbert space

$$V := \overline{\left\{v \in C^\infty(\overline{D});\; v(\underline{r},\underline{\Omega}) = 0 \text{ for } \underline{r} \in \partial\mathcal{R} \text{ and } \underline{\Omega}\cdot\underline{n}(\underline{r}) < 0\right\}}, \tag{3.20}$$

where the closure is taken with respect to the norm $\|\cdot\|_V$ defined in (3.19).

Analogous to the slab geometry case, the continuity of $a(\cdot,\cdot)$ with respect to $\|\cdot\|_V$ follows immediately with $C_c = 3$.

The V-ellipticity of $a(\cdot,\cdot)$ is proven again via the $\widehat{V}$-ellipticity of $\widehat{a}(\cdot,\cdot)$; the latter can be established by a straightforward generalization of the slab geometry case to three-dimensional geometry. First, we modify Lemmas 3.1–3.3 as follows.

LEMMA 3.6. *For all* $u, v \in V$, $\widehat{v} \in \widehat{V}$, *and* $\varepsilon \leq 1$, *we have*

(i) $\|P\widehat{v}\| \leq \|v\|$,
(ii) $\langle \underline{\Omega} \cdot \underline{\nabla} v, v \rangle \geq 0$,
(iii) $\langle \underline{Q} \cdot \underline{\nabla} \widehat{v}, \widehat{v} \rangle \geq 0$,
(iv) $\|v\| \leq \|\frac{1}{\varepsilon} S v\|$.

Proof.

(i) $\|v\|^2 = \|S\widehat{v}\|^2 = \|P\widehat{v}\|^2 + \|\varepsilon(I-P)\widehat{v}\|^2 \geq \|P\widehat{v}\|^2$.

(ii) By applying Green's formula we obtain

$$\langle \underline{\Omega} \cdot \underline{\nabla} v, v \rangle = -\langle \underline{\Omega} \cdot \underline{\nabla} v, v \rangle + \int_{\partial\mathcal{R}} \int_{\mathcal{S}^1} v^2 \, \underline{\Omega} \cdot \underline{n} \, d\Omega ds.$$

Splitting the boundary $\partial\mathcal{R} \times \mathcal{S}^1$ into the parts $\Gamma^+ := \{(\underline{r}, \underline{\Omega}) \in \partial\mathcal{R} \times \mathcal{S}^1;\ \underline{\Omega} \cdot \underline{n}(\underline{r}) \geq 0\}$ and $\Gamma^- := \{(\underline{r}, \underline{\Omega}) \in \partial\mathcal{R} \times \mathcal{S}^1;\ \underline{\Omega} \cdot \underline{n}(\underline{r}) < 0\}$, it follows that

$$\langle \underline{\Omega} \cdot \underline{\nabla} v, v \rangle = \frac{1}{2} \int_{\Gamma^+} v^2 \underline{\Omega} \cdot \underline{n} \, d\Omega ds \geq 0,$$

since $v(\underline{r}, \underline{\Omega}) = 0$ for $(\underline{r}, \underline{\Omega}) \in \Gamma^-$ by the boundary conditions.

(iii) This follows directly from (ii), since S is self-adjoint with respect to $\langle \cdot, \cdot \rangle$.

(iv) This is the same proof as (iv) of Lemma 3.1. ∎

In three-dimensional geometry, a much stronger Poincaré–Friedrichs inequality can be established. Heuristically, this is due to the fact that, when integrating backwards along any direction Ω, the boundary is encountered within a distance less than the diameter of the domain $\mathcal{R}$, whereas in the semi-infinite slab the domain has unbounded diameter.

LEMMA 3.7 (Poincaré–Friedrichs inequality). *Suppose* $\varepsilon \leq 1$ *and let* $\mathcal{R} \subset \mathbb{R}^3$ *be a bounded domain. Then, for any* $v \in V$, *we have*

$$\|v\| \leq \operatorname{diam}(\mathcal{R}) \ \|\underline{\Omega} \cdot \underline{\nabla} v\| \leq \operatorname{diam}(\mathcal{R}) \ \left\|\underline{Q} \cdot \underline{\nabla} \widehat{v}\right\|. \tag{3.21}$$

Proof. For $\underline{r}_i, \underline{r}_k \in \mathcal{R}$, let

$$[\underline{r}_i, \underline{r}_k] := \{\underline{r} : \ \underline{r} = s\underline{r}_i + (1-s)\underline{r}_k, \text{ for } 0 \leq s \leq 1\}$$

denote the line segment between $\underline{r}_i$ and $\underline{r}_k$. Let $\underline{r} \in \mathcal{R}$ and $\underline{\Omega} \in \mathcal{S}^1$ be given arbitrarily. We define

$$\begin{aligned} t_1 &:= \min\{t \in \mathbb{R}:\ [\underline{r}, \underline{r} + t\underline{\Omega}] \subset \mathcal{R}\}, \\ t_2 &:= \max\{t \in \mathbb{R}:\ [\underline{r}, \underline{r} + t\underline{\Omega}] \subset \mathcal{R}\}, \\ \underline{r}_1 &:= \underline{r} + t_1\underline{\Omega}, \qquad \underline{r}_2 := \underline{r} + t_2\underline{\Omega}. \end{aligned}$$

Then it is easy to see that $\underline{\Omega} \cdot \underline{n}(\underline{r}_1) \leq 0$. Taking into account the boundary conditions for $v \in V$, we have that $v(\underline{r}_1, \underline{\Omega}) = 0$; hence,

$$v(\underline{r}, \underline{\Omega}) = \int_{\underline{r}_1}^{\underline{r}} \frac{\partial v}{\partial s} \, ds = \int_{\underline{r}_1}^{\underline{r}} \underline{\Omega} \cdot \underline{\nabla} v \, ds,$$

where ds denotes the arc-length differential along the line $\{\underline{r} + t\underline{\Omega}, t \in \mathbb{R}\}$. Therefore, we conclude from Jensen's inequality that

$$|v(\underline{r},\underline{\Omega})| \le \int_{\underline{r}_1}^{\underline{r}} |\underline{\Omega}\cdot\underline{\nabla} v|\, ds \le \int_{\underline{r}_1}^{\underline{r}_2} |\underline{\Omega}\cdot\underline{\nabla} v|\, ds \le \operatorname{diam}(\mathcal{R})^{1/2} \left(\int_{\underline{r}_1}^{\underline{r}_2} |\underline{\Omega}\cdot\underline{\nabla} v|^2\, ds \right)^{1/2}.$$

Applying Fubini's theorem, it follows that

$$\begin{aligned} \int_{\mathcal{R}}\int_{\mathcal{S}^1} |v(\underline{r},\underline{\Omega})|^2\, d\Omega dr &\le \operatorname{diam}(\mathcal{R}) \int_{\underline{r}_1}^{\underline{r}_2}\int_{\mathcal{R}}\int_{\mathcal{S}^1} |\underline{\Omega}\cdot\underline{\nabla} v|^2\, d\Omega dr ds \\ &\le \operatorname{diam}(\mathcal{R})^2 \int_{\mathcal{R}}\int_{\mathcal{S}^1} |\underline{\Omega}\cdot\underline{\nabla} v|^2\, d\Omega dr. \end{aligned}$$

Thus, from the relation $v = S\widehat{v}$ and (iv) of Lemma 3.6, we have that

$$\begin{aligned} \|v\| &\le \operatorname{diam}(\mathcal{R})\, \|\underline{\Omega}\cdot\underline{\nabla} v\| = \operatorname{diam}(\mathcal{R})\, \|\underline{\Omega}\cdot\underline{\nabla} S\widehat{v}\| \\ &\le \operatorname{diam}(\mathcal{R}) \left\| \frac{1}{\varepsilon} S\underline{\Omega} S\cdot\underline{\nabla}\widehat{v} \right\| = \operatorname{diam}(\mathcal{R}) \left\| \underline{Q}\cdot\underline{\nabla}\widehat{v} \right\|, \end{aligned}$$

which proves the lemma. ∎

For a proof of the next lemma we refer to [29]. The only difference from Lemma 3.3 is that now $\frac{s}{3}$ is replaced by δ.

LEMMA 3.8. *Suppose* $0 \le \varepsilon \le 1$. *Then, for any* $\delta \in [0,1]$, *there exists* $b \ge 0$ *such that*

$$\widehat{H}(b,\delta) := \frac{1}{2}\left\{ \sqrt{(\delta - (1+\delta)b)^2 + 4\delta(1-\delta)} + (1-\delta)\, b + \delta \right\} < 0.863.$$

In particular, for $\delta < 0.4302$, *we can choose* $b = 0$.

Proof. The proof of this lemma is attained by the same procedure as the proof of Lemma 3.3, which appears in [29]. ∎

We are now ready to prove the $\widehat{V}$-ellipticity of the bilinear form $\widehat{a}(\cdot,\cdot)$.

