

A CONVERGENT METHOD FOR LINEAR HALF-SPACE KINETIC EQUATIONS ^{*}, ^{**}

QIN LI¹, JIANFENG LU² AND WEIRAN SUN³

Abstract. We give a unified proof for the well-posedness of a class of linear half-space equations with general incoming data and construct a Galerkin method to numerically resolve this type of equations in a systematic way. Our main strategy in both analysis and numerics includes three steps: adding damping terms to the original half-space equation, using an inf-sup argument and even-odd decomposition to establish the well-posedness of the damped equation, and then recovering solutions to the original half-space equation. The proposed numerical methods for the damped equation is shown to be quasi-optimal and the numerical error of approximations to the original equation is controlled by that of the damped equation. This efficient solution to the half-space problem is useful for kinetic-fluid coupling simulations.

Mathematics Subject Classification. 35F15, 35Q79.

Received May 13, 2015. Revised July 12, 2016. Accepted November 29, 2016.

1. INTRODUCTION

In this paper we propose a Galerkin method for computing a class of half-space kinetic equation with given incoming data:

$$\begin{aligned} (v_1 + u)\partial_x f + \mathcal{L}f &= 0, & x \in [0, +\infty), v \in \mathbb{V} \subseteq \mathbb{R}^d, \\ f|_{x=0} &= \phi(v), & v_1 + u > 0. \end{aligned} \tag{1.1}$$

where $u \in \mathbb{R}$ is a given constant, x is the spatial variable and v is the velocity variable. Typical examples for the velocity space \mathbb{V} are $\mathbb{V} = [-1, 1]$ and $\mathbb{V} = \mathbb{R}^d$. The density function f is vector-valued when the system has

Keywords and phrases. Half-space equations, boundary layer, kinetic-fluid coupling, Galerkin method.

^{*} We would like to express our gratitude to the NSF grant RNMS11-07444 (KI-Net), whose activities initiated our collaboration. The research of Q.L. was supported in part by the AFOSR MURI grant FA9550-09-1-0613 and the National Science Foundation under award DMS-1318377. The research of J.L. was supported in part by the Alfred P. Sloan Foundation and the National Science Foundation under award DMS-1312659. The research of W.S. was supported in part by the Simon Fraser University President's Research Start-up Grant PRSG-877723 and NSERC Discovery Individual Grant #611626.

^{**} J.L. would also like to thank Zheng Chen, Jian-Guo Liu, Chi-Wang Shu for helpful discussions. W.S. would like to thank Cory Hauck for pointing out the reference [ES12].

¹ Computing and Mathematical Sciences, California Institute of Technology, 1200 E California Blvd. MC 305-16, Pasadena, CA 91125 USA. Present address: Department of Mathematics, University of Wisconsin-Madison, Madison, WI, 53705 USA. qinli@math.wisc.edu

² Department of Mathematics, Department of Physics, and Department of Chemistry, Duke University, Box 90320, Durham, NC 27708 USA. jianfeng@math.duke.edu

³ Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC V5A 1S6, Canada. weirans@sfu.ca

multiple species. The integral operator \mathcal{L} only acts on the velocity variable v . The specific structure and main assumptions regarding \mathcal{L} will be given in Section 2.

In asymptotic analysis, half-space equations arise as leading-order boundary-layer equations for kinetic equations with multi-scales. Their solutions bridge the gap between the fluid and kinetic boundary conditions. One motivation of our work is to study the kinetic-fluid coupling using the domain-decomposition method, where the half-space equation serves as the intermediate equation between the fluid and kinetic regimes. In this case, understanding the well-posedness of (1.1) and constructing accurate and efficient numerical schemes to resolve it will provide explicit characterization of the couplings.

In the literature the well-posedness of equation (1.1) has long been investigated [3–5, 7, 16, 22] for various models. For example, when \mathcal{L} is the linearized Boltzmann operator, the well-posedness of such half-space equation is fully proved in the fundamental work by Coron, Golse, and Sulem [7]. In this work, it is shown that depending on the choices of u , one needs to prescribe various numbers of additional boundary conditions such that (1.1) is well-posed. These numbers of boundary conditions correspond to the counting of the incoming Euler characteristics at $x = \infty$. The proof in [7] relies mainly on the energy method. Subsequently, a different proof using a variational formulation of (1.1) for the linearized Boltzmann equation is given in [22]. The key idea in [22] is to revise (1.1) by adding certain damping terms. The revised collision operator thus obtained is coercive and it enforces the end-state of f at $x = \infty$ to be zero. By the conservation properties of \mathcal{L} , the authors then show that (1.1) is well-posed for a large class of incoming data. One restriction in [22] is that u cannot be chosen in the way such that the Mach number of the system is $-1, 1$, or 0 . This restriction was later removed in [16].

The variational formulation is also a common tool in proving the well-posedness of the neutron transport equations over general bounded domains Ω in \mathbb{R}_x^d . There is a vast literature in this direction and we will only review some of the main framework and results in [14] which are most relevant to us. In [14], the linear operator \mathcal{L} is the subcritical neutron transport operator. Hence it has a trivial null space. The main novelty of [14] is that one decomposes the solution f into its even and odd parts in v and imposes different regularities for these two parts. Using this mixed regularity, the authors of [14] write the kinetic equation into a variational form and verify that the bilinear operator involved satisfies an inf-sup condition over a properly chosen function space. Moreover, they show that for appropriately constructed Galerkin approximations, the bilinear operator satisfies the inf-sup condition over finite-dimensional approximation spaces as well. This then shows the Galerkin approximation is quasi-optimal. Note that the even-odd parity was widely used for transport equations, see for example [18].

There are two main goals in our paper: first, we will generalize the analysis in [14, 16, 22] to obtain a unified proof for the well-posedness of half-space equations in the form of (1.1). Second, we will develop a systematic Galerkin method to numerically resolve (1.1) and obtain accuracy estimates for our scheme.

We now briefly explain our main results and compare them with previous ones in the literature. In terms of analysis, we show that with appropriate additional boundary conditions at $x = \infty$ given in [7], equation (1.1) has a unique solution. The basic framework we use is the even-odd variational formulation developed in [14]. Compared with [14], here we allow the linear operator \mathcal{L} to have a nontrivial null space and the background velocity u to be any arbitrary constant for general models. The number of additional boundary conditions will change with u .

Due to the loss of coercivity of \mathcal{L} , if one directly applies the variational method in [14] then the bilinear operator \mathcal{B} ceases to satisfy the inf-sup condition. To overcome this degeneracy, we utilize the ideas in [16, 22] by adding damping terms to (1.1) and reconstructing solutions to (1.1) from the damped equation. In the case of linearized Boltzmann equation with a single species, we thus recover the results (in the L^2 spaces) in [16, 22].

The main differences between our work and [16, 22] are: first, we use a different variational formulation which is convenient for performing numerical analysis. Second, the reconstruction in [16, 22] is restricted to a set of incoming data with a finite codimension such that the damping terms are identically zero. Here we use slightly different damping terms and we recover solutions to (1.1) from the damped equation for any incoming data.

On the other hand, our main concern is the convergence and accuracy of the numerical scheme and the basic L^2 -spaces are sufficient for this purpose. Therefore, except for the hard sphere case, we do not try to achieve decay rates estimates of the half-space solution to its end-state at $x = \infty$, while in the literature there are a lot of works that show subexponential or superpolynomial decay of the solution to its end-state for hard or soft potentials for the linearized Boltzmann equation (see for example [8, 23, 24]).

Our analysis also applies to linearized Boltzmann equations with multiple species and linear neutron transport equations with critical or subcritical scatterings, thus providing an alternative proof to the well-posedness result (in the L^2 -space) in [5].

In parallel with the analysis, numerically we first solve the damped half-space equation and then recover the solution to the original equation. We will use a spectral method and achieve quasi-optimal accuracy (for the damped equation) as in [14]. The spectral method dates back to Degond and Mas-Gallic [13] for solving radiative transfer equations, and was later extended by Coron [9] to solving the linearized BGK equation as well. Compared with these works, our approach differs in three ways: First, as a result of using the even-odd formulation, we can derive explicit boundary conditions for the approximate equations. In particular, the number of these boundary conditions is shown to be consistent with the number of the unknowns. Hence our discrete systems are always well-posed. This was not the case in [9] where a least square method was used to solve a potentially overdetermined problem. Second, the method in [9] used Hermite functions defined on the whole velocity space as their basis functions. This leads to severe Gibbs phenomenon, since in general the solution to the half-space equation has a finite jump at $x = 0$ and $v = -u$. Here we choose to use basis functions with jumps at $v = -u$ which naturally fit into the even-odd formulation. This idea is inline with the double P_N method. Third, we will treat the cases with arbitrary bulk velocities u in a uniform way while in [9] different schemes are used for the cases $u = 0$ and $u \neq 0$.

Since the main purpose of the current work is to establish the basic theoretical framework for solving the half-space equations, we only present two numerical examples in this paper. Both of them are for 1D velocity space and a single species. More extensive tests for multi-dimensional velocity space, multi-species, and multi-frequency cases will be done in a forthcoming paper [19] where general boundary conditions including various reflections at the boundary are considered.

There are also non-spectral methods developed for solving the half-space equations. For example, the work by Golse and Klar [15] uses Chapman–Enskog approximation with diffusive closures. The accuracy of these approximations would be hard to analyze: the iterative approach couples the error from the systematic expansion truncation with the numerical error. Moreover, this work ([15]) also treats the cases $u = 0$ and $u \neq 0$ separately. A positivity-preserving DG method was proposed in [6] to treat the Vlasov–Boltzmann transport equation where algebraical convergence is proved. The recent work by Besse *et al.* [2] treats the half-space problem as a boundary layer matching kinetics with the limiting fluid equation, where a Marshak type approximation [20] is applied for boundary fluxes. Similar idea was also used in [12]. As shown already in [9], in general the Marshak approximation does not yield accurate approximations to the half-space problem.

The layout of this paper as follows: in Section 2, we gather the basic information related to the linear operator \mathcal{L} and the properties of the damped operator we will be using in the proof, together with the variational formulation we use. Section 3 is devoted to show the well-posedness of the damped equation and the recovery of the original equation. In Section 4 we show its numerical counterpart and present the result on the Galerkin approximation. Section 5 collects all numerical schemes and results for the linearized BGK and linear transport equations.

2. LINEAR OPERATOR AND BASIC SETTING

In this section we will set the framework for our analysis and numerics. In particular, we will show the basic assumptions about the collision operator \mathcal{L} and the structure of the damped operator and present the variational formulation of a damped version of (1.1).

2.1. Linear collision operator

In order to state the main assumptions imposed on \mathcal{L} , we first introduce some notations. Denote $\text{Null } \mathcal{L}$ as the null space of \mathcal{L} . Let $\mathcal{P} : (L^2(dv))^m \rightarrow \text{Null } \mathcal{L}$ be the projection onto $\text{Null } \mathcal{L}$. Define the weight function

$$a(v) = (1 + |v|)^{\omega_0}, \tag{2.1}$$

for some $0 \leq \omega_0 \leq 1$. Throughout the paper we use

$$\langle f, g \rangle = \langle f, g \rangle_v = \int_{\mathbb{V}} f \cdot g \, dv, \quad \langle f, g \rangle_{x,v} = \int_{\mathbb{R}^d} \int_{\mathbb{V}} f \cdot g \, dv \, dx. \tag{2.2}$$

2.1.1. Assumptions on \mathcal{L}

The main assumptions on \mathcal{L} are as follows:

(A1) $\mathcal{L} : \mathcal{D}(\mathcal{L}) \rightarrow (L^2(dv))^m$ is self-adjoint, nonnegative, and its domain is given by

$$\mathcal{D}(\mathcal{L}) = \{f \in (L^2(dv))^m \mid a(v)f \in (L^2(dv))^m\} \subseteq (L^2(dv))^m,$$

where $a(v)$ is defined in (2.1).

(A2) $\mathcal{L} : (L^2(a \, dv))^m \rightarrow (L^2(\frac{1}{a} \, dv))^m$ is bounded, that is, there exists a constant $\sigma_0 > 0$ such that

$$\|\mathcal{L}f\|_{(L^2(\frac{1}{a} \, dv))^m} \leq \sigma_0 \|f\|_{(L^2(a \, dv))^m}.$$

(A3) $\text{Null } \mathcal{L}$ is finite dimensional and $\text{Null } \mathcal{L} \subseteq (L^p(dv))^m$ for all $p \in [1, \infty)$.

(A4) \mathcal{L} has a spectral gap: there exists $\sigma_0 > 0$ such that

$$\langle f, \mathcal{L}f \rangle \geq \sigma_0 \|\mathcal{P}^\perp f\|_{(L^2(a \, dv))^m}^2 \quad \text{for any } f \in (L^2(a \, dv))^m,$$

where $\mathcal{P}^\perp = \mathcal{I} - \mathcal{P}$ is the projection (in $(L^2(dv))^m$) onto the null orthogonal space $(\text{Null } \mathcal{L})^\perp$.

Note that Assumption (A4) guarantees that \mathcal{L} has a bounded inverse on $(\text{Null } \mathcal{L})^\perp$. Throughout this paper, we denote \mathcal{L}^{-1} as its pseudo-inverse on $(L^2(dv))^m$.

One operator that is of particular importance is $\mathcal{P}_1 : \text{Null } \mathcal{L} \rightarrow \text{Null } \mathcal{L}$ which is defined by

$$\mathcal{P}_1(f) = \mathcal{P}((v_1 + u)f) \quad \text{for any } f \in \text{Null } \mathcal{L}.$$

Note that \mathcal{P}_1 is a symmetric operator on the finite dimension space $\text{Null } \mathcal{L}$. Therefore, its eigenfunctions form a complete set of basis of $\text{Null } \mathcal{L}$. Denote H^+, H^-, H^0 as the eigenspaces of \mathcal{P}_1 corresponding to positive, negative, and zero eigenvalues respectively and denote their dimensions as

$$\dim H^+ = \nu_+, \quad \dim H^- = \nu_-, \quad \dim H^0 = \nu_0.$$

Let $X_{+,i}, X_{-,j}, X_{0,k}$ be the associated unit eigenfunctions with $1 \leq i \leq \nu_+, 1 \leq j \leq \nu_-,$ and $1 \leq k \leq \nu_0$ for $\nu_\pm, \nu_0 \neq 0$. Note that if any of ν_\pm, ν_0 is equal to zero, then we simply do not have any eigenfunction associated with the corresponding eigenspace. By their definitions, these eigenfunctions satisfy

$$\begin{aligned} \langle X_{\alpha,\gamma}, X_{\alpha',\gamma'} \rangle_v &= \delta_{\alpha\alpha'} \delta_{\gamma\gamma'}, & \langle (v_1 + u)X_{\alpha,\gamma}, X_{\alpha',\gamma'} \rangle_v &= 0 \text{ if } \alpha \neq \alpha' \text{ or } \gamma \neq \gamma', \\ \langle (v_1 + u)X_{0,j}, X_{0,k} \rangle_v &= 0, & \langle (v_1 + u)X_{+,j}, X_{+,i} \rangle_v &> 0, & \langle (v_1 + u)X_{-,j}, X_{-,i} \rangle_v &< 0, \end{aligned} \tag{2.3}$$

where $\alpha \in \{+, -, 0\}, \gamma \in \{i, j, k\}, 1 \leq i \leq \nu_+, 1 \leq j \leq \nu_-,$ and $1 \leq k \leq \nu_0$. These relations in particular give that

$$(v_1 + u)X_{0,j} \in (\text{Null } \mathcal{L})^\perp, \quad j = 1, \dots, \nu_0.$$

Therefore $\mathcal{L}^{-1}((v_1 + u)X_{0,j}) \in (\text{Null } \mathcal{L})^\perp$ is well-defined.

2.1.2. Examples of \mathcal{L}

Many well-known linear or linearized kinetic models satisfy the assumptions (A1)–(A4) for the collision operators. These include the classical linearized Boltzmann equations for either single-species system or multi-species with hard-sphere collisions and the linear neutron transport equations. The particular equations that we use as numerical examples are the isotropic neutron transport equation (NTE) with slab geometry and the linearized BGK equation. Similar analysis can be carried out to models satisfying (A1)–(A4) without extra difficulties. The main structure of these two equations are as follows. The linear operator of the isotropic NTE is the simplest scattering operator which has the form

$$\mathcal{L}f = f - \frac{1}{2} \int_{-1}^1 f(v) \, dv. \tag{2.4}$$

In this case, $a(v) = 1 + |v| = \mathcal{O}(1)$ and $(L^2(a \, dv))^m$ coincides with $(L^2(dv))^m$.

The linearized BGK operator is the linearization of the nonlinear BGK operator, which is introduced as a simplified model that captures some fundamental behavior of the nonlinear Boltzmann equation. The collision operator of the nonlinear BGK is defined as

$$\mathcal{Q}[F] = F - \mathcal{M}[F],$$

where $\mathcal{M}[F]$ is the local Maxwellian associated with F defined by

$$\mathcal{M}[F] = \frac{\rho}{\sqrt{2\pi\theta}} e^{-\frac{|v-u|^2}{2\theta}},$$

where

$$\rho = \int_{\mathbb{R}} F \, dv, \quad \rho u = \int_{\mathbb{R}} vF \, dv, \quad \rho u^2 + \rho\theta = \int_{\mathbb{R}} v^2 F \, dv.$$

For a given bulk velocity $u \in \mathbb{R}$, define the global Maxwellian with the steady state $(\rho, u, \theta) = (1, u, 1/2)$ as

$$M_u = \frac{1}{\sqrt{\pi}} e^{-|v-u|^2}.$$

Linearizing the operator \mathcal{Q} around M by setting

$$F = M_u + \sqrt{M_u} f,$$

we obtain the linearized BGK operator

$$\mathcal{L}_u f = f - m_u,$$

where $m_u(v)$ is f projected onto the kernel space of \mathcal{L}_u . In the case of the 1D linearized BGK, one has:

$$\text{Null } \mathcal{L}_u = \text{span} \left\{ \sqrt{M_u}, v\sqrt{M_u}, v^2\sqrt{M_u} \right\}.$$

Therefore, $m_u(v)$ is a quadratic function associated with a Maxwellian to 1/2 power:

$$m_u(v) = (\tilde{\rho} + \tilde{u}(v - u) + \frac{\tilde{\theta}}{2}((v - u)^2 - 1))\sqrt{M_u},$$

where $(\tilde{\rho}, \tilde{u}, \tilde{\theta})$ are defined in the way such that first three moments of $m(v)$ agree with those of f :

$$\left\langle f - m_u, v^k \sqrt{M_u} \right\rangle = \int_{\mathbb{R}} (f - m_u) v^k \sqrt{M_u} \, dv = 0, \quad k = 0, 1, 2.$$

The half-space equation with the linearized BGK operator that centered at bulk velocity u is:

$$\begin{aligned} v\partial_x f + \mathcal{L}_u f &= 0, \\ f|_{x=0} &= \phi(v), \quad v > 0. \end{aligned} \tag{2.5}$$

Following the classical treatment of the half-space equations, we shift the center of the Maxwellian M_u to the origin by performing the change of variable $v - u \rightarrow v$. The half-space equation (2.5) then becomes

$$\begin{aligned} (v + u)\partial_x f + \mathcal{L}f &= 0, \\ f|_{x=0} &= \phi(v + u), \quad v + u > 0, \end{aligned} \tag{2.6}$$

where

$$\mathcal{L}f = f - m(v), \quad m(v) = m_0, \tag{2.7}$$

and the null space of \mathcal{L} becomes

$$\text{Null } \mathcal{L} = \text{span}\{\sqrt{M}, v\sqrt{M}, v^2\sqrt{M}\},$$

where M is the global Maxwellian centered at the origin such that

$$M = M_0 = \frac{1}{\sqrt{\pi}}e^{-v^2}.$$

As defined in (2.3), we look for $H^{\pm,0}$ decomposition of $\text{Null } \mathcal{L}$. For this particular case one could write down the basis functions explicitly. Following [7], we define

$$\begin{cases} \chi_0 = \frac{1}{6^{1/2}\pi^{1/4}}(2v^2 - 3)\exp(-v^2/2) \\ \chi_{\pm} = \frac{1}{6^{1/2}\pi^{1/4}}(\sqrt{6}v \pm 2v^2)\exp(-v^2/2). \end{cases} \tag{2.8}$$

It is easy to show that

$$\begin{cases} \langle \chi_{\alpha}, \chi_{\beta} \rangle_v = \int_{\mathbb{R}} \chi_{\alpha} \chi_{\beta} \, dv = \delta_{\alpha\beta}, \\ \langle (v + u)\chi_{\alpha}, \chi_{\beta} \rangle_v = 0, \quad \alpha \neq \beta, \\ \langle (v + u)\chi_0, \chi_0 \rangle_v = u_0 = u, \\ \langle (v + u)\chi_+, \chi_+ \rangle_v = u_+ = u + c, \\ \langle (v + u)\chi_-, \chi_- \rangle_v = u_- = u - c, \end{cases} \tag{2.9}$$

where $\alpha, \beta \in \{+, -, 0\}$, $c = \sqrt{3/2}$, and

$$\langle f, g \rangle_v = \int_{\mathbb{R}} fg \, dv.$$

Using these new basis functions, we can decompose $\text{Null } \mathcal{L}$ into subspaces: $\text{Null } \mathcal{L} = H^+ \oplus H^- \oplus H^0$ with:

$$H^+ = \text{span}\{\chi_{\beta} \mid u_{\beta} > 0\}, \quad H^- = \text{span}\{\chi_{\beta} \mid u_{\beta} < 0\}, \quad H^0 = \text{span}\{\chi_{\beta} \mid u_{\beta} = 0\},$$

where again $\beta \in \{+, -, 0\}$. For each fixed $u \in \mathbb{R}$, denote the dimensions of these subspaces as

$$\dim H^+ = \nu_+, \quad \dim H^- = \nu_-, \quad \dim H^0 = \nu_0.$$

Note that ν_{\pm}, ν_0 change with u . In particular, we have the following categories:

$$\begin{cases} u < -c : & (\dim H^+, \dim H^-, \dim H^0) = (0, 3, 0), \\ u = -c : & (\dim H^+, \dim H^-, \dim H^0) = (0, 2, 1), \\ -c < u < 0 : & (\dim H^+, \dim H^-, \dim H^0) = (1, 2, 0), \\ u = 0 : & (\dim H^+, \dim H^-, \dim H^0) = (1, 1, 1), \\ 0 < u < c : & (\dim H^+, \dim H^-, \dim H^0) = (2, 1, 0), \\ u = c : & (\dim H^+, \dim H^-, \dim H^0) = (2, 0, 1), \\ u > c : & (\dim H^+, \dim H^-, \dim H^0) = (3, 0, 0). \end{cases} \tag{2.10}$$

This gives an explicit example that shows the structure of $\text{Null } \mathcal{L}$ changes with u .

2.2. Damped linear operator \mathcal{L}_d

The main difficulty in both analysis and numerics is the non-coercivity of \mathcal{L} . Although in some cases this degeneracy of \mathcal{L} can be handled by carefully choosing appropriate function spaces for the variational formulation, we prefer to work with strictly dissipative operators. To this end, we utilize the idea developed in [16, 22] to modify the original equation (2.6) by adding in damping terms. The particular damping terms are chosen in the way such that we can easily recover the undamped equation (2.6) for any incoming data and such that the damped operator is symmetric. The particular damped operator we introduce is

$$\begin{aligned} \mathcal{L}_d f &= \mathcal{L} f + \alpha \sum_{k=1}^{\nu_+} (v_1 + u) X_{+,k} \langle (v_1 + u) X_{+,k}, f \rangle_v \\ &+ \alpha \sum_{k=1}^{\nu_-} (v_1 + u) X_{-,k} \langle (v_1 + u) X_{-,k}, f \rangle_v + \alpha \sum_{k=1}^{\nu_0} (v_1 + u) X_{0,k} \langle (v_1 + u) X_{0,k}, f \rangle_v \\ &+ \alpha \sum_{k=1}^{\nu_0} (v_1 + u) \mathcal{L}^{-1}((v_1 + u) X_{0,k}) \langle (v_1 + u) \mathcal{L}^{-1}((v_1 + u) X_{0,k}), f \rangle_v. \end{aligned} \tag{2.11}$$

Here the constant α satisfies that $0 < \alpha \ll 1$. The size of α only depends on \mathcal{L} . The main property of \mathcal{L}_d is its coercivity as stated in the following lemma:

Lemma 2.1. *Let \mathcal{L} be the linear operator that satisfies Assumptions (A1)–(A4). Then there exist two constants $\sigma_1, \alpha_0 > 0$ such that for any $0 < \alpha \leq \alpha_0$ we have*

$$\langle f, \mathcal{L}_d f \rangle \geq \sigma_1 \|f\|_{(L^2(\text{adv}))^m}^2 \quad \text{for any } f \in \mathcal{D}(\mathcal{L}).$$

Proof. By the definition of \mathcal{L}_d , we have

$$\begin{aligned} \langle f, \mathcal{L}_d f \rangle &= \langle f, \mathcal{L} f \rangle + \alpha \sum_{k=1}^{\nu_+} \langle (v_1 + u) X_{+,k}, f \rangle^2 + \alpha \sum_{k=1}^{\nu_-} \langle (v_1 + u) X_{-,k}, f \rangle^2 + \alpha \sum_{k=1}^{\nu_0} \langle (v_1 + u) X_{0,k}, f \rangle^2 \\ &+ \alpha \sum_{k=1}^{\nu_0} \langle (v_1 + u) \mathcal{L}^{-1}((v_1 + u) X_{0,k}), f \rangle^2. \end{aligned}$$

Write

$$f = f^\perp + \sum_{i=1}^{\nu_+} f_{+,i} X_{+,i} + \sum_{j=1}^{\nu_-} f_{-,j} X_{-,j} + \sum_{k=1}^{\nu_0} f_{0,k} X_{0,k},$$

where $f^\perp = \tilde{\mathcal{P}} f \in (\text{Null } \mathcal{L})^\perp$. By Assumption (A4), if we chose $0 < \alpha \ll 1$, then

$$\begin{aligned} \langle f, \mathcal{L}_d f \rangle &\geq \sigma_0 \|f^\perp\|_{(L^2(\text{adv}))^m}^2 + \frac{\alpha}{4} \sum_{k=1}^{\nu_+} \gamma_{+,k}^2 f_{+,k}^2 + \frac{\alpha}{4} \sum_{k=1}^{\nu_+} \gamma_{-,k}^2 f_{-,k}^2 \\ &- \frac{\alpha}{4} \sum_{k=1}^{\nu_+} \langle (v_1 + u) X_{+,k}, f^\perp \rangle^2 - \frac{\alpha}{4} \sum_{k=1}^{\nu_-} \langle (v_1 + u) X_{-,k}, f^\perp \rangle^2 \\ &\geq \frac{\sigma_0}{2} \|f^\perp\|_{(L^2(\text{adv}))^m}^2 + \frac{\alpha}{4} \sum_{k=1}^{\nu_+} \gamma_{+,k}^2 f_{+,k}^2 + \frac{\alpha}{4} \sum_{k=1}^{\nu_-} \gamma_{-,k}^2 f_{-,k}^2, \end{aligned} \tag{2.12}$$

where γ_\pm 's are defined as

$$\begin{aligned} \gamma_{+,i} &:= \langle (v_1 + u) X_{+,i}, X_{+,i} \rangle_v > 0, & 0 \leq i \leq \nu_+, \\ \gamma_{-,j} &:= -\langle (v_1 + u) X_{-,j}, X_{-,j} \rangle_v > 0, & 0 \leq j \leq \nu_-. \end{aligned} \tag{2.13}$$