THEOREM 3.9 (V-ellipticity of $\widehat{a}(\cdot,\cdot)$). *Suppose that* $0 \le \alpha \le 1$, $0 \le \varepsilon \le 1$ *and that the diameter,* $\operatorname{diam}(\mathcal{R})$, *of the domain* $\mathcal{R}$ *is* 1. *Let* $\widehat{a}(\cdot,\cdot)$ *be given as in* (2.11), *where* $\mathcal{L}S$ *is defined in* (2.8). *Let* $\|\cdot\|_{\widehat{V}}$ *be given as in* (3.18). *Then there exists a constant* $C_e \ge 0.137$ *independent of* ε *and* α *such that for all* $\widehat{v} \in \widehat{V}$,

$$\begin{aligned} \widehat{a}(\widehat{v},\widehat{v}) &= \left\| \underline{Q}\cdot\underline{\nabla}\widehat{v} + (I-P)\widehat{v} + \alpha P\widehat{v} \right\|^2 \\ &\ge C_e \left(\left\| \underline{Q}\cdot\underline{\nabla}\widehat{v} \right\|^2 + \|\widehat{v}\|^2 \right) = C_e\, \|\widehat{v}\|_{\widehat{V}}^2 . \end{aligned} \tag{3.22}$$

Proof. The proof begins as in the proof of Theorem 3.4 with $Q\frac{\partial\widehat{v}}{\partial z}$ replaced by $\underline{Q}\cdot\underline{\nabla}\widehat{v}$ and with P defined as in (1.2). Now, however, the stronger Poincaré–Friedrichs inequality (3.21) of Lemma 3.7, together with (i) of Lemma 3.6, yields

$$\left\| \underline{Q}\cdot\underline{\nabla}\widehat{v} \right\|^2 \ge \delta\, \|\widehat{v}\|^2 . \tag{3.23}$$

As in (3.11) we have

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) \quad &\geq \quad \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2 + (1-\delta)\|\widehat{v}\|^2 - 2d\sqrt{\delta}\sqrt{\gamma}\left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|\cdot\|\widehat{v}\| \\ &\quad -2(1-d)\sqrt{(1-\delta)}\sqrt{(1-\gamma)}\left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|\cdot\|\widehat{v}\|.\end{aligned} \tag{3.24}$$

Again, we divide the region $(\delta,\gamma)\in[0,1]\times[0,1]$ into two triangles and choose d as follows:

$$d=\begin{cases} 1 & \text{for} \quad \delta+\gamma\leq 1, \\ 0 & \text{for} \quad \delta+\gamma\geq 1. \end{cases}$$

Next, we consider these two cases separately.

Case 1. $\delta+\gamma\leq 1$: setting $d=1$ in (3.24) yields

$$\widehat{a}(\widehat{v},\widehat{v})\geq(1-\gamma\eta)\ \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2+\left(1-\delta-\frac{\delta}{\eta}\right)\ \|\widehat{v}\|^2. \tag{3.25}$$

If $\delta\leq 0.4302$, we choose η so that

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

which implies

$$\eta\gamma=\frac{\delta+\sqrt{\delta^2+4\delta\gamma}}{2}\leq\frac{\delta+\sqrt{\delta^2+4\delta(1-\delta)}}{2}=\widehat{H}(0,\delta).$$

By inspection, we see that for $\delta\leq 0.4302$ the function $\widehat{H}(0,\delta)$ is an increasing function in δ taking on the value 0.755 for $\delta=0.4302$.

Now assume $\delta\geq 0.4302$ and let $b\geq 0$. Using (3.23) we write (3.25) as

$$\widehat{a}(\widehat{v},\widehat{v})\ \geq\ (1-b-\gamma\eta)\ \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2\ +\ \left(1-\delta+\delta b-\frac{\delta}{\eta}\right)\ \|\widehat{v}\|^2. \tag{3.26}$$

We choose $\eta(b)$ so that

$$(1-b-\gamma\eta)=\left(1-\delta+\delta b-\frac{\delta}{\eta}\right),$$

that is

$$\eta=\frac{-\left(b+\delta b-\delta\right)+\sqrt{\left(b+\delta b-\delta\right)^2+4\delta\gamma}}{2\gamma}, \tag{3.27}$$

and then choose $b\geq 0$ to minimize

$$G_1(b,\,\delta,\,\gamma):=(b+\gamma\eta)=\frac{((1-\delta)b+\delta)+\sqrt{\left(b+\delta b-\delta\right)^2+4\delta\gamma}}{2}.$$

Since $\delta+\gamma\leq 1$, inspection of (3.27) reveals that G_1 is largest when $\gamma=1-\delta$, and we conclude

$$G_1(b,\,\delta,\,\gamma)\leq\widehat{H}(b,\delta)\leq 0.863 \tag{3.28}$$

by Lemma 3.8.

Case 2. $\delta + \gamma \geq 1$: setting $d = 0$ in (3.24) and proceeding as in Case 1 results in

$$\widehat{a}(\widehat{v},\widehat{v}) \geq (1-(1-\gamma)\eta)\ \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2 + \left(1-\delta-\frac{1-\delta}{\eta}\right)\ \|\widehat{v}\|^2. \tag{3.29}$$

As in Case 1, we first assume $\delta \leq 0.4302$, choose η so that

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

and show that

$$(1-\gamma)\eta \leq \widehat{H}(0,\delta) \leq \widehat{H}(0,0.4302) = 0.755.$$

Next, we assume $\delta \geq 0.4302$. Then, for $b \geq 0$, we use (3.23) to obtain

$$\widehat{a}(\widehat{v},\widehat{v})\ \geq\ (1-b-(1-\gamma)\eta)\ \left\|\underline{Q}\cdot\underline{\nabla}\widehat{v}\right\|^2\ +\ \left(1-\delta+\delta b-\frac{1-\delta}{\eta}\right)\ \|\widehat{v}\|^2\,. \tag{3.30}$$

Finally, we choose η to make the coefficients on the right in (3.30) equal and choose b to minimize

$$\begin{aligned} G_2(b,\,\delta,\,\gamma) &:= b+(1-\gamma)\eta \\ &= \frac{1}{2}\left\{\sqrt{(b+\delta b-\delta)^2+4(1-\delta)(1-\gamma)}+(b+\delta-\delta b)\right\}. \end{aligned}$$

By inspection $G_2(b,\,\delta,\,\gamma) \leq G_2(b,\delta,\,(1-\delta)) = \widehat{H}(b,\,\delta)$ for $\delta+\gamma \geq 1$. The theorem now follows from Lemma 3.8. □

From the $\widehat{V}$-ellipticity of the bilinear form $\widehat{a}(\cdot,\cdot)$, the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ follows immediately. We summarize this result in the following corollary.

COROLLARY 3.10 (V-ellipticity of $a(\cdot,\cdot)$). *Suppose that* $0 \leq \alpha \leq 1$, $0 \leq \varepsilon \leq 1$, *and that* $\operatorname{diam}(\mathcal{R}) = 1$. *Let* $a(\cdot,\cdot)$ *be given as in* (2.10), *where for* $\mathcal{L}$ *the definition* (2.5) *is used. Further, let* $\|\cdot\|_V$ *be defined as in* (3.19). *Then there exists a constant* $C_e \geq 0.137$ *independent of* α *and* ε *such that for all* $v \in V$,

$$a(v,v)\ \geq\ C_e\ \|v\|_V^2\,. \tag{3.31}$$

□

Together with the lemma of Lax and Milgram [7] the existence and the uniqueness of a solution for problem (2.10) and its discrete version, where V is replaced by a finite-dimensional subspace $V^h \subset V$, follows directly. In the next section we will use the continuity and the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ to prove discretization error bounds for a variety of discrete spaces V^h.

4. Discretization error bounds. In this section we establish bounds for the discretization error $\psi - \psi_h$. Here $\psi \in V$ denotes the solution of (2.10) and $\psi_h \in V^h \subset V$ denotes the solution of the discrete problem: find $\psi_h \in V^h$ such that

$$a(\psi_h, v_h) = \langle q_s, \mathcal{L}v_h\rangle \qquad \forall v_h \in V^h. \tag{4.1}$$

The continuity and the V-ellipticity of $a(\cdot,\cdot)$ lead directly to Céa's lemma [6]:

$$a(\psi-\psi_h,\psi-\psi_h) \leq a(\psi-v_h,\psi-v_h) \qquad \forall v_h \in V^h \qquad \text{or}$$

$$\|\psi-\psi_h\|_V \leq \sqrt{\frac{C_c}{C_e}}\ \min_{v_h\in V^h}\ \|\psi-v_h\|_V\,. \tag{4.2}$$

Therefore, bounding $\|\psi - \psi_h\|_V$ is reduced to the problem of bounding $\min_{v_h \in V^h} \|\psi - v_h\|_V$, which is a problem of approximation theory and depends on the space V^h. Here, we consider discrete spaces V^h that are formed by functions that can be expanded into the first N Legendre polynomials (spherical harmonics) with respect to the direction angle μ ($\underline{\Omega}$) and are piecewise polynomials of degree k in z ($\underline{r}$) on a partition $\mathcal{T}_h$ of the slab $[z_l, z_r]$ (region $\mathcal{R}$). This class of finite-dimensional subspaces corresponds to a discretization by a spectral method in angle and a finite-element discretization in space.

The spectral discretization in angle with Legendre polynomials (spherical harmonics) is common for transport problems [18] and also called the P_N discretization. Compared to a discrete ordinates or a finite-element (e.g., with piecewise constant basis functions) discretization in angle, the P_N discretization has the advantage of avoiding ray effects [18]. However, it is well known that the treatment of the boundary conditions with this angular discretization is problematic [8]. Most common is the use of the P_N discretization together with Mark or Marshak boundary conditions [9], which are based on rather heuristic arguments.

Here we are mainly interested in diffusive transport problems and in accurate solutions in the interior of the computational domain. Therefore, we avoid the problem of the correct treatment of the boundary conditions by assuming that

$$\psi(\underline{r}, \underline{\Omega}) \equiv 0 \ \text{ for } \ \underline{r} \in \partial\mathcal{R} \quad \text{and} \quad \psi(z_l, \mu) \equiv \psi(z_r, \mu) \equiv 0 \ \text{ for } \ \mu \in [-1, 1] \tag{4.3}$$

in the slab geometry case, respectively. Because of (1.9) this is fulfilled up to an order ε. Moreover, it is shown in [12] and [28] that in diffusive regimes the part of the solution that is influenced by the boundary conditions (the boundary layer solution part) is decaying exponentially with the distance from the boundary. Therefore, the assumption (4.3) is quite reasonable for diffusive regimes, in order to demonstrate that the least-squares discretization in combination with the scaling transformation will yield an accurate solution in the interior. We indicate in section 6 how the boundary conditions can be treated very naturally in the least-squares context in order to extend this solution approach also to nondiffusive regimes.

4.1. Slab geometry. Let $\mathcal{T}_h = \{z_r =: z_0, z_1, \ldots, z_M := z_r\}$ be a partition of the slab $[z_l, z_r]$ with maximum mesh size h and let $\mathbb{P}_k(\mathcal{T}_h)$ denote the space of piecewise polynomials of degree $\leq k$ on the partition $\mathcal{T}_h$. Further, let $P_l(\mu)$ denote the lth Legendre polynomial. The scaled Legendre polynomials $p_l(\mu) := \sqrt{2l+1} P_l(\mu)$ form an orthogonal basis of $L^2([-1, 1])$. Thus, any $\psi \in V$ has the following expansion in angle,

$$\psi(z, \mu) = \sum_{l=0}^{\infty} \phi_l(z)\, p_l(\mu), \tag{4.4}$$

where the Fourier coefficients $\phi_l(z)$, which are called *moments* in transport theory, are given by

$$\phi_l(z) = \frac{1}{2} \int_{-1}^{1} \psi(z, \mu)\, p_l(\mu) d\mu. \tag{4.5}$$

For the discretization we truncate the expansion in (4.4) and approximate the moments $\phi_l(z)$ by piecewise polynomials of degree $\leq k$ on the partition $\mathcal{T}_h$. This

results in the discrete space

$$V^h := \Bigg\{ v_h \in C^0(D);\ v_h = \sum_{l=0}^{N-1} \phi_l^h(z) p_l(\mu);\ \ \phi_l^h(z) \in \mathbb{P}_k(\mathcal{T}_h) \\ \text{for } \ l = 0, \ldots, N-1;\, v_h(z_l, \mu) = 0 \ \text{ for } \mu > 0; \quad v_h(z_r, \mu) = 0 \text{ for } \mu < 0 \Bigg\}. \tag{4.6}$$

For any $\psi \in V \cap \big(H^{\nu+1}([z_l, z_r]) \times L^2([-1,1])\big)$, let $\Pi_z \psi$ denote the interpolant of ψ with respect to z by piecewise polynomials of degree $\le k$ on the partition $\mathcal{T}_h$. For $1 \le \nu \le k$ and $0 \le m \le \nu$ it can be shown [14, p. 91] that

$$\|\psi - \Pi_z \psi\|_{m,0} \le C h^{\nu+1-m}\ |\psi|_{\nu+1,0}\,, \tag{4.7}$$

where

$$\|\psi\|_{m,0}^2 := \sum_{i=0}^{m} \left\| \frac{\partial^i \psi}{\partial z^i} \right\|^2, \qquad |\psi|_{m,0}^2 := \left\| \frac{\partial^m \psi}{\partial z^m} \right\|^2. \tag{4.8}$$

Further, for any $\psi \in V \cap \big(L^2([z_l, z_r]) \times H^2([-1,1])\big)$ we define

$$\Pi_N \psi := \sum_{l=0}^{N-1} \phi_l(z)\, p_l(\mu) \qquad \text{with} \qquad \phi_l(z) := \frac{1}{2} \int_{-1}^{1} \psi(z, \mu)\, p_l(\mu)\, d\mu, \tag{4.9}$$

as the truncated expansion of ψ into the first N normalized Legendre polynomials. Note that by assumption (4.3) we have $\Pi_N \psi \in V$.

By using the fact that the Legendre polynomials are eigenfunctions of the Sturm–Liouville operator [10, p. 37]

$$\mathcal{L}_S p_l(\mu) := \frac{d}{d\mu} \left[\left(1 - \mu^2\right) \frac{dp_l(\mu)}{d\mu} \right] = -l(l+1) p_l(\mu), \tag{4.10}$$

the error $\psi - \Pi_N \psi$ of the truncated expansion into Legendre polynomials can be bounded. In the following lemma we sketch a proof for this bound and establish further a bound for the commutator $[\Pi_N \mathcal{L} - \mathcal{L} \Pi_N]$, which is used later in the bound for the discretization error $\psi - \psi_h$.

LEMMA 4.1. *Suppose $\psi \in V \cap \big(H^{k+1}([z_l, z_r]) \times H^2([-1,1])\big)$ where $N \ge 2$ and $k \ge 1$. Then we have:*

(i) $\left\| \dfrac{d^\nu \phi_l}{dz^\nu} \right\| \le \dfrac{1}{l(l+1)} \left\| \mathcal{L}_S \dfrac{\partial^\nu \psi}{\partial z^\nu} \right\|$ *for* $0 \le \nu \le k+1$,

(ii) $\left\| \frac{\partial^\nu \psi}{\partial z^\nu} - \Pi_N \frac{\partial^\nu \psi}{\partial z^\nu} \right\| \le \frac{C}{N} \left\| \mathcal{L}_S \frac{\partial^\nu \psi}{\partial z^\nu} \right\|$ *for* $0 \le \nu \le k+1$,

(iii) $\left\| [\Pi_N \mathcal{L} - \mathcal{L} \Pi_N] \frac{\partial^\nu \psi}{\partial z^\nu} \right\| \le \frac{C}{N^2} \left\| \mathcal{L}_S \frac{\partial^{\nu+1} \psi}{\partial z^{\nu+1}} \right\|$ *for* $0 \le \nu \le k$,

with C independent of N, ε, α.

Proof.

(i) By (4.5), (4.10), and integration by parts we have

$$\frac{d^\nu \phi_l}{dz^\nu} = \frac{-1}{2l(l+1)} \int_{-1}^{1} \frac{\partial^\nu \psi}{\partial z^\nu} \mathcal{L}_{\mathcal{S}} p_l(\mu)\, d\mu = \frac{-1}{2l(l+1)} \int_{-1}^{1} p_l(\mu)\, \mathcal{L}_{\mathcal{S}} \frac{\partial^\nu \psi}{\partial z^\nu}\, d\mu;$$

thus,

$$\left| \frac{d^\nu \phi_l}{dz^\nu} \right| \leq \frac{1}{\sqrt{2} l(l+1)} \left\| \mathcal{L}_{\mathcal{S}} \frac{\partial^\nu \psi}{\partial z^\nu} \right\|_{L^2([-1,1])}, \tag{4.11}$$

which, after squaring and integrating over D, leads directly to (i).

(ii) Since the normalized Legendre polynomials form an orthogonal basis of $L^2([-1,1])$, we obtain from (4.11):

$$\begin{aligned} \left\| \frac{\partial^\nu \psi}{\partial z^\nu} - \Pi_N \frac{\partial^\nu \psi}{\partial z^\nu} \right\|^2_{L^2([-1,1])} &= 2 \sum_{l=N}^{\infty} \left| \frac{d^\nu \phi_l}{dz^\nu} \right|^2 \\ &\leq \left\| \mathcal{L}_{\mathcal{S}} \frac{\partial^\nu \psi}{\partial z^\nu} \right\|^2_{L^2([-1,1])} \sum_{l=N}^{\infty} \frac{1}{[l(l+1)]^2}. \end{aligned} \tag{4.12}$$

For $l \geq 1$ we have $\frac{1}{l} \leq \frac{2}{l+1}$, so that the sum can be bounded by

$$\sum_{l=N}^{\infty} \frac{1}{[l(l+1)]^2} \leq 4 \sum_{l=N+1}^{\infty} \frac{1}{l^4} \leq 4 \int_{N}^{\infty} \frac{1}{l^4}\, dl = \frac{4}{3N^3} \leq \frac{4}{3N^2}.$$

Integrating (4.12) over $[z_l, z_r]$ yields,

$$\left\| \frac{\partial^\nu \psi}{\partial z^\nu} - \Pi_N \frac{\partial^\nu \psi}{\partial z^\nu} \right\|^2 \leq \frac{4}{3N^2} \left\| \mathcal{L}_S \frac{\partial^\nu \psi}{\partial z^\nu} \right\|^2.$$