In addition, if $\nu_0 \neq 0$, then

$$\begin{aligned}
 \langle f, \mathcal{L}_d f \rangle &\geq \sigma_0 \|f^\perp\|_{(L^2(a \, dv))^m}^2 + \alpha \sum_{k=1}^{\nu_0} \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), f \rangle^2 \\
 &\geq \frac{\sigma_0}{2} \|f^\perp\|_{(L^2(a \, dv))^m}^2 + \frac{\alpha}{4\nu_0} \left(\sum_{k,m=1}^{\nu_0} \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), X_{0,m} \rangle_v f_{0,m} \right)^2 \\
 &\quad - \frac{\alpha}{4} \sum_{k=1}^{\nu_0} \left\langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), \sum_{m=1}^{\nu_+} f_{+,k} X_{+,k} \right\rangle^2 \\
 &\quad - \frac{\alpha}{4} \sum_{k=1}^{\nu_0} \left\langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), \sum_{m=1}^{\nu_+} f_{-,k} X_{-,k} \right\rangle^2 \\
 &\quad - \frac{\alpha}{4} \sum_{k=1}^{\nu_0} \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), f^\perp \rangle^2. \tag{2.14}
 \end{aligned}$$

Since the matrix $(\langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), X_{0,m} \rangle_v)$ is strictly positive, there exists a constant $c_0 > 0$ such that

$$\sum_{k=1}^{\nu_0} \left(\sum_{m=1}^{\nu_0} \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), X_{0,m} \rangle_v f_{0,m} \right)^2 \geq c_0 \sum_{k=1}^{\nu_0} f_{0,m}^2. \tag{2.15}$$

Hence by multiplying (2.12) by a large enough number and adding it to (2.14), we have

$$\langle f, \mathcal{L}_d f \rangle \geq \sigma_1 \|f\|_{(L^2(a \, dv))^m}^2 \quad \text{for some } \sigma_1 > 0. \tag{2.16}$$

provided $0 < \alpha \ll 1$. □

2.3. Variational formulation

In this part we present the variational formulation for the half-space equation. First, we state the full equation that we want to study in this paper using the notation of H^\pm, H^0 . Suppose \mathcal{L} is a linear operator in v that satisfies (A1)–(A4). Our goal is to prove the well-posedness of the following equation and then construct efficient numerical schemes and obtain estimate of its accuracy:

$$\begin{aligned}
 (v_1 + u)\partial_x f + \mathcal{L}f &= 0, & x \in [0, +\infty), \quad v \in \mathbb{V}, \\
 f|_{x=0} &= \phi(v), & v_1 + u > 0, \\
 f - f_\infty &\in (L^2(dv \, dx))^m,
 \end{aligned} \tag{2.17}$$

for some $f_\infty \in H^+ \oplus H^0$. The particular formulation about the end-state f_∞ was given in [7] (for single species $m = 1$) where the authors proved the well-posedness of the half-space linearized Boltzmann equation:

Theorem 2.2 [7]. *Let \mathcal{L} be the linearized Boltzmann operator with a hard-sphere collision kernel and the incoming data $\phi \in L^2(a(v)\mathbf{1}_{v_1+u>0} \, dv)$. Then there exists a constant $\beta > 0$ and a unique $f_\infty \in H^+ \oplus H^0$ such that equation (1.1) has a unique solution f which satisfies*

$$f - f_\infty \in L^2(e^{2\beta x} \, dx; L^2(a \, dv)),$$

where $a(v) = 1 + |v|$.

Remark 2.3. The main result in [7] is actually stronger than Theorem 2.2 where $f - f_\infty$ is shown to be in $L^\infty(e^{2\beta x} dx; L^2(dv))$. Here we content ourselves with the L^2 -weighted space (in x) since L^2 suffices our needs in proving the quasi-optimal convergence of our numerical scheme.

We will use $\mathbb{V} = \mathbb{R}^3$ as the setting to explain the variational formulation. Other spaces for v will work in a similar way. Let $u \in \mathbb{R}$ be given. We use the damped operator \mathcal{L}_d and obtain the modified equation as

$$\begin{aligned} (v_1 + u)\partial_x f + \mathcal{L}_d f &= 0, \\ f|_{x=0} &= \phi(v_1), \quad v_1 + u > 0. \end{aligned} \tag{2.18}$$

We define the shifted “even” and “odd” parts of a function as

$$f^+(v) = \frac{f(v_1, v_2, v_3) + f(-2u - v_1, v_2, v_3)}{2}, \quad f^-(v) = \frac{f(v_1, v_2, v_3) - f(-2u - v_1, v_2, v_3)}{2} \tag{2.19}$$

such that $f = f^+ + f^-$ and

$$f^\pm(-u + v_1, v_2, v_3) = \pm f^\pm(-u - v_1, v_2, v_3).$$

Define the function space

$$\Gamma = \{f \in (L^2(a \, dv \, dx))^m \mid (v_1 + u)\partial_x f^+ \in (L^2(\frac{1}{a} \, dv \, dx))^m\}, \tag{2.20}$$

which is a Hilbert space with the inner product

$$\langle f, g \rangle_\Gamma = \int_{\mathbb{R}} \int_{\mathbb{R}^3} f \cdot g \, a \, dv \, dx + \int_{\mathbb{R}} \int_{\mathbb{R}^3} (v_1 + u)\partial_x f^+ \cdot (v_1 + u)\partial_x g^+ \frac{1}{a} \, dv \, dx.$$

Thus the norm of Γ is equivalent to

$$\|f\|_{(L^2(a \, dv \, dx))^m} + \|(v_1 + u)\partial_x f^+\|_{(L^2(\frac{1}{a} \, dv \, dx))^m}.$$

Moreover, every element $g \in \Gamma$ has a well-defined trace:

$$\mathcal{T} : \Gamma \rightarrow (L^2(|v_1 + u| \, dv))^m \tag{2.21}$$

such that

$$\mathcal{T}g = g^+|_{x=0}, \quad \text{for all } g \in C([0, \infty); (L^2(a \, dv))^m), \tag{2.22}$$

and

$$\int_{\mathbb{R}^3} |v_1 + u| |g^+|^2 \, dv < \infty. \tag{2.23}$$

Now we define a bilinear operator $\mathcal{B} : \Gamma \times \Gamma \rightarrow \mathbb{R}$ such that

$$\begin{aligned} \mathcal{B}(f, \psi) &= -\langle f^-, (v_1 + u)\partial_x \psi^+ \rangle_{x,v} + \langle (v_1 + u)\partial_x f^+, \psi^- \rangle_{x,v} + \langle \psi, \mathcal{L}_d f \rangle_{x,v} + \langle |v_1 + u| f^+, \psi^+ \rangle_{x=0} \\ &= -\langle f^-, (v_1 + u)\partial_x \psi^+ \rangle_{x,v} + \langle (v_1 + u)\partial_x f^+, \psi^- \rangle_{x,v} + \langle \psi, \mathcal{L} f \rangle_{x,v} \\ &\quad + \alpha \sum_{k=1}^{\nu_+} \langle \langle (v_1 + u)X_{+,k}, \psi \rangle_v, \langle (v_1 + u)X_{+,k}, f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_-} \langle \langle (v_1 + u)X_{-,k}, \psi \rangle_v, \langle (v_1 + u)X_{-,k}, f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_0} \langle \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), \psi \rangle_v, \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_0} \langle \langle (v_1 + u)X_{0,k}, \psi \rangle_v, \langle (v_1 + u)X_{0,k}, f \rangle_v \rangle_x + \langle |v_1 + u| f^+, \psi^+ \rangle_{x=0}. \end{aligned} \tag{2.24}$$

Recall that the inner product $\langle \cdot, \cdot \rangle_{x,v}$ is defined in (2.2). It is straightforward to check by using integration by parts and symmetry that the variational formulation of (2.18) has the form

$$\mathcal{B}(f, \psi) = l(\psi), \quad \text{for every } \psi \in \Gamma. \tag{2.25}$$

Here the linear operator $l(\cdot)$ is given by

$$l(\psi) = 2 \int_{v_1+u>0} (v_1 + u) \phi \psi^+ dv, \tag{2.26}$$

where ϕ is the given incoming data and ψ^+ is the even (with respect to $-u$) part of ψ as defined in (2.19).

3. WELL-POSEDNESS

In this section we show the well-posedness of the half-space equation (2.17). The proof will be done in two steps: first, we use the variational form (2.25) to show the well-posedness of the damped equation (2.18). Then we construct recovering procedures to find the solution to the original half-space equation.

3.1. Solution of the damped equation

The main tool we use to show the well-posedness of the weak formulation (2.25) is to use the Babuška–Aziz lemma [1]. There are two parts in this lemma and we recall its statement below.

Theorem 3.1 Babuška–Aziz. *Suppose Γ is a Hilbert space and $\mathcal{B} : \Gamma \times \Gamma \rightarrow \mathbb{R}$ is a bilinear operator on Γ . Let $l : \Gamma \rightarrow \mathbb{R}$ be a bounded linear functional on Γ .*

- (a) *If \mathcal{B} satisfies the boundedness and inf-sup conditions on Γ such that*
- *there exists a constant $c_0 > 0$ such that $|\mathcal{B}(f, g)| \leq c_0 \|f\|_{\Gamma} \|g\|_{\Gamma}$ for all $f, g \in \Gamma$;*
 - *there exists a constant $\kappa_0 > 0$ such that*

$$\begin{aligned} \sup_{\|f\|_{\Gamma}=1} \mathcal{B}(f, \psi) &\geq \kappa_0 \|\psi\|_{\Gamma}, & \text{for any } \psi \in \Gamma, \\ \sup_{\|\psi\|_{\Gamma}=1} \mathcal{B}(f, \psi) &\geq \kappa_0 \|f\|_{\Gamma}, & \text{for any } f \in \Gamma \end{aligned} \tag{3.1}$$

*for some constant $\kappa_0 > 0$.
then there exists a unique $f \in \Gamma$ which satisfies*

$$\mathcal{B}(f, \psi) = l(\psi), \quad \text{for any } \psi \in \Gamma.$$

- (b) *Suppose Γ_N is a finite-dimensional subspace of Γ . If in addition $\mathcal{B} : \Gamma_N \times \Gamma_N \rightarrow \mathbb{R}$ satisfies the inf-sup condition on Γ_N , then there exists a unique solution f_N such that*

$$\mathcal{B}(f_N, \psi_N) = l(\psi_N), \quad \text{for any } \psi_N \in \Gamma_N.$$

Moreover, f_N gives a quasi-optimal approximation to the solution f in (a), that is, there exists a constant κ_1 such that

$$\|f - f_N\|_{\Gamma} \leq \kappa_1 \inf_{w \in \Gamma_N} \|f - w\|_{\Gamma}.$$

It is clear that the inf-sup condition of \mathcal{B} is essential to the solvability of (2.25). We thus first show that \mathcal{B} satisfies this condition.

Proposition 3.2 Inf-sup. *Let Γ and \mathcal{B} be the function space and the bilinear operator defined in (2.20) and (2.24) respectively. Then $\mathcal{B} : \Gamma \times \Gamma \rightarrow \mathbb{R}$ satisfies the inf-sup condition (3.1).*

Proof. Note that \mathcal{B} is symmetric in its variables. Hence it suffices to show that the second condition in (3.1) holds. To this end, let $f \in \Gamma$ be arbitrary. We only need to find an appropriate ψ such that

$$\mathcal{B}(f, \psi) \geq \kappa_0 \|f\|_{\Gamma}^2, \quad \|\psi\|_{\Gamma} \leq \kappa_1 \|f\|_{\Gamma}. \tag{3.2}$$

Indeed, if ψ satisfies (3.2), then one can simply let $\Psi = \frac{\psi}{\|\psi\|_{\Gamma}}$ and obtain the second inequality in (3.1) (with a different constant). The construction of such ψ will be carried out in two steps. First, let $\psi_1 = f$. Then by Lemma 2.1,

$$\mathcal{B}(f, \psi_1) = \langle f, \mathcal{L}_d f \rangle_{x,v} + \langle (v_1 + u|f^+, f^+) \rangle_{x=0} \geq \sigma_1 \|f\|_{(L^2(a \, dv \, dx))^m}^2.$$

Next, let

$$\psi_2 = \frac{1}{(1 + |v_1 + u| + |v_2| + |v_3|)^{\omega_0}} (v_1 + u) \partial_x f^+.$$

We claim that $\psi_2 \in \Gamma$. Indeed, by the definition of $a(v)$, one can find two constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{a(v)} \leq \frac{1}{(1 + |v_1 + u| + |v_2| + |v_3|)^{\omega_0}} \leq \frac{c_2}{a(v)}.$$

Here the constants c_1, c_2 depend on u . Thus $\psi_2 \in (L^2(a \, dv \, dx))^m$ because

$$\|\psi_2\|_{(L^2(a \, dv \, dx))^m} \leq \|(v_1 + u) \partial_x f^+\|_{(L^2(\frac{1}{a} \, dv \, dx))^m} \leq \|f\|_{\Gamma}.$$

Moreover the definition of ψ_2 implies that

$$\psi_2^+ = 0 \in (L^2(\frac{1}{a} \, dv \, dx))^m.$$

Hence $\psi_2 \in \Gamma$ and it satisfies

$$\|\psi_2\|_{\Gamma} \leq \|f\|_{\Gamma}. \tag{3.3}$$

Using ψ_2 in \mathcal{B} , we have

$$\begin{aligned} \mathcal{B}(f, \psi_2) &= \langle (v_1 + u) \partial_x f^+, \psi_2 \rangle + \langle \psi_2, \mathcal{L} f \rangle + \alpha \sum_{k=1}^{\nu_+} \langle \langle (v_1 + u) X_{+,k}, \psi_2 \rangle_v \langle (v_1 + u) X_{+,k}, f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_-} \langle \langle (v_1 + u) X_{-,k}, \psi_2 \rangle_v \langle (v_1 + u) X_{-,k}, f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_0} \langle \langle (v_1 + u) X_{0,k}, \psi_2 \rangle_v \langle (v_1 + u) X_{0,k}, f \rangle_v \rangle_x \\ &\quad + \alpha \sum_{k=1}^{\nu_0} \langle \langle (v_1 + u) \mathcal{L}^{-1}((v_1 + u) X_{0,k}), \psi_2 \rangle_v \langle (v_1 + u) \mathcal{L}^{-1}((v_1 + u) X_{0,k}), f \rangle_v \rangle_x \\ &\geq \|(v_1 + u) \partial_x f^+\|_{(L^2(\frac{1}{a} \, dv \, dx))^m}^2 - \kappa_2 \|f\|_{(L^2(a \, dv \, dx))^m}^2, \end{aligned}$$

for some constant $\kappa_2 > 0$. Hence by taking $\kappa_3 > 0$ large enough, we have that

$$\mathcal{B}(f, \kappa_3 \psi_1 + \psi_2) \geq \kappa_0 \|f\|_{\Gamma}^2, \tag{3.4}$$

for some $\kappa_0 > 0$. Recall that by the definition of ψ_1 and (3.3), we also have

$$\|\kappa_3 \psi_1 + \psi_2\|_{\Gamma} \leq \sqrt{1 + \kappa_3} \|f\|_{\Gamma},$$

which, together with (3.4), shows the inf-sup property of \mathcal{B} on $\Gamma \times \Gamma$. □

Using the inf-sup property of \mathcal{B} and the Babuška–Aziz Lemma, we can now show the solvability of the variational form (2.25).