(iii) Since $\Pi_N P \frac{\partial^\nu \psi}{\partial z^\nu} = \frac{d^\nu \phi_0}{dz^\nu} = P \Pi_N \frac{\partial^\nu \psi}{\partial z^\nu}$ and $\Pi_N P \mu \frac{\partial^\nu \psi}{\partial z^\nu} = \frac{1}{\sqrt{3}} \frac{d^\nu \phi_1}{dz^\nu} = P \mu \Pi_N \frac{\partial^\nu \psi}{\partial z^\nu}$ we see directly that

$$[\Pi_N \mathcal{L} - \mathcal{L} \Pi_N] = \Pi_N \mu \frac{\partial}{\partial z} - \mu \Pi_N \frac{\partial}{\partial z}. \tag{4.13}$$

Using the moment expansion (4.4) and the following recurrence relation for the Legendre polynomials [3, p. 540]

$$\mu\, p_l(\mu) = b_{l-1}\, p_{l-1}(\mu) + b_l\, p_{l+1}(\mu),$$

with $b_l := \frac{l+1}{\sqrt{4(l+1)^2 - 1}}$ and $p_{-1}(\mu) \equiv 0$, we obtain from (4.13)

$$[\Pi_N \mathcal{L} - \mathcal{L} \Pi_N] \frac{\partial^\nu \psi}{\partial z^\nu} = b_{N-1} \left(\frac{d^{\nu+1} \phi_N}{dz^{\nu+1}}\, p_{N-1} - \frac{d^{\nu+1} \phi_{N-1}}{dz^{\nu+1}}\, p_N \right). \tag{4.14}$$

Now we notice that for any integers $k, l \geq 0$,

$$\left\| \frac{d^{\nu+1} \phi_l}{dz^{\nu+1}}\, p_k \right\|^2 = \left\| \frac{d^{\nu+1} \phi_l}{dz^{\nu+1}} \right\|^2 \leq \frac{1}{[l(l+1)]^2} \left\| \mathcal{L}_{\mathcal{S}} \frac{\partial^{\nu+1} \psi}{\partial z^{\nu+1}} \right\|^2,$$

where the last inequality follows from (i). Since $b_l \leq \frac{1}{\sqrt{3}}$, we can bound (4.14) as follows:

$$\left\| [\Pi_N \mathcal{L} - \mathcal{L}\Pi_N] \frac{\partial^\nu \psi}{\partial z^\nu} \right\| \leq \frac{1}{\sqrt{3}} \frac{8}{3} \frac{1}{N^2} \left\| \mathcal{L}_S \frac{\partial^{\nu+1}\psi}{\partial z^{\nu+1}} \right\|,$$

since $\frac{2}{N^2-1} \leq \frac{8}{3}\frac{1}{N^2} \Leftrightarrow \frac{2}{1-\frac{1}{N^2}} \leq \frac{8}{3}$, which is valid for $N \geq 2$. □

We are now in a position to bound the discretization error $\psi - \psi_h$ in the following theorem.

THEOREM 4.2 (discretization error bound for slab geometry). *Suppose* $0 \leq \alpha \leq 1$, $0 \leq \varepsilon \leq \frac{1}{\sqrt{3}}$, *and* $1 \leq \nu \leq k$. *Let* $\psi \in V \cap \left(H^{\nu+1}([z_l, z_r]) \times H^2([-1,1])\right)$ *be the solution of* (2.10) *with right-hand side* $q_s \in H^\nu([z_l, z_r]) \times H^2([-1,1])$. *Further, let* $\psi_h \in V^h$ *be the solution of* (4.1) *with* V^h *defined as in* (4.6). *Then we have*

$$\begin{aligned} \|\psi - \psi_h\|_V \leq \quad & \left\{ \frac{1}{\sqrt{C_e}} \left(\frac{C_1}{N} \|\mathcal{L}_S q_s\| + \frac{C_2}{N^2} \left\| \mathcal{L}_S \frac{\partial \psi}{\partial z} \right\| \right) \right. \\ & \left. + \frac{1}{C_e} \left(C_3 h^\nu \left\| \frac{\partial^\nu q_s}{\partial z^\nu} \right\| + C_4 \frac{h^\nu}{N^2} \left\| \mathcal{L}_S \frac{\partial^{\nu+1}\psi}{\partial z^{\nu+1}} \right\| \right) \right\} =: e_h, \end{aligned} \tag{4.15}$$

with C_1, C_2, C_3, C_4 *independent of* α *and* ε. *In particular,*

$$\begin{aligned} \left\| P\left(\mu \frac{\partial(\psi - \psi_h)}{\partial z} \right) \right\| &\leq \varepsilon e_h, \qquad \|(I-P)(\psi - \psi_h)\| \leq \varepsilon e_h, \\ \left\| (I-P)\left(\mu \frac{\partial(\psi - \psi_h)}{\partial z} \right) \right\| &\leq e_h, \qquad \|P(\psi - \psi_h)\| \leq e_h. \end{aligned}$$

Proof. Using the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ from Corollary 3.5 and Céa's lemma (4.2) we conclude that

$$\begin{aligned} \|\psi - \psi_h\|_V &\leq \frac{1}{\sqrt{C_e}} \sqrt{a(\psi - \psi_h, \psi - \psi_h)} \leq \frac{1}{\sqrt{C_e}} \min_{v_h \in V^h} \sqrt{a(\psi - v_h, \psi - v_h)} \\ &\leq \frac{1}{\sqrt{C_e}} \sqrt{a(\psi - \Pi_z\Pi_N\psi, \psi - \Pi_z\Pi_N\psi)} = \frac{1}{\sqrt{C_e}} \|\mathcal{L}(\psi - \Pi_z\Pi_N\psi)\| \\ &\leq \frac{1}{\sqrt{C_e}} \left\{ \|\mathcal{L}\psi - \mathcal{L}\Pi_N\psi\| + \|\mathcal{L}\Pi_N\psi - \mathcal{L}\Pi_z\Pi_N\psi\| \right\}, \end{aligned} \tag{4.16}$$

since $\Pi_z\Pi_N\psi \in V^h$.

In order to bound the first term in the last line of (4.16), we use (ii) and (iii) of Lemma 4.1 to get

$$\begin{aligned} \|\mathcal{L}\psi - \mathcal{L}\Pi_N\psi\| &\leq \|\mathcal{L}\psi - \Pi_N\mathcal{L}\psi\| + \|\Pi_N\mathcal{L}\psi - \mathcal{L}\Pi_N\psi\| \\ &\leq \|q_s - \Pi_N q_s\| + \frac{C_2}{N^2} \left\| \mathcal{L}_S \frac{\partial \psi}{\partial z} \right\| \\ &\leq \frac{C_1}{N} \|\mathcal{L}_S q_s\| + \frac{C_2}{N^2} \left\| \mathcal{L}_S \frac{\partial \psi}{\partial z} \right\|. \end{aligned} \tag{4.17}$$

To bound the second term in the last line of (4.16), observe that $P\Pi_z\psi = \Pi_z P\psi$ for any $\psi \in V$, since P operates only on μ. Therefore, denoting $T := \frac{1}{\varepsilon}(I-P)+\alpha P$, we have

$$\begin{aligned}\|\mathcal{L}\Pi_N\psi - \mathcal{L}\Pi_z\Pi_N\psi\| &\leq \left\| \frac{\partial}{\partial z}\left(\frac{1}{\varepsilon}S\mu\Pi_N\psi - \frac{1}{\varepsilon}S\mu\Pi_z\Pi_N\psi\right)\right\| + \|T\Pi_N\psi - T\Pi_z\Pi_N\psi\| \\ &= \left\| \frac{\partial}{\partial z}\left(\frac{1}{\varepsilon}S\mu\Pi_N\psi - \Pi_z\frac{1}{\varepsilon}S\mu\Pi_N\psi\right)\right\| + \|T\Pi_N\psi - \Pi_z T\Pi_N\psi\|\,.\end{aligned}$$

Applying (4.7) to bound the last two terms leads to

$$\begin{aligned}\|\mathcal{L}\Pi_N\psi - \mathcal{L}\Pi_z\Pi_N\psi\| &\leq Ch^\nu \left\| \frac{\partial^{\nu+1}}{\partial z^{\nu+1}}\frac{1}{\varepsilon}S\mu\Pi_N\psi\right\| + Ch^\nu \left\| \frac{\partial^\nu}{\partial z^\nu}T\Pi_N\psi\right\| \\ &\leq Ch^\nu \left\| \Pi_N\frac{\partial^\nu\psi}{\partial z^\nu}\right\|_V \leq \frac{C}{\sqrt{C_e}}h^\nu \left\| \mathcal{L}\Pi_N\frac{\partial^\nu\psi}{\partial z^\nu}\right\|,\end{aligned}$$

where we used the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ in the last step. We proceed further by using (iii) of Lemma 4.1 and the fact that Π_N is an orthogonal projection to get

$$\begin{aligned}\|\mathcal{L}\Pi_N\psi - \mathcal{L}\Pi_z\Pi_N\psi\| &\leq \frac{C}{\sqrt{C_e}}h^\nu\left(\left\|\Pi_N\mathcal{L}\frac{\partial^\nu\psi}{\partial z^\nu}\right\| + \left\|[\mathcal{L}\Pi_N - \Pi_N\mathcal{L}]\,\frac{\partial^\nu\psi}{\partial z^\nu}\right\|\right) \\ &\leq \frac{C_3}{\sqrt{C_e}}h^\nu\left\|\frac{\partial^\nu q_s}{\partial z^\nu}\right\| + \frac{C_4}{\sqrt{C_e}}\frac{h^\nu}{N^2}\left\|\mathcal{L}_S\frac{\partial^{\nu+1}\psi}{\partial z^{\nu+1}}\right\|.\end{aligned} \tag{4.18}$$

Finally, substituting (4.17) and (4.18) into (4.16) results in (4.15). ∎

Remark 4.3 (interpretation of error bound). For diffusive regimes ($\varepsilon \ll 1$) the exact solution has the diffusion expansion (1.7). Therefore, it follows that $\mathcal{L}_S\psi = O(\varepsilon)$ and $\mathcal{L}_S q_s = O(\varepsilon)$. Taking this into account, we get from (4.15) that $e_h = O(h^\nu)+O(\varepsilon)$. Thus, the error in the zeroth moment $\|P(\psi-\psi_h)\|$ is bounded by $O(h^\nu)+O(\varepsilon)$ and the error in the higher moments $\|(I-P)(\psi-\psi_h)\|$ is bounded by $O(\varepsilon h^\nu)+O(\varepsilon^2)$. In particular, for diffusive regimes, where ε is very small, convergence of the discrete solution is also assured by the above bound for small N, which is a reasonable choice in this case, since the exact solution is nearly independent of μ.

On the other hand, if ε is close to $\frac{1}{\sqrt{3}}$, so that $\varepsilon > h^\nu$ and $\varepsilon > \frac{1}{N}$, an error bound can be obtained more easily, since $\|v\|_V^2 \leq \left(1+\frac{1}{\varepsilon^2}\right)\|v\|_{1,0}^2$. Therefore, for any $\psi \in V \cap (H^{\nu+1}([z_l,z_r]) \times H^2([-1,1]))$, Céa's lemma in combination with the bounds for the interpolation error lead directly to

$$\|\psi-\psi_h\|_V \leq \sqrt{\frac{C_c}{C_e}}\,\left(1+\frac{1}{\varepsilon^2}\right)^{1/2}\,\left(\frac{C_1}{N}\|\mathcal{L}_S\psi\|_{1,0} + C_2h^\nu\left\|\frac{\partial^\nu\psi}{\partial z^\nu}\right\|_{1,0}\right).$$

However, we point out that this bound is not useful in the diffusion limit $\varepsilon \to 0$ for fixed N and h. ∎

Remark 4.4 (V-ellipticity constant C_e). The error bound (4.15) depends on the reciprocal of the V-ellipticity constant C_e. According to Theorem 3.4 and Corollary 3.5, $C_e = 0.012$, so $1/C_e = 83.3$, which is fairly large. However, we would like to

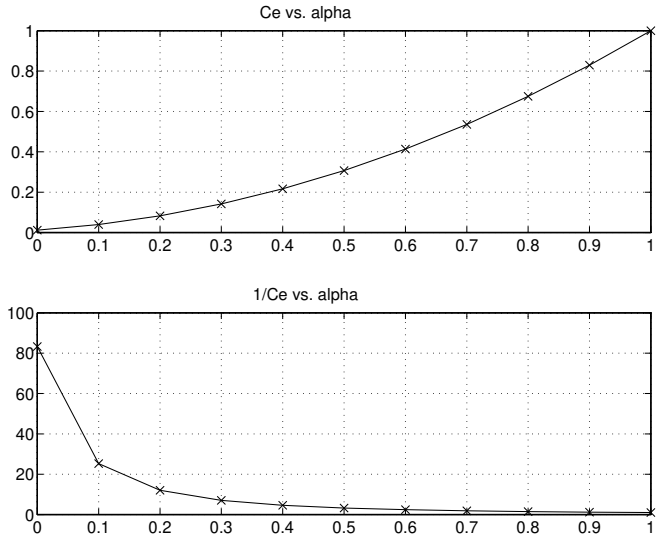

FIG. 4.1. *V-ellipticity constant C_e and its reciprocal versus the absorption parameter α.*

point out that we simplified the proof of Theorem 3.4 by considering only the worst case, $\alpha = 0$. Without setting $\alpha = 0$, (3.11) would change to

$$\begin{aligned}\widehat{a}(\widehat{v},\widehat{v}) \geq \quad & \left\| Q\frac{\partial \widehat{v}}{\partial z}\right\|^2 + (1+\alpha^2\delta-\delta)\,\|\widehat{v}\|^2 - 2d(1-\alpha)\sqrt{\delta}\sqrt{\gamma}\left\| Q\frac{\partial \widehat{v}}{\partial z}\right\|\|\widehat{v}\| \\ & -2(1-d)(1-\alpha)\sqrt{(1-\delta)}\sqrt{(1-\gamma)}\left\| Q\frac{\partial \widehat{v}}{\partial z}\right\|\|\widehat{v}\|\,,\end{aligned} \tag{4.19}$$

which clearly shows that the V-ellipticity constant C_e increases with α.

To judge the quantitative behavior, we computed C_e for certain values of α using (4.19). The results are plotted in Figure 4.1. Already for $\alpha = 0.3$, $1/C_e$ drops down to 7.04. □

4.2. Three-dimensional geometry. In this section we extend the error bounds of the previous section to three-dimensional geometry. Let $\mathcal{T}_h$ be a triangulation of $\mathcal{R}$ into tetrahedrons or hexahedrons of maximum diameter h. Further, for any $v(\underline{r},\underline{\Omega}) \in H^{k+1}(\mathcal{R}) \times L^2(\mathcal{S}^1)$, let $\Pi_h v$ denote the interpolant of v by piecewise polynomials of degree $\leq k$ on the triangulation $\mathcal{T}_h$. Then it can be shown [7] that

$$\|v - \Pi_h v\|_{1,0} \leq Ch^k\,|v|_{k+1,0}\,, \tag{4.20}$$

where $\|\cdot\|_{\nu,0}$ ($|\cdot|_{\nu,0}$) denotes the standard norm (semi norm) of $H^\nu(\mathcal{R}) \times L^2(\mathcal{S}^1)$ [1].

Further, we recall that the *spherical harmonics* [3, p. 571] are defined by

$$Y_l^m(\theta,\varphi) := (-1)^m\, C_{l,m}\, P_l^m(\cos(\theta))\, e^{im\varphi},$$

where $C_{l,m}$ is a constant, depending on l and m, $P_l^m(\mu)$ denotes the associated Legendre polynomials, and θ denotes the polar angle with respect to the z-axis, while φ denotes the azimuthal angle about the z-axis. The spherical harmonics form an

orthonormal basis of $L^2(\mathcal{S}^1)$; therefore, any $v \in L^2(\mathcal{R} \times \mathcal{S}^1)$ has an expansion of the form

$$v(\underline{r}, \underline{\Omega}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \phi_{l,m}(\underline{r}) Y_l^m(\underline{\Omega}), \text{ with } \phi_{l,m}(\underline{r}) = \int_{\mathcal{S}^1} v(\underline{r}, \underline{\Omega}) Y_l^{m*}(\underline{\Omega})\, d\Omega. \tag{4.21}$$

Similar to the slab geometry case, we truncate this expansion for the discretization and approximate the moments $\phi_{l,m}$ by a polynomial $\phi_{l,m}^h \in \mathbb{P}_k(\mathcal{T}_h)$, where $\mathbb{P}_k(\mathcal{T}_h)$ is the space of piecewise polynomials of degree $\leq k$ on the triangulation $\mathcal{T}_h$. This truncation leads to the following class of discrete spaces:

$$V^h := \left\{ v_h \in V : v_h(\underline{r}, \underline{\Omega}) = \sum_{l=0}^{N-1} \sum_{m=-l}^{l} \phi_{l,m}^h(\underline{r})\, Y_l^m(\underline{\Omega}), \quad \phi_{l,m}^h(\underline{r}) \in \mathbb{P}_k(\mathcal{T}_h) \right\}, \tag{4.22}$$

which corresponds to a finite-element discretization in space and a P_N discretization [18] in angle.

As in the slab geometry case, we denote by

$$\Pi_N v(\underline{r}, \underline{\Omega}) := \sum_{l=0}^{N-1} \sum_{m=-l}^{l} \phi_{l,m}(\underline{r}) Y_l^m(\underline{\Omega}), \tag{4.23}$$

with $\phi_{l,m}(\underline{r})$ defined as in (4.21), the truncated expansion of v into spherical harmonics. To bound the error of the truncated expansion, we use the fact that the spherical harmonics are the eigenfunctions of the Laplacian operator on the unit sphere, so

$$\begin{aligned} \Delta_\Omega Y_l^m(\underline{\Omega}) &:= \left[\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \left(\sin\theta \frac{\partial}{\partial\theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial\varphi^2} \right] Y_l^m(\underline{\Omega}) \\ &= -l(l+1) Y_l^m(\underline{\Omega}) \end{aligned}$$

for $l \geq 0$ and $m = -l, -l+1, \ldots, 0, \ldots, l$. This is summarized together with other inequalities that are needed for the bound of the discretization error in the following lemma. Since the proof of this lemma is analogous to the slab geometry case and involves only more tedious calculus, we refer the reader to Lemmas 3.7, 3.8, and 3.9 in [29] for the proof.