Proposition 3.3 Well-posedness of the damped equation. *Suppose \mathcal{L} satisfies Assumption (A1)–(A4) and \mathcal{L}_d is defined as in (2.11) with α small enough such that the coercivity in Lemma 2.1 holds. Let $\phi \in (L^2(a(v)\mathbf{1}_{v_1+u>0} dv))^m$ and Γ be the function space defined in (2.20). Then*

- (a) *There exists a unique $f \in \Gamma$ such that (2.25) holds.*
- (b) *Moreover, f satisfies that*

$$(v_1 + u)\partial_x f \in (L^2(\frac{1}{a} dv dx))^m$$

and it solves the damped half-space equation in the sense of distributions

$$\begin{aligned} (v_1 + u)\partial_x f + \mathcal{L}_d f &= (v_1 + u)\partial_x f + \mathcal{L}f + \alpha \sum_{k=1}^{\nu_+} (v_1 + u)X_{+,k} \langle (v_1 + u)X_{+,k}, f \rangle_v \\ &+ \alpha \sum_{k=1}^{\nu_-} (v_1 + u)X_{-,k} \langle (v_1 + u)X_{-,k}, f \rangle_v + \alpha \sum_{k=1}^{\nu_0} (v_1 + u)X_{0,k} \langle (v_1 + u)X_{0,k}, f \rangle_v \quad (3.5) \\ &+ \alpha \sum_{k=1}^{\nu_0} (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}) \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,k}), f \rangle_v = 0 \end{aligned}$$

with the boundary conditions (defined in the trace sense at $x = 0$)

$$f|_{x=0} = \phi(v), \quad v_1 + u > 0. \tag{3.6}$$

- (c) *If $a(v) = 1 + |v|$, then there exists $\beta > 0$ such that $(L^2(e^{2\beta x} dx; L^2(a dv)))^m$.*

Proof.

- (a) It is straightforward to verify the boundedness of \mathcal{B} and l as defined in (2.24) and (2.26). The well-posedness of the variational form is then an immediate consequence of Proposition 3.2 and part (a) of the Babuška–Aziz lemma.
- (b) In order to show that $(v_1 + u)\partial_x f \in (L^2(\frac{1}{a} dv dx))^m$, we note that the damped equation (3.5) holds in the sense of distributions by choosing the test function $\psi \in C_c^\infty((0, \infty) \times \mathbb{R})$. Thus

$$(v_1 + u)\partial_x f = \beta(v_1 + u)f - \mathcal{L}_d(f) \in (L^2(\frac{1}{a} dv dx))^m.$$

By the density argument this implies that

$$\langle (v_1 + u)\partial_x f^-, \psi^+ \rangle_{x,v} + \langle (v_1 + u)\partial_x f^+, \psi^- \rangle_{x,v} - \beta \langle (v_1 + u)\psi, f \rangle_{x,v} + \langle \psi, \mathcal{L}_d f \rangle_{x,v} = 0,$$

for all $\psi \in C^\infty(0, \infty)$. Therefore, if we choose $\phi \in C^\infty[0, \infty)$ and integrate by parts in the variational form (2.25), then boundary terms satisfy

$$\langle (v_1 + u)f^-, \psi^+ \rangle_v + \langle |v_1 + u|f^+, \psi^+ \rangle_v = 2 \int_{v_1+u>0} (v_1 + u) \phi \psi^+ dv \quad \text{at } x = 0,$$

which implies,

$$\int_{v_1+u>0} (v_1 + u)f\psi^+ dv = \int_{v_1+u>0} (v_1 + u) \phi \psi^+ dv \quad \text{at } x = 0.$$

Since $\psi^+ \in C^\infty(0, \infty)$ is arbitrary, we have $f = \phi$ at $x = 0$ when $v_1 + u > 0$.

(c) If $a(v) = 1 + |v|$, then there exists $\beta > 0$ such that $f \in (L^2(e^{2\beta x} dx; L^2(a dv)))^m$. The proof will be along the same line for the general case of $a(v)$. We use the standard way to incorporate the exponential into the bilinear form by changing f by $g = e^{\beta x} f$. The new bilinear form \mathcal{B}_β is

$$\mathcal{B}_\beta(g, \psi) = \mathcal{B}(g, \psi) - \beta \langle (v_1 + u)g, \psi \rangle_{x,v},$$

where $\mathcal{B}(g, \psi)$ is defined in (2.24). Note that by Cauchy–Schwartz, if we choose $0 < \beta \ll \alpha$, then by the spectral gap assumption (A4), we have

$$\begin{aligned} |\beta \langle (v_1 + u)g, \psi_1 \rangle_{x,v}| &\leq \frac{1}{2} \mathcal{B}(g, \psi), \\ |\beta \langle (v_1 + u)g, \psi_2 \rangle_{x,v}| &\leq \frac{1}{2} \mathcal{B}(g, \psi) + \frac{1}{2} \langle (v_1 + u)\partial_x g^+, \psi_2 \rangle_{x,v}. \end{aligned}$$

Hence, this extra β -term will not affect the inf-sup estimate. Since $g \in (L^2(dv dx))^m$, we have that $f \in (L^2(e^{2\beta x} dx; L^2(dv)))^m$. □

Remark 3.4. Note that $f - f_\infty$ for the neutron transport equations satisfy the exponential decay as $x \rightarrow \infty$ since $a(v) \sim 1$ in this case.

3.2. Recovery of the undamped solution

Using the solution of the damped equation (3.5), we now explicitly construct solutions to the original undamped equation (2.17). First we introduce the following notations: for any solution f to the damped equation (3.5), denote

$$\begin{aligned} \mathbf{U}_+(f) &= (\langle (v_1 + u)X_{+,1}, f \rangle_v, \dots, \langle (v_1 + u)X_{+,\nu_+}, f \rangle_v)^T, \\ \mathbf{U}_-(f) &= (\langle (v_1 + u)X_{-,1}, f \rangle_v, \dots, \langle (v_1 + u)X_{-,\nu_-}, f \rangle_v)^T, \\ \mathbf{U}_0(f) &= (\langle (v_1 + u)X_{0,1}, f \rangle_v, \dots, \langle (v_1 + u)X_{0,\nu_0}, f \rangle_v)^T, \\ \mathbf{U}_{\mathcal{L},0}(f) &= (\langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,1}), f \rangle_v, \dots, \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,1})X_{0,\nu_0}, f \rangle_v)^T, \end{aligned} \tag{3.7}$$

and

$$\mathbf{U}(f) = (\mathbf{U}_+^T(f), \mathbf{U}_-^T(f), \mathbf{U}_0^T(f), \mathbf{U}_{\mathcal{L},0}^T(f))^T. \tag{3.8}$$

Next we define some auxiliary functions. For each $1 \leq i \leq \nu_+$, let $g_{+,i}$ be the solution to (2.18) with boundary conditions given by $X_{+,i}$:

$$g_{+,i}|_{x=0} = X_{+,i}, \quad v_1 + u > 0.$$

Similarly, for each $1 \leq j \leq \nu_0$, denote $g_{0,j}$ as the solution to (2.18) where

$$g_{0,j}|_{x=0} = X_{0,j}, \quad v_1 + u > 0.$$

Let C be the block matrix defined by

$$C = \begin{pmatrix} C_{++} & C_{+0} \\ C_{0+} & C_{00} \end{pmatrix}, \tag{3.9}$$

where

$$\begin{aligned} C_{++,ii'} &= \langle (v_1 + u)X_{+,i}, g_{+,i'} \rangle|_{x=0}, & C_{+,ij'} &= \langle (v_1 + u)X_{+,i}, g_{0,j'} \rangle|_{x=0}, \\ C_{0+,ji'} &= \langle (v_1 + u)X_{0,j}, g_{+,i'} \rangle|_{x=0}, & C_{00,jj'} &= \langle (v_1 + u)X_{0,j}, g_{0,j'} \rangle|_{x=0} \end{aligned}$$

for $1 \leq i, i' \leq \nu_+$ and $1 \leq j, j' \leq \nu_0$. In the case where $\dim H^0 = 0$, we have

$$C = C_{++}. \tag{3.10}$$

The main property we will show about C is that C is non-singular. This will be an easy consequence of the following lemma:

Lemma 3.5. *Let f be a solution to the damped equation (3.5) and $\mathbf{U}(f)$ be defined as in (3.8). Suppose*

$$\mathbf{U}_+(f) = \mathbf{U}_0(f) = 0, \quad \text{at } x = 0. \tag{3.11}$$

Then $\mathbf{U}(f) = 0$ for all x .

Proof. We separate the proof in two parts according to $\dim H^0$.

Case 1. $\dim(H^0) = 0$. In this case condition (3.11) reduces to

$$\mathbf{U}_+(f) = 0, \quad \text{at } x = 0. \tag{3.12}$$

Moreover, the damped equation (3.5) reduces to

$$\begin{aligned} (v_1 + u)\partial_x f + \mathcal{L}f + \alpha \sum_{k=1}^{\nu_+} (v_1 + u)X_{+,k} \langle (v_1 + u)X_{+,k}, f \rangle_v \\ + \alpha \sum_{k=1}^{\nu_-} (v_1 + u)X_{-,k} \langle (v_1 + u)X_{-,k}, f \rangle_v = 0, \end{aligned} \tag{3.13}$$

and $\mathbf{U}(f)$ becomes

$$\mathbf{U}(f) = (\mathbf{U}_+^T(f), \mathbf{U}_-^T(f))^T.$$

Multiplying (3.13) by $X_{+,k}, X_{-,j}$ and integrating over $v \in \mathbb{V}$, we obtain a linear system for \mathbf{U} :

$$\partial_x \mathbf{U} + A_1 \mathbf{U} = 0, \tag{3.14}$$

where the coefficient matrix is diagonal:

$$A_1 = \left(\begin{array}{c|c} \alpha D_+ & 0 \\ \hline 0 & -\alpha D_- \end{array} \right), \tag{3.15}$$

where D_+, D_- are positive definite and

$$D_+ = \text{diag}(\gamma_{+,1}, \dots, \gamma_{+,\nu_+}), \quad D_- = \text{diag}(\gamma_{-,1}, \dots, \gamma_{-,\nu_-}),$$

where $\gamma_{\pm,k} > 0$ are defined as in (2.13). Since solutions to (3.13) are in $(L^2(dv dx))^m$, it is clear that

$$\langle (v_1 + u)X_{-,j}, f(0, \cdot) \rangle_v = \langle (v_1 + u)X_{-,j}, f(x, \cdot) \rangle_v = 0, \quad \text{for all } 1 \leq j \leq \nu_- \text{ and } x \geq 0.$$

Hence $\mathbf{U}_-(f) = 0$ holds for all x . Moreover, by the structure of A_1 in (3.15) and the initial condition (3.12), we have $\mathbf{U}_+(f) = 0$ for all x . Thus $\mathbf{U}(f) = 0$ for all x .