LEMMA 4.5 (truncated expansion into spherical harmonics). *Suppose $N \geq 2$. Let β be any multi-index and recall that $D^\beta v := \frac{\partial^{|\beta|} v}{\partial x^{\beta_1} \partial y^{\beta_2} \partial z^{\beta_3}}$. Suppose that $v(\underline{r}, \underline{\Omega}) \in V \cap \left(H^{|\beta|}(\mathcal{R}) \times H^2(\mathcal{S}^1) \right)$. Then:*

(i) $\left\| D^\beta \phi_{l,m} \right\| \leq \dfrac{1}{l(l+1)} \left\| \Delta_\Omega D^\beta v \right\|$ *for* $l \geq 0$; $-l \leq m \leq l$;

(ii) $\|v - \Pi_N v\| \leq \dfrac{C}{N} \|\Delta_\Omega v\|$;

(iii) $\|[\Pi_N \mathcal{L} - \mathcal{L}\Pi_N]\, v\| \leq \dfrac{C}{N} |\Delta_\Omega v|_{1,0}$ *with C independent of ε and α;*

(iv) $\|\widehat{v}\|_{\widehat{V}} \leq C \|\widehat{v}\|_{1,0}$ *with C independent of ε and α.* □

We can now establish the following discretization error bound.

THEOREM 4.6 (discretization error bound for three-dimensional geometry). *Suppose $0 \leq \alpha \leq 1$ and $0 \leq \varepsilon \leq 1$. Let $\psi \in V \cap \left(H^{k+1}(\mathcal{R}) \times H^2(\mathcal{S}^1) \right)$ be the solution of* (2.10) *with $q_s \in H^k(\mathcal{R}) \times H^2(\mathcal{S}^1)$. Further, let $\psi_h \in V^h$ be the solution of* (4.1)

with V^h defined as in (4.22)*. In addition, assume that ψ has the diffusion expansion $\psi(\underline{r},\underline{\Omega}) = \phi_0(\underline{r}) + \varepsilon\phi_R(\underline{r},\underline{\Omega})$ (compare* (1.7)*). Then we have*

$$\begin{aligned}\|\psi-\psi_h\|_V \leq\ & \frac{1}{\sqrt{C_e}}\frac{C_1}{N}\left(\|\Delta_\Omega q_s\| + |\Delta_\Omega\psi|_{1,0}\right)\\ &+\sqrt{\frac{C_c}{C_e}}\left\{C_2h^k\left(|\phi_0|_{k+1,0}+|\phi_R|_{k+1,0}\right)\right\} =: e_h,\end{aligned} \tag{4.24}$$

with C_1, C_2 independent of ε and α. In particular,

$$\begin{aligned}\|P\,(\Omega\cdot\nabla(\psi-\psi_h))\| &\leq \varepsilon e_h, \qquad \|(I-P)(\psi-\psi_h)\| \leq \varepsilon e_h,\\ \|(I-P)\,(\Omega\cdot\nabla(\psi-\psi_h))\| &\leq e_h, \qquad \|P(\psi-\psi_h)\| \leq e_h.\end{aligned}$$

Proof. From the V-ellipticity of $a(\cdot,\cdot) = \|\mathcal{L}\cdot\|^2$ and Céa's lemma (4.2), we conclude

$$\begin{aligned}\|\psi-\psi_h\|_V &\leq \frac{1}{\sqrt{C_e}}\|\mathcal{L}\,(\psi-\Pi_h\Pi_N\psi)\|\\ &\leq \frac{1}{\sqrt{C_e}}\left(\|\mathcal{L}\,(\psi-\Pi_N\psi)\| + \|\mathcal{L}\,(\Pi_N\psi-\Pi_h\Pi_N\psi)\|\right).\end{aligned} \tag{4.25}$$

The first term in (4.25) can be bounded by using (ii) and (iii) of Lemma 4.5 in the following way:

$$\begin{aligned}\|\mathcal{L}(\psi-\Pi_N\psi)\| &\leq \left(\|\mathcal{L}\psi-\Pi_N\mathcal{L}\psi\| + \|[\Pi_N\mathcal{L}-\mathcal{L}\Pi_N]\,\psi\|\right)\\ &\leq \left(\frac{C}{N}\,\|\Delta_\Omega q_s\| + \frac{C}{N}|\Delta_\Omega\psi|_{1,0}\right).\end{aligned} \tag{4.26}$$

To bound the second term in (4.25), recall that, by definition, $\|v\|_V = \|\widehat{v}\|_{\widehat{V}} = \left\|S^{-1}v\right\|_{\widehat{V}}$ and $S^{-1}\Pi_h\Pi_N\psi = \Pi_hS^{-1}\Pi_N\psi$. Therefore, together with the continuity of $a(\cdot,\cdot)$ and (iv) of Lemma 4.5, we get

$$\begin{aligned}&\|\mathcal{L}\,(\Pi_N\psi-\Pi_h\Pi_N\psi)\|\\ &\qquad\leq \sqrt{C_c}\,\|\Pi_N\psi-\Pi_h\Pi_N\psi\|_V = \sqrt{C_c}\,\left\|S^{-1}\Pi_N\psi-\Pi_hS^{-1}\Pi_N\psi\right\|_{\widehat{V}}\\ &\qquad\leq \sqrt{C_c}C\left\|S^{-1}\Pi_N\psi-\Pi_hS^{-1}\Pi_N\psi\right\|_{1,0} \leq \sqrt{C_c}Ch^k\left|S^{-1}\Pi_N\psi\right|_{k+1,0},\end{aligned} \tag{4.27}$$

where the last inequality follows from (4.20). Since $\Pi_NP = P\Pi_N$ and $\|\Pi_N\| \leq 1$, we can now exploit the diffusion expansion of ψ to get

$$\begin{aligned}\left|S^{-1}\Pi_N\psi\right|_{k+1,0} &= |P\Pi_N\psi|_{k+1,0} + \left|\frac{1}{\varepsilon}(I-P)\Pi_N\psi\right|_{k+1,0}\\ &\leq |\phi_0|_{k+1,0} + |\phi_R|_{k+1,0}\,.\end{aligned} \tag{4.28}$$

Finally, combining (4.25), (4.26), (4.27), and (4.28) leads to the bound (4.24).

The last set of inequalities follows from the definition of the norm $\|\cdot\|_V$ in (3.19) and the definition of S in (2.4). ∎

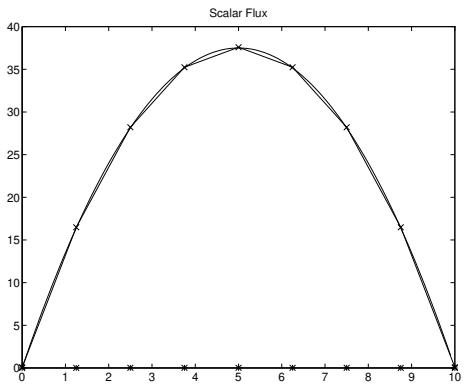

FIG. 5.1. *Scalar flux of exact (solid) and least-squares solution with scaling transformation (crosses) and without (asterisks) for problem* (5.1).

5. Numerical results. In this section we confirm the theoretical error bounds of the previous section by numerical results and demonstrate that the resulting discrete system can be solved very efficiently by a full multigrid solver. The following tests are all based on problems in slab geometry. For the discretization in space we use linear basis functions, whereas the angle dependence is discretized by a P_N approximation. We first validate the accuracy of this discretization and describe the full multigrid solver later.

5.1. Accuracy of the least-squares discretization. To test the accuracy of the least-squares finite-element discretization in diffusive regimes we chose the problem used by Larsen, Morel, and Miller in [16], which is given by

$$\left\{ \begin{array}{rcll} \mu\dfrac{\partial\psi}{\partial z} + 100(I-P)\psi = & 0.01 & \text{for } (z,\mu) \in [0,10]\times[-1,\,1] \\ \psi(0,\mu) = & 0 & \text{for } \mu > 0 \\ \psi(10,\mu) = & 0 & \text{for } \mu < 0 \end{array} \right\}. \tag{5.1}$$

In our notation this means $\varepsilon = 0.01$ and $\alpha = 0.0$. The scalar flux $\phi_0 = P\psi$ of the discrete solution for this problem, resulting from a least-squares P_4 discretization in combination with the scaling transformation, is shown in Figure 5.1 by the crosses. The solid line in Figure 5.1 shows the exact solution of the corresponding diffusion equation, which is $\phi(z) = -3/2z^2 + 15z$. Taking into account that the mesh size of 1.25 is two orders of magnitude larger than ε, we see that this result is very accurate. Moreover, we would like to mention that the least-squares discretization using piecewise linear elements without the scaling transformation results in the zero solution, indicated by the asterisks in Figure 5.1. An explanation of this result is given in [22] and by a different analysis also in [29].