Case 2. $\dim(H^0) \neq 0$. In this case, we multiply $X_{+,j}, X_{-,i}, X_{0,k}, \mathcal{L}^{-1}(v_1 X_{0,m})$ to (3.5) and integrate over $v \in \mathbb{V}$. This gives

$$\partial_x \mathbf{U} + A_2 \mathbf{U} = 0, \tag{3.16}$$

where the coefficient matrix A_2 is

$$A_2 = \left(\begin{array}{cc|c|c} \alpha D_+ & & 0 & \alpha A_{21} \\ & -\alpha D_- & & \alpha A_{22} \\ \hline & 0 & 0 & \alpha B \\ \hline \alpha A_{21}^T & \alpha A_{22}^T & I + \alpha B & \alpha D \end{array} \right), \tag{3.17}$$

where again D_{\pm} are positive diagonal matrices such that

$$D_+ = \text{diag}(\gamma_{+,1}, \dots, \gamma_{+,\nu_+})_{\nu_+ \times \nu_+}, \quad D_- = \text{diag}(\gamma_{-,1}, \dots, \gamma_{-,\nu_-})_{\nu_- \times \nu_-}.$$

The other matrices are

$$\begin{aligned} A_{21,ik} &= (\langle (v_1 + u)X_{+,i}, \mathcal{L}^{-1}((v_1 + u)X_{0,k}) \rangle_v)_{\nu_+ \times \nu_0}, \\ A_{22,jk} &= (\langle (v_1 + u)X_{-,j}, \mathcal{L}^{-1}((v_1 + u)X_{0,k}) \rangle_v)_{\nu_- \times \nu_0}, \\ B_{ij} &= \langle (v_1 + u)X_{0,i}, \mathcal{L}^{-1}((v_1 + u)X_{0,j}) \rangle_{v, \nu_0 \times \nu_0}, \\ D_{ij} &= \langle (v_1 + u)\mathcal{L}^{-1}((v_1 + u)X_{0,i}), \mathcal{L}^{-1}((v_1 + u)X_{0,j}) \rangle_{v, \nu_0 \times \nu_0}, \end{aligned}$$

where B is symmetric positive definite and D is symmetric. Note that if we define

$$Q = \left(\begin{array}{c|cc} I & & 0 \\ \hline I & & 0 \\ \hline 0 & (\alpha B)^{1/2}(I + \alpha B)^{-1/2} & 0 \\ \hline 0 & & 0 \\ \hline & & I \end{array} \right),$$

and

$$\tilde{A}_2 = \left(\begin{array}{cc|cc} \alpha D_+ & & & \\ \hline -\alpha D_- & & 0 & \\ \hline 0 & & 0 & \\ \hline \alpha A_{21}^T & \alpha A_{22}^T & (I + \alpha B)^{1/2}(\alpha B)^{1/2} & \\ \hline & & & \alpha D \end{array} \right).$$

Then

$$A_2 = Q^{-1}\tilde{A}_2Q. \tag{3.18}$$

Thus A_2 and \tilde{A}_2 have the same signature. In particular, they have the same number of negative eigenvalues. Now we count the number of negative eigenvalues of \tilde{A}_2 . Let

$$P = \left(\begin{array}{ccc|cc} & I & & 0 & 0 \\ \hline & & I & 0 & 0 \\ \hline & 0 & & I & 0 \\ \hline -A_{21}^T D_+^{-1} & A_{22}^T D_-^{-1} & & 0 & I \end{array} \right).$$

Then P is non-singular and

$$A_3 = P\tilde{A}_2P^T = \left(\begin{array}{cc|cc} \alpha D_+ & & & \\ \hline -\alpha D_- & & 0 & \\ \hline 0 & & 0 & \\ \hline 0 & & (I + \alpha B)^{1/2}(\alpha B)^{1/2} & \\ \hline & & & \alpha D_1 \end{array} \right),$$

where D_1 is symmetric and

$$D_1 = D - A_{21}^T D_+^{-1} A_{21} + A_{22}^T D_-^{-1} A_{22}.$$

By Sylvester's law of inertia, the matrices \tilde{A}_2 and A_3 , thus A_2 and A_3 , have the same number of negative eigenvalues. The total number of negative eigenvalues of A_3 is determined by that of the submatrix $\begin{pmatrix} 0 & (I + \alpha B)^{1/2}(\alpha B)^{1/2} \\ (I + \alpha B)^{1/2}(\alpha B)^{1/2} & \alpha D_1 \end{pmatrix}$. Define

$$P_1 = \begin{pmatrix} (I + \alpha B)^{-1/4}(\alpha B)^{-1/4} & 0 \\ 0 & (I + \alpha B)^{-1/4}(\alpha B)^{-1/4} \end{pmatrix}.$$

Then

$$A_4 = P_1 \begin{pmatrix} 0 & (I + \alpha B)^{1/2}(\alpha B)^{1/2} \\ (I + \alpha B)^{1/2}(\alpha B)^{1/2} & \alpha D_1 \end{pmatrix} P_1^T = \begin{pmatrix} 0 & I \\ I & \alpha D_2 \end{pmatrix},$$

where

$$D_2 = (I + \alpha B)^{-1/4}(\alpha B)^{-1/4}D_1(I + \alpha B)^{-1/4}(\alpha B)^{-1/4}.$$

Note that D_2 is symmetric. Hence, D_2 has a complete set of eigenvectors. Let $(\lambda, \mathbf{E}) = (\lambda, (\mathbf{e}_1, \mathbf{e}_2)^T)$ be an eigenpair of A_4 such that

$$\begin{pmatrix} 0 & I \\ I & \alpha D_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix}. \tag{3.19}$$

This is equivalent to

$$\mathbf{e}_2 = \lambda \mathbf{e}_1, \quad \mathbf{e}_1^T + \alpha D_2 \mathbf{e}_2^T = \lambda \mathbf{e}_2^T.$$

Note that $\lambda \neq 0$. Since D_2 is symmetric, it has a complete set of orthogonal eigenvectors. Let \mathbf{e} be an arbitrary eigenvector of D with eigenvalue $\lambda_{\mathbf{e}}$ and take $\mathbf{e}_2 = \mathbf{e}$. Then

$$\mathbf{e}_1 = \frac{1}{\lambda} \mathbf{e}, \quad \frac{1}{\lambda} \mathbf{e}^T + \alpha D_2 \mathbf{e}^T = \lambda \mathbf{e}^T. \tag{3.20}$$

Thus

$$\frac{1}{\lambda} + \alpha \lambda_{\mathbf{e}} - \lambda = 0, \tag{3.21}$$

which has exactly one negative solution for λ . Since the set of eigenvectors of D_2 is complete, the matrix A_4 has exactly ν_0 negative eigenvalues. Together with D_- , we have that A_3 , thus A_2 , has exactly $\nu_- + \nu_0$ negative eigenvalues, which prescribes $\nu_- + \nu_0$ conditions on \mathbf{U} such that

$$\mathbf{E}_k \cdot \mathbf{U}(x) = 0, \quad 1 \leq k \leq \nu_- + \nu_0, \quad x \geq 0, \tag{3.22}$$

where \mathbf{E}_k are the eigenvectors associated with negative eigenvalues. Write each \mathbf{E}_k as

$$\mathbf{E}_k = (\mathbf{e}_{k,+}, \mathbf{e}_{k,-}, \mathbf{e}_{k,0}, \mathbf{e}_{k,\mathcal{L},0})^T$$

and define the matrix \mathbf{E} by

$$\mathbf{E} = \begin{pmatrix} \mathbf{e}_{1,-} & & \mathbf{e}_{1,\mathcal{L},0} \\ & \dots & \\ \mathbf{e}_{\nu_- + \nu_0,-} & & \mathbf{e}_{\nu_- + \nu_0,\mathcal{L},0} \end{pmatrix}_{(\nu_- + \nu_0) \times (\nu_- + \nu_0)}$$

By (3.11) we have

$$\mathbf{E} \begin{pmatrix} \mathbf{U}_- \\ \mathbf{U}_{\mathcal{L},0} \end{pmatrix} = 0, \quad \text{at } x = 0. \tag{3.23}$$

Now we show that \mathbf{E} is nonsingular. Suppose not. Let \mathcal{N}_2 be the space spanned by the eigenvectors of A_2 with negative eigenvalues. Then there exists a nontrivial vector in \mathcal{N}_2 which takes the form

$$\widehat{\mathbf{E}} = (\widehat{\mathbf{e}}_+, 0, \widehat{\mathbf{e}}_0, 0)^T.$$

By (3.18), if $\mathbf{E} = (\mathbf{e}_+, \mathbf{e}_-, \mathbf{e}_0, \mathbf{e}_{\mathcal{L},0})^T$ is an eigenvector of A_2 with eigenvalue λ , then $\mathbf{F} = Q(\mathbf{e}_+, \mathbf{e}_-, \mathbf{e}_0, \mathbf{e}_{\mathcal{L},0})^T$ is an eigenvector of \tilde{A}_2 with the same eigenvalue. By the definition of Q , if we denote

$$\mathbf{F} = (\mathbf{f}_+, \mathbf{f}_-, \mathbf{f}_0, \mathbf{f}_{\mathcal{L},0})^T,$$

then

$$e_- = f_-, \quad e_{\mathcal{L},0} = f_{\mathcal{L},0}.$$

Let \mathcal{QN}_2 as the space spanned by the eigenvectors of \tilde{A}_2 with negative eigenvalues. Then there exists a nontrivial $\widehat{F} \in \mathcal{QN}_2$ such that

$$\widehat{F} = (\widehat{f}_+, 0, \widehat{f}_0, 0)^T.$$

Since \mathcal{QN}_2 is an invariant subspace of \tilde{A}_2 , we have that

$$\tilde{A}_2 \widehat{F} = (\alpha D_+ \widehat{f}_+^T, 0, 0, \widehat{f}_2)^T \in \mathcal{QN}_2$$

where $\widehat{f}_2 = \alpha A_{21}^T \widehat{f}_+^T + (I + \alpha B)^{1/2} (\alpha B)^{1/2})^T \widehat{f}_0^T$. By the symmetry and non-degeneracy of \tilde{A}_2 , the quadratic form given by \tilde{A}_2 on \mathcal{QN}_2 is strictly negative. Therefore,

$$\widehat{F}^T \tilde{A}_2 \widehat{F} = \alpha \widehat{f}_+^T D_+ \widehat{f}_+ \leq 0.$$

Since D_+ is strictly positive definite, we have that $\widehat{f}_+ = 0$ and

$$\widehat{F}^T \tilde{A}_2 \widehat{F} = 0,$$

which implies that $\widehat{F} = 0$. This contradicts the assumption that \widehat{F} is non-trivial. Hence the matrix E is non-singular. By (3.23) we derive that

$$U_-(f) = 0, \quad U_{\mathcal{L},0}(f) = 0, \quad \text{at } x = 0.$$

Together with (3.11), we have the initial data for the ODE (3.16) as $U(f) = 0$ at $x = 0$. Thus the only solution to this ODE is $U(f) = 0$ for all x . □

Using Lemma 3.5 we can now show

Lemma 3.6. *The matrix C defined in (3.9) is non-singular.*

Proof. First we recall [7] the uniqueness property of the solution to (3.29): if f is a solution to (3.29) which satisfies $f \in (L^2(a \, dv \, dx))^m$ and $(v_1 + u) \partial_x f \in (L^2(\frac{1}{a} \, dv \, dx))^m$, then f must be unique. For the convenience of the reader, we brief explain its proof: Suppose h is a solution to the half-space equation (3.29) with incoming data $\phi = 0$. Then $\int_{\mathbb{R}^3} (v_1 + u) h^2 \, dv$ is decreasing in x . Since there exists $h_\infty \in H^+ \oplus H^0$ such that $h - h_\infty \in (L^2(dv \, dx))^m$, we can find a sequence x_k such that

$$\int_{\mathbb{R}^3} (v_1 + u) h^2(x_k, v) \, dv \rightarrow \int_{\mathbb{R}^3} (v_1 + u) h_\infty^2(v) \, dv \geq 0.$$

Hence $\int_{\mathbb{R}^3} (v_1 + u) h^2(x, v) \, dv \geq 0$ for all $x \geq 0$. This holds in particular at $x = 0$. Since the incoming data is zero at $x = 0$, the outgoing data at $x = 0$ must also be zero and $\int_{\mathbb{R}^3} (v_1 + u) h^2(x_k, v) \, dv = 0$ for all $x \geq 0$. The conservation property of the half-space equation then implies that $h(x, \cdot) \in (\text{Null } \mathcal{L})^\perp$. By multiplying the equation by h and integrate over v , we have $\langle h, Lh \rangle = 0$ for all $x \geq 0$. Hence $\mathcal{P}h = 0$ for all $x \geq 0$ by the spectral gap of \mathcal{L} in (A4). Therefore $h \equiv 0$ and the solution to the half-space equation is unique.

Now suppose C is singular. Then there exist constants

$$(\eta_{+,1}, \dots, \eta_{+,\nu_+}, \eta_{0,1}, \dots, \eta_{0,\nu_0}) \neq 0$$

such that we can find incoming data

$$\phi_g = \sum_{j=1}^{\nu_+} \eta_{+,j} X_{+,j} + \sum_{k=1}^{\nu_0} \eta_{0,k} X_{0,k}$$

that gives rise to a solution g satisfying that

$$\begin{aligned} \langle (v_1 + u)X_{+,1}, g \rangle_v = \dots = \langle (v_1 + u)X_{+,\nu_+}, g \rangle_v = 0, \\ \langle (v_1 + u)X_{0,1}, g \rangle_v = \dots = \langle (v_1 + u)X_{0,\nu_0}, g \rangle_v = 0, \end{aligned} \quad \text{at } x = 0. \tag{3.24}$$

By Lemma 3.5, we have

$$\mathbf{U}(g) = 0 \quad \text{for all } x. \tag{3.25}$$

Thus the solution g satisfies both the damped and the original half-space equation (1.1) with the end-state $g_\infty = 0$. By the uniqueness of solutions to (1.1), we have $\eta_{+,1} = \dots = \eta_{+,\nu_+} = \eta_{0,1} = \dots = \eta_{0,\nu_0} = 0$ which is a contraction. Thus C must be non-singular. \square

Now we state and prove the main recovery theorem.