For piecewise linear elements ($k = 1$) in space, the error bound in Theorem 4.2 indicates an $O(h)$ behavior of the least-squares discretization error measured in the V-norm (3.4) for a sufficiently smooth solution. To analyze the order of the least-squares discretization numerically, we used the problem

$$\left\{ \begin{array}{rcll} \left[\mu\frac{\partial}{\partial z} + \frac{1}{\varepsilon}(I-P) + \varepsilon\alpha P\right]\psi(z,\mu) & = q & \text{for } (z,\mu) \in [0,1]\times[-1,1] \\ \psi(0,\mu) & = 0 & \text{for } \mu > 0 \\ \psi(1,\mu) & = 0 & \text{for } \mu < 0 \end{array} \right\}, \tag{5.2}$$

TABLE 5.1
Order of least-squares discretization for linear elements in space.

	$\varepsilon = 1.0$				$\varepsilon = 0.001$			
	$\alpha = 0.0$		$\alpha = 1.0$		$\alpha = 0.0$		$\alpha = 1.0$	
$\frac{1}{h}$	$\|e_h\|_2$	$\frac{\|e_{2h}\|_2}{\|e_h\|_2}$	$\|e_h\|_2$	$\frac{\|e_{2h}\|_2}{\|e_h\|_2}$	$\|e_h\|_2$	$\frac{\|e_{2h}\|_2}{\|e_h\|_2}$	$\|e_h\|_2$	$\frac{\|e_{2h}\|_2}{\|e_h\|_2}$
4	$8.5{\cdot}10^{-2}$		$2.6{\cdot}10^{-2}$		$1.5{\cdot}10^{-2}$		$1.2{\cdot}10^{-2}$	
8	$2.1{\cdot}10^{-2}$	4.0	$6.7{\cdot}10^{-3}$	3.9	$3.8{\cdot}10^{-3}$	3.9	$2.9{\cdot}10^{-3}$	4.1
16	$5.4{\cdot}10^{-3}$	3.9	$1.7{\cdot}10^{-3}$	3.9	$9.7{\cdot}10^{-4}$	3.9	$7.5{\cdot}10^{-4}$	3.9
32	$1.3{\cdot}10^{-3}$	4.1	$4.2{\cdot}10^{-4}$	4.0	$2.4{\cdot}10^{-4}$	4.0	$1.8{\cdot}10^{-4}$	4.1
64	$3.4{\cdot}10^{-4}$	3.8	$1.0{\cdot}10^{-4}$	4.2	$6.1{\cdot}10^{-5}$	3.9	$4.7{\cdot}10^{-5}$	3.8
128	$8.5{\cdot}10^{-5}$	4.0	$2.6{\cdot}10^{-5}$	3.8	$1.5{\cdot}10^{-5}$	4.0	$1.2{\cdot}10^{-5}$	3.9
256	$2.1{\cdot}10^{-5}$	4.0	$6.6{\cdot}10^{-6}$	3.9	$3.8{\cdot}10^{-6}$	3.9	$2.9{\cdot}10^{-6}$	4.1
512	$5.3{\cdot}10^{-6}$	3.9	$1.6{\cdot}10^{-6}$	4.1	$9.5{\cdot}10^{-7}$	4.0	$7.3{\cdot}10^{-7}$	3.9
1024	$1.3{\cdot}10^{-6}$	4.0	$4.1{\cdot}10^{-7}$	3.9	$2.4{\cdot}10^{-7}$	3.9	$1.8{\cdot}10^{-7}$	4.0
2048	$3.3{\cdot}10^{-7}$	3.9	$1.0{\cdot}10^{-7}$	4.1	$6.0{\cdot}10^{-8}$	4.0	$4.6{\cdot}10^{-8}$	3.9
4096	$8.2{\cdot}10^{-8}$	4.0	$2.5{\cdot}10^{-8}$	4.0	$1.4{\cdot}10^{-8}$	4.2	$1.1{\cdot}10^{-8}$	4.2
8192	$2{,}0{\cdot}10^{-8}$	4.1	$6.5{\cdot}10^{-9}$	3.8	$3.5{\cdot}10^{-9}$	4.0	$2.8{\cdot}10^{-9}$	3.9

with $q := \mu\pi\cos(\pi z) + \varepsilon\alpha\sin(\pi z)$, which has the known exact solution $\sin(\pi z)$. We then computed the discrete L^2-error of the least-squares discretization with linear elements in space for a sequence of grids that were created from the coarsest grid by halving the mesh size from one to another grid. Table 5.1 depicts the ratio of these errors for each of two consecutive grids measured in the L^2 norm. The value of approximately 4 of this quotient indicates an $O(h^2)$ behavior of the discretization error for linear elements. On the other hand, our theory predicts $O(h)$ behavior in the V norm, which is similar to an H^1 norm. To formally prove the observed L^2 norm behavior would require a technique like the famous Nitsche trick for second-order elliptic problems.

5.2. Multigrid solver. In this section we describe the multigrid solvers for solving the discrete problem resulting from a least-squares discretization. We refer the reader who is not familiar with multigrid methods to [5] for an introduction and to [13] and [25] for more advanced topics.

The proper choice of the components, namely, the intergrid transfer operators, coarse grid problems, and relaxation schemes, is essential for the efficiency of a multigrid solver. The choice of the first two components is naturally given by the least-squares variational formulation. The sequence of discrete spaces $V_1 \subset V_2 \subset \cdots \subset V_l = V^h$ determines the coarse grid problems since they are just the restriction of the variational problem to these discrete subspaces. For the discrete subspaces, we use finite-element spaces with linear basis elements on increasingly finer partitions (halving the spatial cells) of the slab.

The prolongation operator, which is a mapping from a coarse grid to the next finer grid in the grid sequence, is formed directly by composing the isomorphisms between the discrete spaces and their corresponding coordinate spaces with the injection mapping between V_{k-1} and V_k [4], [27]. To be more precise, let us assume that $V_k = \operatorname{span}\{\phi_i\}$, where ϕ_i denotes a standard linear nodal basis function on an irregular grid $G_h := \{z_0, z_i = z_{i-1} + h_i, \text{ for } i = 1, \ldots, M\}$. Similarly, let $V_{k-1} = \operatorname{span}\{\widehat{\phi}_i\}$ be the discrete space corresponding to the next coarser grid $G_{2h} = \{z_i \in G_h : i \text{ even }\}$. Let $\underline{\Psi}_{2h}$ be a given coarse grid function on G_{2h}, which is the coordinate vector of the

TABLE 5.2
Multigrid convergence factors.

$V(1,1)$-cycle					
σ_t	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.25$	$\alpha = 0.1$	$\alpha = 0.0$
10^0	0.052	0.086	0.083	0.118	0.169
10^1	0.091	0.092	0.091	0.117	0.136
10^2	0.056	0.056	0.071	0.106	0.131
10^3	0.092	0.093	0.092	0.105	0.127
10^4	0.095	0.094	0.094	0.106	0.129
10^5	0.095	0.094	0.093	0.107	0.130
10^6	0.095	0.092	0.092	0.107	0.130
10^7	0.095	0.092	0.092	0.107	0.130
10^8	0.095	0.092	0.092	0.107	0.130
10^9	0.095	0.094	0.092	0.107	0.130
10^{10}	0.095	0.094	0.092	0.106	0.130

function

$$\psi_{2h} := \sum_{z_i \in G_{2h}} \underline{\Psi}_{2h}(z_i)\, \widehat{\phi}_i \quad \in V_{k-1}.$$

Since $V_{k-1} \subset V_k$, ψ_{2h} has also the expansion $\psi_{2h} := \sum_{z_i \in G_h} \underline{\Psi}_h(z_i)\, \phi_i$, which defines a fine grid function $\underline{\Psi}_h$ on G_h. The prolongation operator I_{2h}^h can then be defined by $I_{2h}^h \underline{\Psi}_{2h} := \underline{\Psi}_h$. By expressing the coarse grid basis functions $\widehat{\phi}_i$ in terms of the fine grid basis functions ϕ_i we obtain:

$$\left(I_{2h}^h \underline{\Psi}_{2h}\right)(z_i) = \begin{cases} \underline{\Psi}_{2h}(z_i) & \text{if } z_i \in G_{2h}, \\ \frac{h_{i+1}}{h_i + h_{i+1}} \underline{\Psi}_{2h}(z_{i-1}) + \frac{h_i}{h_i + h_{i+1}} \underline{\Psi}_{2h}(z_{i+1}) & \text{if } z_i \in G_h \backslash G_{2h}. \end{cases}$$

The restriction operator, which is a mapping from a finer grid to the next coarser grid, is just the adjoint of the prolongation operator. For $z_i \in G_{2h}$ this yields

$$\underline{\Psi}_{2h}(z_i) = \left(I_h^{2h} \underline{\Psi}_h\right)(z_i) = \frac{h_{i-1}}{h_{i-1} + h_i} \underline{\Psi}_h(z_{i-1}) + \underline{\Psi}_h(z_i) + \frac{h_{i+2}}{h_{i+1} + h_{i+2}} \underline{\Psi}_h(z_{i+1}).$$

As relaxation we employ a line moment relaxation that updates all moments simultaneously for a given spatial point. Our computational tests showed essentially no differences in the error reduction and smoothing properties of this line relaxation scheme for various different orderings of the spatial points. To save computation, we use this line relaxation scheme in a red-black fashion, since then the residual after one relaxation sweep is zero at the black points and need not be computed for the restriction to the next coarser grid. This scheme is also more amenable to parallel computer architectures.