Proposition 3.7 *Recovery.* Let $\phi \in (L^2(a(v)\mathbf{1}_{v_1+u>0} dv))^m$ and f be the solution to the damped equation (3.5) with incoming data ϕ . Let $C, g_{+,i}, g_{0,j}$ be the matrix and the family of auxiliary functions defined in (3.10) and (3.9). Define the coefficient vector $\eta = (\eta_{+,1}, \dots, \eta_{+,\nu_+}, \eta_{0,1}, \dots, \eta_{0,\nu_0})^T$ such that

$$\eta = C^{-1}(\mathbf{U}_+(f), \mathbf{U}_0(f))^T \Big|_{x=0} \tag{3.26}$$

and

$$g = \sum_{i=1}^{\nu_+} \eta_{+,i} g_{+,i} + \sum_{j=1}^{\nu_0} \eta_{0,j} g_{0,j}, \quad \Phi = \sum_{i=1}^{\nu_+} \eta_{+,i} X_{+,i} + \sum_{j=1}^{\nu_0} \eta_{0,j} X_{0,j}. \tag{3.27}$$

Define

$$f_\phi = f - g + \Phi = f - \sum_{i=1}^{\nu_+} \eta_{+,i} (g_{+,i} - X_{+,i}) - \sum_{j=1}^{\nu_0} \eta_{0,j} (g_{0,j} - X_{0,j}). \tag{3.28}$$

Then f_ϕ is the unique solution to the half-space equation

$$\begin{aligned} (v_1 + u)\partial_x f_\phi + \mathcal{L}f_\phi &= 0, \\ f_\phi|_{x=0} &= \phi(v), \quad v_1 + u > 0, \\ f_\phi - f_{\phi,\infty} &\in L^2(dx; L^2(dv))^m, \end{aligned} \tag{3.29}$$

where $f_{\phi,\infty} \in H^+ \oplus H^0$ is the end-state given by

$$f_{\phi,\infty} = \sum_{j=1}^{\nu_+} \eta_{+,j} X_{+,j} + \sum_{k=1}^{\nu_0} \eta_{0,k} X_{0,k}.$$

Proof. We directly show that f_ϕ satisfies (3.29). First, by the definitions of $g_{+,i}, g_{0,j}$, we have $f_\phi|_{x=0} = \phi(v)$ for $v_1 + u > 0$. Second, it follows from the definition in (3.27) that g , thus $f - g$, are both solutions to the damped equation (3.5). By the definition of η we have

$$\mathbf{U}_+(f - g) = 0, \quad \mathbf{U}_0(f - g) = 0.$$

Hence by Lemma 3.5, we have $\mathbf{U}(f - g) = 0$. This shows $f - g$ is in fact a solution to the undamped equation (3.29). Since every $X_{+,i}$ and $X_{0,j}$ are solutions to (3.29), we have f_ϕ as a solution to (3.29). \square

4. GALERKIN APPROXIMATION AND NUMERICAL SCHEME

Let us now use the variational formulation (2.25) to design a Galerkin method to approximate the solution to the damped equation (3.5). There are two parts in this section: first we show the construction of the finite-dimensional approximation and its error estimate. Then we transform the finite-dimensional variational form into an ODE system which will set base for our numerical scheme.

4.1. Galerkin approximation

First we use both parts of the Babuška–Aziz lemma to show the validity of the Galerkin approximation and its quasi-optimality.

Proposition 4.1 Approximations in \mathbb{R}^3 . Suppose $\{\psi_n^{(1)}\}_{n=1}^\infty$ is an orthonormal basis of $L^2(dv_1)$ such that

- $\psi_{2n-1}^{(1)}(v_1)$ is odd and $\psi_{2n}^{(1)}(v_1)$ is even in v_1 with respect to $-u$ for any $n \geq 1$;
- $(v_1 + u)\psi_{2n}^{(1)}(v_1) \in \text{span}\{\psi_1^{(1)}, \dots, \psi_{2n+1}^{(1)}\}$ for each $n \geq 1$.

Suppose $\{\psi_n^{(2)}\}, \{\psi_n^{(3)}\}_{n=1}^\infty$ are orthonormal bases for $L^2(dv_2)$ and $L^2(dv_3)$ respectively. Define the closed subspace Γ_{NK} as

$$\Gamma_{NK} = \left\{ g(x, v) \in \Gamma \mid g(x, v) = \sum_{i=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} g_{kln}^{(i)}(x) \psi_k^{(1)}(v_1) \psi_l^{(2)}(v_2) \psi_n^{(3)}(v_3) \mathbf{e}_i, g_{kln}^{(i)} \in H^1(dx) \right\},$$

where $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ is the standard i^{th} basis vector of \mathbb{R}^m with $1 \leq i \leq m$. Then

(a) there exists a unique $f_{NK} \in \Gamma_{NK}$ such that

$$f_{NK}(x, v) = \sum_{i=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} a_{kln}^{(i)}(x) \psi_k^{(1)}(v_1) \psi_l^{(2)}(v_2) \psi_n^{(3)}(v_3) \mathbf{e}_i, \tag{4.1}$$

which satisfies

$$\mathcal{B}(f_{NK}, g) = l(g) \quad \text{for every } g \in \Gamma_{NK}, \tag{4.2}$$

where \mathcal{B} and l for the damped equation and are defined in (2.24) and (2.26) respectively. The coefficients $\{a_{kln}^{(i)}(x)\}$ satisfy

$$a_{kln}^{(i)}(\cdot) \in C^1[0, \infty) \cap H^1(0, \infty), \quad 1 \leq k \leq 2N + 1, 1 \leq l, n \leq K, 1 \leq i \leq m.$$

(b) There exists a constant C_0 such that

$$\|f - f_{NK}\|_\Gamma \leq C_0 \inf_{w \in \Gamma_{NK}} \|f - w\|_\Gamma,$$

where $\|\cdot\|_\Gamma$ is the norm defined in (2.20).

Proof. Both (a) and (b) directly follow from the Babuška–Aziz lemma as long as we verify the inf-sup condition of \mathcal{B} on the finite-dimensional subspace Γ_{NK} . Since it is similar as the continuum case in Proposition 3.2, we only explain the modification in choosing the test functions ψ_1 and ψ_2 . For any $f \in \Gamma_N$, we choose

$$\psi_1 = f, \quad \psi_2 = \mathcal{P}_N \left(\frac{1}{(1 + |v_1 + u| + |v_2| + |v_3|)^{\omega_0}} (v_1 + u) \partial_x f^+ \right),$$

where $\mathcal{P}_N : (L^2(dv))^m \rightarrow \Gamma_N$ is the projection onto Γ_N . The rest of the estimates are similar to the proof in Section 3, and thus omitted. \square

Since our numerical examples are both in one-dimension for a single species, we apply Proposition 4.1 to $\mathbb{V} \subseteq \mathbb{R}^1$ and $m = 1$ to obtain the following corollary for two special cases:

Corollary 4.2 Approximations in \mathbb{R}^1 . Let $\mathbb{V} = \mathbb{R}^1$ or $\mathbb{V} = [-1, 1]$. Let $u \in \mathbb{R}$ be arbitrary if $\mathbb{V} = \mathbb{R}^1$ and $u = 0$ if $\mathbb{V} = [-1, 1]$. Suppose $\{\psi_n\}_{n=1}^\infty$ is an orthonormal basis of $L^2(dv)$ such that

- ψ_{2n-1} is odd and ψ_{2n} is even in v with respect to $-u$ for any $n \geq 1$;
- $(v + u)\psi_{2n}(v) \in \text{span}\{\psi_1, \dots, \psi_{2n+1}\}$ for each $n \geq 1$.

Define the closed subspace Γ_N as

$$\Gamma_N = \left\{ g(x, v) \in \Gamma \mid g(x, v) = \sum_{k=1}^{2N+1} g_k(x)\psi_k(v), \quad g_k \in H^1(dx) \right\}.$$

Then there exists a unique $f_N \in \Gamma_N$ such that

$$f_N(x, v) = \sum_{k=1}^{2N+1} a_k(x)\psi_k(v), \quad a_k(x) \in C^1[0, \infty), \quad 1 \leq k \leq 2N + 1, \tag{4.3}$$

which satisfies

$$\mathcal{B}(f_N, g) = l(g), \quad \text{for every } g \in \Gamma_N, \tag{4.4}$$

where \mathcal{B} and l are defined in (2.24) and (2.26) respectively.

The approximate solution to the undamped solution is constructed similarly as for the continuous case: let $C, g_{+,i}, g_{0,j}$ be the same matrix and auxiliary functions as in (3.10) and (3.9). Let $g_{+,NK}^{(i)}, g_{0,NK}^{(j)}$ be the Galerkin approximate solutions to $g_{+,i}$ and $g_{0,j}$ respectively. Let

$$\begin{aligned} \eta_{NK} &= \left(\eta_{+,NK}^{(1)}, \dots, \eta_{+,NK}^{(\nu_+)}, \eta_{0,NK}^{(1)}, \dots, \eta_{0,NK}^{(\nu_0)} \right)^T = C^{-1}(\mathbf{U}_+(f_{NK}), \mathbf{U}_0(f_{NK}))^T \Big|_{x=0}, \\ g_{NK} &= \sum_{i=1}^{\nu_+} \eta_{+,NK}^{(i)} g_{+,NK}^{(i)} + \sum_{j=1}^{\nu_0} \eta_{0,NK}^{(j)} g_{0,NK}^{(j)}, \quad \Phi_{NK} = \sum_{i=1}^{\nu_+} \eta_{+,NK}^{(i)} X_{+,i} + \sum_{j=1}^{\nu_0} \eta_{0,NK}^{(j)} X_{0,j}, \end{aligned} \tag{4.5}$$

Let f_ϕ be the solution to the undamped half-space equation (3.29). Define its approximation $f_{\phi,NK}$ as

$$f_{\phi,NK} = f_\phi - g_{NK} + \Phi_{NK}, \tag{4.6}$$

which is an analog of the continuous version in (3.28). The following proposition shows the above approximation is almost quasi-optimal with a correction term.

Proposition 4.3. Let f_ϕ be the solution to the undamped half-space equation (3.29). Suppose $f_{\phi,NK}$ is constructed as in (4.6). Suppose f_ϕ is the unique solution to the equation (2.6). Then there exists a constant C_0 such that

$$\|f_\phi - f_{\phi,NK}\|_\Gamma \leq C_0 \left(\inf_{w \in \Gamma_N} \|f_\phi - w\|_\Gamma + \inf_{w \in \Gamma_N} \|f - w\|_\Gamma + \delta_N \|f\|_{(L^2(a dv dx))^m} \right),$$

where $\|\cdot\|_\Gamma$ is the norm defined in (2.20) and

$$\delta_N := \sum_{i=1}^{\nu_+} \inf_{w \in \Gamma_N} \|g_{+,i} - w\|_\Gamma + \sum_{j=1}^{\nu_0} \inf_{w \in \Gamma_N} \|g_{0,j} - w\|_\Gamma.$$

Proof. Let f be the solution the damped equation (3.5) with incoming data ϕ . Let g, Φ be defined as in (3.27). Then there exist constants $\kappa_4, \tilde{\kappa}_4 > 0$ such that

$$\begin{aligned} \|\Phi - \Phi_{NK}\|_\Gamma &= \|\Phi - \Phi_{NK}\|_{(L^2(dv))^m} \\ &\leq \kappa_4 \|f - f_N\|_\Gamma + \tilde{\kappa}_4 \|f\|_{(L^2(dv dx))^m} \left(\sum_{i=1}^{\nu_+} \|g_{+,i} - g_{+,NK}^{(i)}\|_\Gamma + \sum_{j=1}^{\nu_0} \|g_{0,j} - g_{0,NK}^{(j)}\|_\Gamma \right), \end{aligned}$$

Second, since $f - g$ is a solution to the damped equation, we have

$$\|(f - g) - (f_N - g_{NK})\|_\Gamma \leq \kappa_5 \inf_{w \in \Gamma_{NK}} \|w - (f - g)\|_\Gamma,$$

since $f_{NK}, g_{NK} \in \Gamma_{NK}$. Therefore,

$$\begin{aligned} \|(f - g + \Phi) - (f_{NK} - g_{NK} + \Phi_{NK})\|_\Gamma &\leq \kappa_5 \inf_{w \in \Gamma_{NK}} \|w - (f - g)\|_\Gamma + \|\Phi - \Phi_{NK}\|_\Gamma \\ &\leq \kappa_5 \inf_{w \in \Gamma_{NK}} \|w - (f - g + \tilde{\Phi})\|_\Gamma + \kappa_4 \|f - f_{NK}\|_\Gamma \\ &\leq \kappa_6 \left(\inf_{w \in \Gamma_{NK}} \|f_\phi - w\|_\Gamma + \inf_{w \in \Gamma_{NK}} \|f - w\|_\Gamma + \delta_{NK} \|f\|_{(L^2(dv dx))^m} \right), \end{aligned}$$

where

$$\delta_{NK} = \sum_{i=1}^{\nu_+} \inf_{w \in \Gamma_{NK}} \|g_{+,i} - w\|_\Gamma + \sum_{j=1}^{\nu_0} \inf_{w \in \Gamma_{NK}} \|g_{0,j} - w\|_\Gamma.$$

Note that the second inequality holds because $\Phi \in H^+ \oplus H^0 \subseteq \Gamma_{NK}$. □

Remark 4.4. Note that in the above reconstruction scheme, the solutions $g_{+,NK}^{(i)}$ for $1 \leq i \leq \nu_+$ and $g_{0,NK}^{(j)}$ for $1 \leq j \leq \nu_0$ can be precomputed, as they do not depend on the prescribed incoming data ϕ . In particular, we can use a higher order approximation (larger N, K) for these functions.

4.2. ODE formulation

In this part we reformulate the variational form (4.2) into an ODE with explicit boundary conditions. This ODE will be the system that we solve in numerics; since this is a linear ODE, its solution can be directly obtained by solving the associated generalized eigenvalue problems. To illustrate the idea, we first treat the special case where there is a single species in 1D, that is, $m = K = 1$.

Proposition 4.5. *The variational form (4.4) is equivalent to the following ODE for the coefficients $a_k(x)$ together with the boundary conditions at $x = 0$:*

$$\sum_{k=1}^{2N+1} A_{kl} \partial_x a_k(x) = \sum_{k=1}^{2N+1} B_{kl} a_k(x), \tag{4.7}$$

$$\sum_{k=1}^{N+1} \langle (v+u)\psi_{2k-1}, \psi_{2j} \rangle_v a_{2k-1}(0) + \sum_{k=1}^N \langle |v+u|\psi_{2k}, \psi_{2j} \rangle_v a_{2k}(0) = 2 \int_{v+u>0} (v_1+u) \phi \psi_{2j} dv, \tag{4.8}$$

where $1 \leq j \leq N$ and

$$A_{kl} = \langle (v+u)\psi_k, \psi_l \rangle_v, \quad B_{kl} = -\langle \psi_k, \mathcal{L}_d \psi_l \rangle_v, \quad 1 \leq i, j \leq 2N+1. \tag{4.9}$$

Proof. In order to show that the boundary conditions for the solution to (4.4) are given by (4.8), we first choose test functions $G_{2j}(x, v) = g(x)\psi_{2j}(v)$ where $g(x) \in C_c^\infty([0, \infty))$ and $1 \leq j \leq N$. Applying G_j in (4.4), we get

$$-\langle f_N^-, (v_1 + u)\psi_{2j}(v)\partial_x g(x) \rangle_{x,v} + \langle (\mathcal{L}\psi_{2j})g(x), f_N \rangle_{x,v} = 0, \tag{4.10}$$

where f_N is defined in (4.3) and $f_N = f_N^- + f_N^+$ with

$$f_N^- = \sum_{k=1}^{N+1} a_{2k-1}(x)\psi_{2k-1}, \quad f_N^+ = \sum_{k=1}^N a_{2k}(x)\psi_{2k}.$$

By integration by parts in (4.10) we obtain

$$\left\langle \sum_{k=1}^{N+1} \psi_{2k-1} \partial_x a_{2k-1}(x), (v_1 + u)\psi_{2j}(v)g(x) \right\rangle_{x,v} + \langle (\mathcal{L}\psi_{2j})g(x), f_N \rangle_{x,v} = 0.$$

Since $g \in C_c^\infty([0, \infty))$ is arbitrary, we have

$$\sum_{k=1}^{N+1} \langle \psi_{2k-1}, (v_1 + u)\psi_{2j}(v) \rangle_v \partial_x a_{2k-1}(x) + \langle (\mathcal{L}\psi_{2j}), f_N \rangle_v = 0, \tag{4.11}$$

for each $1 \leq j \leq N$ and $x \in [0, \infty)$. Note we choose $\tilde{G}_{2j} = \tilde{g}(x)\psi_{2j}(x)$ where $\tilde{g} \in C^\infty([0, \infty))$. Then equation (4.4) becomes

$$\begin{aligned} & -\langle f_N^-, (v + u)\psi_{2j}(v)\partial_x g(x) \rangle_{x,v} + \langle (\mathcal{L}\psi_{2j})g(x), f_N \rangle_{x,v} + \langle (v_1 + u)f_N^+, \psi_{2j}\tilde{g}(0) \rangle_{x=0} \\ & = 2 \int_{v_1+u>0} (v_1 + u)\phi(v)\psi_{2j}(v)g(0) dv, \end{aligned} \tag{4.12}$$

for each $1 \leq j \leq N$. The set of N boundary conditions (4.8) then follows from integrating by parts in (4.12) and applying (4.11). \square

The general case follows from the similar idea and we only sketch its proof.

Proposition 4.6. *Let*

$$A = \left(\left\langle (v_1 + u)\psi_k^{(1)}, \psi_j^{(1)} \right\rangle_{v_1} \right)_{(2N+1) \times (2N+1)}.$$

Define two 8-tensors \mathfrak{A} and \mathfrak{B} as

$$\begin{aligned} \mathfrak{A} &= A \otimes I \otimes I \otimes I = (A_{ik}\delta_{lj}\delta_{ns}\delta_{pq})_{(2N+1)^2 \times K^2 \times K^2 \times m^2}, \\ \mathfrak{B}_{klnp}^{ijsq} &= -\left\langle \psi_k^{(1)}(v_1)\psi_l^{(2)}(v_2)\psi_n^{(3)}(v_3)\mathbf{e}_p, \mathcal{L}_d \left(\psi_i^{(1)}(v_1)\psi_j^{(2)}(v_2)\psi_s^{(3)}(v_3)\mathbf{e}_q \right) \right\rangle_v \end{aligned} \tag{4.13}$$

for $1 \leq i, k \leq 2N + 1$, $1 \leq j, l \leq K$, $1 \leq s, n \leq K$, and $1 \leq p, q \leq m$. Then the variational form (4.2) is equivalent to the following ODE for the coefficients $a_{kln}^{(p)}(x)$:

$$\sum_{p=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} \mathfrak{A}_{klnp}^{ijsq} \partial_x a_{kln}^{(p)}(x) = \sum_{p=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} \mathfrak{B}_{klnp}^{ijsq} a_{kln}^{(p)}(x), \tag{4.14}$$

together with the boundary conditions at $x = 0$:

$$\begin{aligned} & \sum_{k=1}^{N+1} \left\langle (v_1 + u)\psi_{2k-1}^{(1)}, \psi_{2i}^{(1)} \right\rangle_{v_1} a_{2k-1,jl}^{(q)}(0) + \sum_{k=1}^N \left\langle |v_1 + u|\psi_{2k}^{(1)}, \psi_{2i}^{(1)} \right\rangle_{v_1} a_{2k,jl}^{(q)}(0) \\ & = 2 \int_{v_1+u>0} (v_1 + u)\phi \cdot \psi_{2i}^{(1)}(v_1)\psi_j^{(2)}(v_2)\psi_k^{(3)}(v_3)\mathbf{e}_q dv \end{aligned} \tag{4.15}$$

for $i = 1, \dots, N$, $j, l = 1, 2, \dots, K$, and $q = 1, \dots, m$.

Proof. Equation (4.14) is obtained by choosing the test function g in (4.2) as the basis functions such that the velocity part is $\psi_i^{(1)}(v_1)\psi_j^{(2)}(v_2)\psi_l^{(3)}(v_3)\mathbf{e}_q$. The boundary condition (4.15) is derived by choosing the test functions as $\psi_{2i}^{(1)}(v_1)\psi_j^{(2)}(v_2)\psi_l^{(3)}(v_3)\mathbf{e}_q$. \square

Numerically, the approximate solutions f_{NK} in (4.14) (or f_N (4.7) in 1D) will be solved using the method of generalized eigenvalues. In particular, we define the generalized eigenvalues and its associated eigen-tensor for $(\mathfrak{A}, \mathfrak{B})$ as $\lambda \in \mathbb{R}$ and $\eta = (\eta_{kln}^{(p)})_{(2N+1) \times K \times K \times m}$ such that

$$\mathfrak{A}\eta = \sum_{p=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} \mathfrak{A}_{klnp}^{ijsq} \eta_{kln}^{(p)} = \lambda \sum_{p=1}^m \sum_{l,n=1}^K \sum_{k=1}^{2N+1} \mathfrak{B}_{klnp}^{ijsq} \eta_{kln}^{(p)} \tag{4.16}$$

for all $1 \leq i \leq 2N + 1$, $1 \leq j, s \leq K$, and $1 \leq q \leq m$. When reduced to 1D system, the generalized eigenvalue problem for (\mathbf{A}, \mathbf{B}) becomes

$$\mathbf{A}\eta = \lambda \mathbf{B}\eta. \tag{4.17}$$

To solve for the coefficient $a(x)$, we take (4.7) as an example. Define $\gamma(x) = \eta^T \mathbf{B} a(x)$ and multiply (4.7) by η^T from the left. We then obtain the equation for γ as

$$\eta^T \mathbf{A} \partial_x a(x) = \eta^T \mathbf{B} a(x) \quad \Rightarrow \quad \lambda \partial_x \gamma(x) = \gamma(x).$$

If $\lambda = 0$, then we immediately get the constraint

$$\gamma(x) = \eta^T \mathbf{B} a = 0. \tag{4.18}$$

If $\lambda \neq 0$, then we have

$$\gamma(x) = e^{x/\lambda} \gamma(0).$$

Depending on the signs of the eigenvalues, γ either grows exponentially to infinity or decays exponentially to zero; as we look for bounded decaying solutions, this gives us constraints to $\gamma(0)$ for the growing modes: If $\lambda > 0$, then we have the constraints

$$\gamma = \eta^T \mathbf{B} a = 0. \tag{4.19}$$

Note that we do not need constraints for modes with negative eigenvalues. The total number of constraints in the form of (4.19) is determined by the number of positive generalized eigenvalues. The following Proposition gives the signature of (\mathbf{A}, \mathbf{B}) :

Proposition 4.7. *Let $\mathfrak{A}, \mathfrak{B}$ be the 8-tensors defined in (4.13) with any arbitrary $u \in \mathbb{R}$ and $N, K \geq 1$. Then*

- (a) *there are mNK^2 positive generalized eigenvalues, mNK^2 negative eigenvalues, and mK^2 zero eigenvalue for the pair $(\mathfrak{A}, \mathfrak{B})$.*
- (b) *In the special case where $m = K = 1$ and \mathbf{A}, \mathbf{B} be the matrices defined in (4.9) with any arbitrary $u \in \mathbb{R}$ and $N \geq 1$, there are N positive generalized eigenvalues, N negative eigenvalues, and one zero eigenvalue for the pair (\mathbf{A}, \mathbf{B}) .*

Proof. We first verify that in the 1D case, (\mathbf{A}, \mathbf{B}) has N positive, N negative, and one zero generalized eigenvalues. By the definition of \mathbf{B} and the strict coercivity of \mathcal{L}_d , the matrix \mathbf{B} is symmetric and strictly positive definite. Hence the numbers of positive, negative, and zero generalized eigenvalues are the same with the signature of the matrix $\mathbf{B}^{-1}\mathbf{A}$. Furthermore, by the Sylvestre’s Law of Inertia, $\mathbf{B}^{-1}\mathbf{A}$ and \mathbf{A} have the same signature. Hence, we only need to count the numbers of positive, negative, and zero eigenvalues of \mathbf{A} . Note that by the definition of the basis functions ψ_k in (5.4), \mathbf{A} is independent of u since one can perform a change of variable $v + u \rightarrow v$

in each entry in \mathbf{A} . Thus we only need to study the matrix \mathbf{A}_0 with $u = 0$. Change the order of the basis functions such that

$$(\tilde{\psi}_1, \tilde{\psi}_2, \dots, \tilde{\psi}_{N+1}, \tilde{\psi}_{N+2}, \dots, \tilde{\psi}_{2N+1}) = (\psi_1, \psi_3, \dots, \psi_{2N+1}, \psi_2, \dots, \psi_{2N}) = P(\psi_1, \psi_2, \dots, \psi_{2N+1}),$$

where P is the similarity matrix. Defined $\tilde{\mathbf{A}}_0 = P\mathbf{A}_0P^{-1}$. Then $\tilde{\mathbf{A}}_0$ and \mathbf{A}_0 have the same signature. By the even/odd properties of $\tilde{\psi}_i$, the matrix $\tilde{\mathbf{A}}_0$ has the form

$$\tilde{\mathbf{A}}_0 = \begin{pmatrix} 0 & A_1 \\ A_1^T & 0 \end{pmatrix},$$

where $A_1 = (\int_{\mathbb{R}} v\psi_{2i}\psi_{2j+1})_{N \times (N+1)}$. Suppose $\eta = (\eta_{1,1}, \dots, \eta_{1,N}, \eta_{2,1}, \dots, \eta_{2,N+1})^T = (\eta_1^T, \eta_2^T)^T$ is an eigenvector of $\tilde{\mathbf{A}}_0$ with eigenvalue λ . Then one has

$$A_1\eta_2 = \lambda\eta_1, \quad A_1^T\eta_2 = \lambda\eta_1.$$

It is clear that $(\eta_1, -\eta_2)$ is also an eigenvector of $\tilde{\mathbf{A}}_0$ and the associated eigenvalue is $-\lambda$. This shows the eigenvalues of $\tilde{\mathbf{A}}_0$ appear in pairs. Since A_1 has a full rank N , we have that $\text{rank } \tilde{\mathbf{A}}_0 = 2N$. Therefore $\tilde{\mathbf{A}}_0$, thus \mathbf{A}_0 and \mathbf{A} , has N positive eigenvalues, N negative eigenvalues, and one zero eigenvalue.

Now we claim that each generalized eigenpair (λ, v) of \mathbf{A} gives rise to mK^2 eigenpairs of \mathfrak{A} . Indeed, let $\{w^{(l)}\}_{l=1}^K$ be a set of basis vectors of \mathbb{R}^K . Choose the 4-tensor $\eta_i^{(ln)} = v \otimes w^{(l)} \otimes w^{(n)} \otimes \mathbf{e}_i$. Then

$$\mathfrak{A}\eta_i^{(ln)} = (\mathbf{A} \otimes I \otimes I \otimes I)(v \otimes w^{(l)} \otimes w^{(n)} \otimes \mathbf{e}_i) = (\mathbf{A}v) \otimes w^{(l)} \otimes w^{(n)} \otimes \mathbf{e}_i = \lambda\eta_i^{(ln)},$$

for any $1 \leq l, n \leq K$. Thus each $(\lambda, v \otimes w^l \otimes w^{(n)} \otimes \mathbf{e}_i)$ is an eigenpair of \mathfrak{A} .

Note that we can also view \mathfrak{A} and \mathfrak{B} as two matrices of size $(m(2N + 1)K^2) \times (m(2N + 1)K^2)$ by defining a bijection between the indices

$$\mathcal{I} : \{(i, j, l, p) \mid i = 1, \dots, 2N + 1, j, l = 1, \dots, K, p = 1, \dots, m\} \rightarrow \{1, \dots, m(2N + 1)K^2\}.$$

Then \mathfrak{B} is symmetric and positive definite and \mathfrak{A} is symmetric. Therefore, by a similar argument as for (\mathbf{A}, \mathbf{B}) using Sylvestre’s Law of Inertia, the number of positive, negative, and zero generalized eigenvalues agree with those of \mathfrak{A} . This shows there are mNK^2 positive, mNK^2 negative, and mK^2 zero generalized eigenvalues for $(\mathfrak{A}, \mathfrak{B})$. \square

By Proposition 4.7, we outline the specific steps that we take in our numerical computation: in total we have $N + 1$ equations for $a(0)$ given by the constraints (4.18) and (4.19). Combining them with the N equations given by the boundary conditions (4.8) for $a(0)$, we get $2N + 1$ equations for $2N + 1$ unknowns $\{a_k(0)\}$. The linear system (4.7) for a is then uniquely solvable, which further uniquely determines the approximate solution $f_N(x, v)$ by (4.3).