The convergence rates for a $V(1,1)$-cycle of this multigrid algorithm, which uses one relaxation before and one relaxation after the coarse grid correction, for a uniform grid are listed in Table 5.2. Even for values of $\sigma_t = 1/\varepsilon \geq 10^6$, we get $V(1,1)$-cycle convergence factors of order 0.1. These convergence factors are sufficient to get a solution with an error on the order of the discretization error by one single full-multigrid cycle [29]. For irregular grids or for discontinuous coefficients σ_t and σ_a (under the assumptions that these coefficients are continuous in each cell of the coarsest grid) the $V(1,1)$-cycle convergence rates are approximately 0.2. For more numerical studies see [23].

6. Conclusions. The least-squares finite-element discretization with piecewise linear basis functions in space directly applied to the neutron transport equation does not yield a correct discrete solution in diffusive regimes. However, in combination with a scaling transformation applied to the transport operator prior to the discretization, the least-squares discretization is accurate for diffusive regimes and represents a systematic, general, solution approach.

This approach, which converts the first-order transport problem into a variational form with a symmetric bilinear form, is systematic because it includes the theory for the existence and uniqueness of the analytical as well as for the discrete solution, bounds for the discretization error, and guidance for the development of an efficient multigrid solver for the resulting discrete system.

The key results are the V-ellipticity and the continuity of the scaled least-squares bilinear form with constants independent of ε and α. They make it possible to establish error bounds that remain valid in diffusive regimes. Together with the freedom to choose a discrete space, this approach yields a general framework for finding discretizations for the transport equation that are accurate in diffusive regimes.

Because of its generality, this approach opens many possibilities for future work. The error bounds for the P_N discretization in this paper are restricted to diffusive regimes because of the assumption (4.3). This assumption will not be reasonable for nondiffusive regimes and the correct treatment of the boundary conditions then become important. In a forthcoming paper we will analyze how the theory in section 3 can be modified for a P_N discretization with Mark or Marshak boundary conditions. These boundary conditions result in nonconforming finite-element spaces, since $V^h \not\subseteq V$. In this case, the discrete ellipticity has to be proven by using a different Poincaré–Friedrichs inequality.

Another, more natural treatment of the boundary conditions would be to incorporate them directly into the least-squares functional. For example, in the case of slab geometry, this could be done by adding to the bilinear form $a(\cdot,\cdot)$ a boundary form

$$b(\psi, v) := C_b \left(\int_0^1 \mu\psi(z_l, \mu)v(z_l, \mu)d\mu - \int_{-1}^0 \mu\psi(z_r, \mu)v(z_r, \mu)\, d\mu \right),$$

with an appropriate constant C_b depending on the regime. For three-dimensional problems the boundary form becomes

$$b(\psi, v) := C_b \int_{\partial\mathcal{R}} \int_{\underline{n}\cdot\underline{\Omega}<0} |\underline{n} \cdot \underline{\Omega}|\psi v d\Omega d\sigma.$$

We are currently working on the derivation of error bounds that are based on this approach [24].

In the case of slab geometry, the problem of the treatment of the boundary conditions does not arise if finite elements are used for the discretization in angle (phase-space finite-element discretization) and if $\mu = 0$ is chosen as a node. Note that finite elements that are discontinuous in μ are admissible in the functional. The theory of section 3 can be directly applied to this kind of discretization. Discretization error bounds and numerical results for a phase-space finite-element discretization of the transport equation will appear in future work. However, in general, this cannot be

extended easily to three-dimensional problems. Even under the assumption that the spacial region is a polygon, we would need to introduce for every face of the polygon nodes on the unit sphere so that incoming and outcoming directions are separated on the resulting triangulation of the sphere. Therefore, also for phase-space finite-element discretization of three-dimensional problems, the treatment of the boundary conditions by means of the above boundary functional seems to be best.

In addition, future work could be directed to the combination of adaptive refinement with the multigrid solver [26] in order to resolve boundary layers. Moreover, it appears to be possible to generalize the scaling transformation to anisotropic transport problems.

Acknowledgments. The early work on this paper was done while both authors were visiting the Center for Nonlinear Studies at Los Alamos National Laboratory. The second author is also grateful to the Center of Computational Mathematics of the University of Colorado at Denver for using their facilities while finishing his graduate studies. Thanks are also due to S.F. McCormick, J.E. Morel, and John McGhee for illuminating discussions. Furthermore, we like to thank the referee, whose many valuable comments improved the paper.

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] R.E. Alcouffe, E.W. Larsen, W.F. Miller, and B.R. Wienke, *Computational efficiency of numerical methods for the multigroup, discrete ordinates neutron transport equations: The slab geometry case*, Nuclear Sci. Engrg., 71 (1979), pp. 111–127.
[3] G.B. Arfken, *Mathematical Methods for Physicists*, 2nd ed., Academic Press, New York, 1971.
[4] J.H. Bramble, *Multigrid Methods*, Pitman Research Notes in Mathematics Series 294, Longman Scientific and Technical, Essex, UK, 1993.
[5] W.L. Briggs, *A Multigrid Tutorial*, SIAM, Philadelphia, 1987.
[6] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics, Springer-Verlag, New York, 1994.
[7] P.G. Ciarlet and J.L. Lions, *Handbook of Numerical Analysis, Vol.* II, *Finite Element Methods*, Elsevier Science Publishers, North–Holland, Amsterdam, 1991.
[8] J.A. Davis, *Variational Boundary Conditions for a P_N Approximation*, Nuclear Sci. Engrg., 25 (1966), pp. 189–197.
[9] J.J. Duderstadt and W.R Martin, *Transport Theory*, John Wiley, New York, 1978.
[10] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1977.
[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Publishing Program, Boston, MA, 1985.
[12] G.J. Habetler and B.J. Matkowskya, *Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation*, J. Math. Phys., 16 (1975), pp. 846–854.
[13] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[14] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1990.
[15] E.W. Larsen, *Diffusion theory as an asymptotic limit of transport theory for nearly critical systems with small mean free path*, Ann. Nuclear Energy, 7 (1980), pp. 249–255.
[16] E.W. Larsen, J.E. Morel, and W.F. Miller, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, J. Comput. Phys., 69 (1987), pp. 283–324.
[17] E.W. Larsen and J.E. Morel, *Asymptotic solutions of numerical transport problems in optically thick diffusive regimes* II, J. Comput. Phys., 83 (1989), p. 212.
[18] E.E. Lewis and W.F. Miller, *Computational Methods of Neutron Transport*, John Wiley, New York, 1984.
[19] F. Malvagi and G.C. Pomraning, *Initial and boundary conditions for diffusive transport problems*, J. Math. Phys., 32 (1991), p. 805.

[20] T.A. Manteuffel, S.F. Mccormick, J.E. Morel, S. Oliveira, and G. Yang, *A fast multigrid algorithm for isotropic transport problems,* I: *Pure scattering*, SIAM J. Sci. Comput., 16 (1995), pp. 601–635.

[21] T.A Manteuffel, S.F. McCormick, J.E. Morel, S. Oliveira, and G. Yang, *A parallel version of a multigrid algorithm for isotropic transport equations*, SIAM J. Sci. Comput., 15 (1994), pp. 474–493.

[22] T.A. Manteuffel and K.J. Ressel, *Multilevel methods for transport equations in diffusive regimes*, in Proc. Copper Mountain Conference on Multigrid Methods, 1993.

[23] T.A. Manteuffel and K.J. Ressel, *A systematic solution approach for neutron transport problems in diffusive regimes*, in Proc. Copper Mountain Conference on Multigrid Methods, 1995.

[24] T.A. Manteuffel, K.J. Ressel, and G. Starke, *A Boundary Functional for the Least-Squares Finite-Element Solution of Neutron Transport Problems*, SIAM J. Numer. Anal., submitted.

[25] S.F. Mccormick, *Multigrid Methods*, Frontiers in Applied Mathematics 3, SIAM, Philadelphia, 1987.

[26] S.F. McCormick, *Multilevel Adaptive Methods for Partial Differential Equations*, Frontiers in Applied Mathematics, SIAM, Philadelphia, 1989.

[27] S.F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, SIAM, Philadelphia, 1992.

[28] G.C. Pomraning, *Diffusive limits for linear transport equations*, Nuclear Sci. Engrg., 112 (1992), pp. 239–255.

[29] K.J. Ressel, *Least-Squares Finite-Element Solution of the Neutron Transport Equation in Diffusive Regimes*, Ph.D. thesis, University of Colorado at Denver, 1994.