4.3. Numerical scheme

Let us now summarize the numerical algorithm for the half space equation. For simplicity, we present the algorithm for the 1D case and the extension to the higher dimensional cases is similar.

The whole procedure consists of two parts: Compute the damped equation, as shown in Algorithm 1 and recover the solution to the original equation, as presented in Algorithm 2. Computing the damped equation itself has discretization set-up step and computation step.

The first substep in Step I requires constructing $2N + 1$ basis functions. Since it depends on the collision operator, we leave the details to numerical example section where we show basis preparation for the linearized BGK and the transport equation. The forth step in Step I requires the number of non-negative eigenvalue being exactly $N + 1$ and this is guaranteed by Proposition 4.6, which is also used in substep 1 in Step II.

Algorithm 1: Compute the damped equation (2.18).

Data: Boundary condition: $\phi(v)$ for $v > 0$ and the discretization N .

Result: f that solves (2.25), the variational formulation of (2.18).

Step I Set up discretization:

1. Construct $2N + 1$ basis functions.
2. Compute two matrices defined in (4.9).
3. Solve the generalized eigenvalue problem (4.17).
4. Store the $N + 1$ eigenvectors associated with non-negative eigenvalues.

Step II Compute the damped equation, seek for $a(0)$.

1. Find $2N + 1$ equations satisfied by $a(0)$:
 - Use (4.19) to find $N + 1$ equations that projects out positive eigenvectors provided in II.4.
 - Impose the boundary condition (4.8), which provides N equations.
2. Compute $a(0)$.

Step III Assemble f using equation(4.3).

The main cost of the numerical scheme lies in solving the eigenvalue problem (4.17), which scales cubically as N increases. Note that this is a common step for different boundary conditions for the damped equation, and thus only needs to be done once. As we employ a spectral discretization, as shown further in the numerical results, accurate results are obtained even with a small number of basis functions $2N + 1$. Therefore, the computational cost is quite low.

Algorithm 2: Recover the solution to the original equation (3.29).

Data: Boundary condition: $\phi(v)$ for $v > 0$ and the positive modes $X_{+,0}$.

Result: f_ϕ that solves (3.29).

1. Use Algorithm 1 to compute (2.18) using ϕ as the boundary condition.
Denote the solution by f .
2. Use Algorithm 1 to compute (2.18) using $X_{+,0}$ as the boundary conditions.
Denote the solution by $g_{+,0}$.
3. Compute C in (3.9) and U in (3.7).
4. Invert C for η as shown in (3.26).
5. f_ϕ given by (3.28) and f_∞ given by the equation below (3.29).

5. NUMERICAL EXAMPLES

As explained in Section 4.3, the overall strategy to solve the half-space equation consists of two steps: First, we solve for the numerical solution to the half-space damped equation (3.5) using the Galerkin approximation; Second, we recover the undamped solution by Proposition 3.7, which involves the solutions of the damped equation with various boundary conditions in order to obtain the matrix C in the linear system (3.26).

Below we consider the linearized BGK equation and a linear transport equation, both restricted to one dimension and single species (more general cases are studied and presented in [19]). As in Proposition 4.2, for the Galerkin approximation, we specify a set of even and odd functions to form the approximation space Γ_N . The choice of these functions depends on the particular equation under study. By Proposition 4.5, the solution of the approximate system (4.7)–(4.8) is reduced to solving the generalized eigenvalue problem (4.17), where we assemble the matrices A and B using Gaussian quadrature. This will be discussed in more details below.

Our algorithm is implemented in MATLAB. The Gaussian quadrature abscissas and weights are obtained using symbolic calculations in order to guarantee the precision.

5.1. Linearized BGK equation

We first consider the case of one-dimension linearized BGK equation. In this case, the basis functions is constructed using the half-space Hermite polynomials. Those are orthogonal polynomials defined on the positive half v -axis with the weight functions $\exp(-v^2)$: $\{B_n(v), v > 0\}$ such that each $B_n(v)$ is a polynomial of order n and

$$\int_0^\infty B_m(v)B_n(v)e^{-v^2} dv = \delta_{nm}. \tag{5.1}$$

The orthogonal polynomials can be constructed using three term recursion formula (see for example [21]). For completeness we recall some details in Appendix 5.2.

The basis functions ψ_k 's we need are either odd or even with respect to $v = -u$. Hence we shift the functions B_n 's by $-u$ and make even and odd extensions:

$$B_n^E(v) = \begin{cases} B_n(v+u)/\sqrt{2}, & v > -u, \\ B_n(-v-u)/\sqrt{2}, & v < -u. \end{cases} \tag{5.2}$$

$$B_n^O(v) = \begin{cases} B_n(v+u)/\sqrt{2}, & v > -u, \\ -B_n(-v-u)/\sqrt{2}, & v < -u. \end{cases} \tag{5.3}$$

Finally, ψ_k 's are obtained by multiplying these functions by the square root of the Maxwellian: for $n \geq 1$

$$\begin{aligned} \psi_{2n-1} &= B_{n-1}^O e^{-(v+u)^2/2}, \\ \psi_{2n} &= B_{n-1}^E e^{-(v+u)^2/2}. \end{aligned} \tag{5.4}$$

By definition, ψ_{2n-1} is odd, ψ_{2n} is even, and they form a orthonormal basis of $L^2(dv)$. For a fixed n , $(v+u)\psi_{2n}(v)$ is a odd function with respect to $v = -u$. For $v > -u$,

$$(v+u)\psi_{2n}(v) = (v+u)B_{n-1}(v+u)e^{-(v+u)^2/2}/\sqrt{2}.$$

Since $(v+u)B_{n-1}(v+u)$ is a n -th order polynomial in $v+u$, there exists an expansion

$$(v+u)B_{n-1}(v+u) = \sum_{i=0}^n \alpha_i B_i(v+u). \tag{5.5}$$

This yields that

$$(v+u)\psi_{2n}(v) = \sum_{i=0}^n \alpha_i \psi_{2i+1} \in \text{span}\{\psi_1, \dots, \psi_{2n+1}\}. \tag{5.6}$$

Therefore, $\Gamma_N = \text{span}\{\psi_1, \dots, \psi_{2N+1}\}$ satisfies the condition of Proposition 4.2 and the variational formulation (4.4)–(4.8) is well-posed. The $(2N+1) \times (2N+1)$ matrices A and B are then given by

$$A_{ij} = \int_{\mathbb{R}} (v+u)\psi_i\psi_j dv \quad \text{and} \quad B_{ij} = - \int_{\mathbb{R}} \psi_i \mathcal{L}_d \psi_j dv.$$

Note that both matrices are symmetric. The matrix A can be obtained by using the recurrence relation of the orthogonal polynomials. For the matrix B , recall that

$$\begin{aligned} \mathcal{L}\psi_i &= \psi_i - m_i = \psi_i - \chi_0 \int_{\mathbb{R}} \psi_i \chi_0 \, dv - \chi_+ \int_{\mathbb{R}} \psi_i \chi_+ \, dv - \chi_- \int_{\mathbb{R}} \psi_i \chi_- \, dv. \\ \mathcal{L}_d \psi_i &= \mathcal{L}\psi_i + \alpha \sum_{k=1}^{\nu_+} (v+u) X_{+,k} \int_{\mathbb{R}} (v+u) X_{+,k} \psi_i \, dv \\ &\quad + \alpha \sum_{k=1}^{\nu_-} (v+u) X_{-,k} \int_{\mathbb{R}} (v+u) X_{-,k} \psi_i \, dv + \alpha \sum_{k=1}^{\nu_0} (v+u) X_{0,k} \int_{\mathbb{R}} (v+u) X_{0,k} \psi_i \, dv \\ &\quad + \alpha \sum_{k=1}^{\nu_0} (v+u) \mathcal{L}^{-1}((v+u) X_{0,k}) \int_{\mathbb{R}} (v+u) \mathcal{L}^{-1}((v+u) X_{0,k}) \psi_i \, dv. \end{aligned}$$

All the integrals involved in calculating B can be easily made exact up to machine precision by using Gaussian quadrature. For simplicity, let us just focus on

$$\int_{\mathbb{R}} \psi_{2j} \chi_0 \, dv$$

and note that the other integrals share the same structure: the integrand is a product of two polynomials and two Gaussians $e^{-v^2/2}$ and $e^{-(v+u)^2/2}$. To evaluate this type of integral using Gaussian quadrature, we first split the integral into two parts:

$$\int_{\mathbb{R}} \psi_{2j} \chi_0 \, dv = \int_{-u}^{\infty} \psi_{2j} \chi_0 \, dv + \int_{-\infty}^{-u} \psi_{2j} \chi_0 \, dv.$$

Note that ψ_{2j} , on either side of $-u$, is a $(j-1)$ -th order polynomial multiplied by $\exp(-(v+u)^2/2)$, while χ_0 is a quadratic function multiplied with a different weight function $\exp(-v^2/2)$. The product of two Gaussians centered at different locations could be combined into a single Gaussian:

$$\begin{aligned} \int_{-u}^{\infty} \psi_{2j} \chi_0 \, dv &= \frac{\sqrt{2}}{2} \int_{-u}^{\infty} B_{j-1}(v+u) \frac{\chi_0(v)}{e^{-v^2/2}} e^{-\frac{(v+u)^2+v^2}{2}} \, dv \\ &= \frac{\sqrt{2}}{2} e^{-u^2/4} \int_0^{\infty} B_{j-1}(v) \frac{\chi_0(v-u)}{e^{-(v-u)^2/2}} e^{-(v-u/2)^2} \, dv. \end{aligned} \tag{5.7}$$

Similarly, for $v < -u$ we have

$$\begin{aligned} \int_{-\infty}^{-u} \psi_{2j} \chi_0 \, dv &= \frac{\sqrt{2}}{2} \int_{-\infty}^{-u} B_{j-1}(-v-u) \frac{\chi_0(v)}{e^{-v^2/2}} e^{-\frac{(v+u)^2+v^2}{2}} \, dv \\ &= \frac{\sqrt{2}}{2} e^{-u^2/4} \int_0^{\infty} B_{j-1}(v) \frac{\chi_0(-v-u)}{e^{-(v+u)^2/2}} e^{-(v+u/2)^2} \, dv. \end{aligned} \tag{5.8}$$

The integrals (5.7) and (5.8) can be evaluated up to machine precision by Gaussian quadrature based on weight $e^{-(v-u/2)^2}$ and $e^{-(v+u/2)^2}$ respectively, as $B_{j-1}\chi_0 e^{v^2/2}$ is a polynomial with its degree up to $N+3$. The boundary condition (4.8) requires the numerical evaluation of the integral

$$\int_{v+u>0} (v+u) \phi \psi_{2j} \, dv.$$

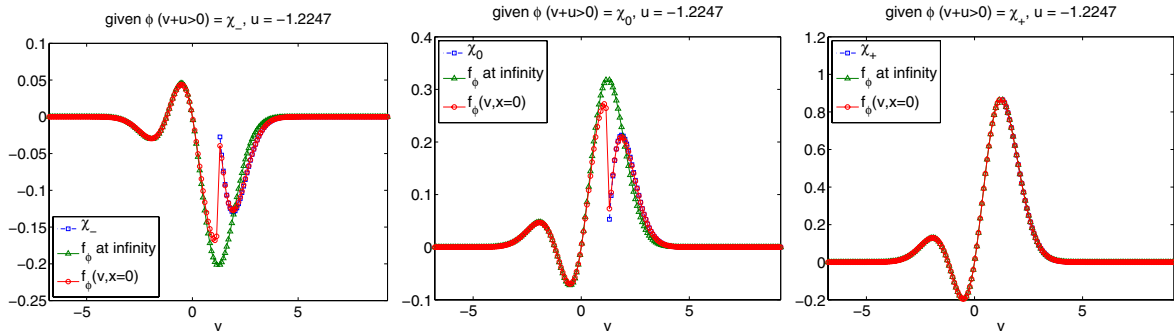


FIGURE 1. $u = -\sqrt{1.5} = -c$. In this case $\chi_+ \in H^0$, and χ_- and χ_0 are in H^- .

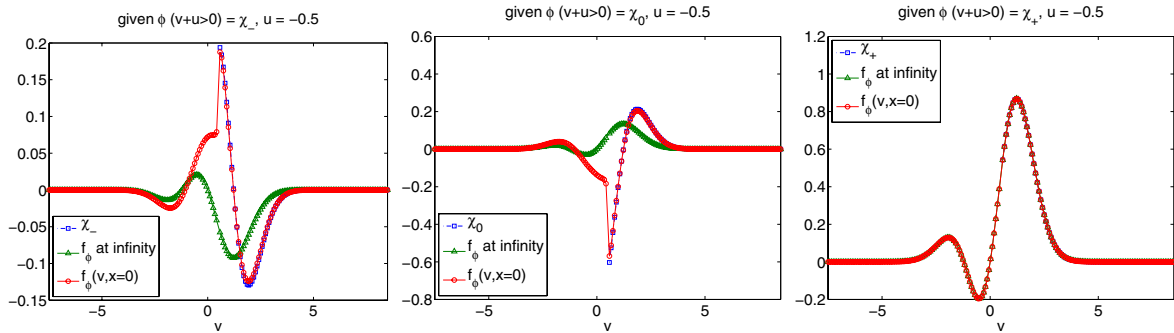


FIGURE 2. $-c < u = -0.5 < 0$. In this case $\chi_+ \in H^+$, and χ_- and χ_0 are in H^- .

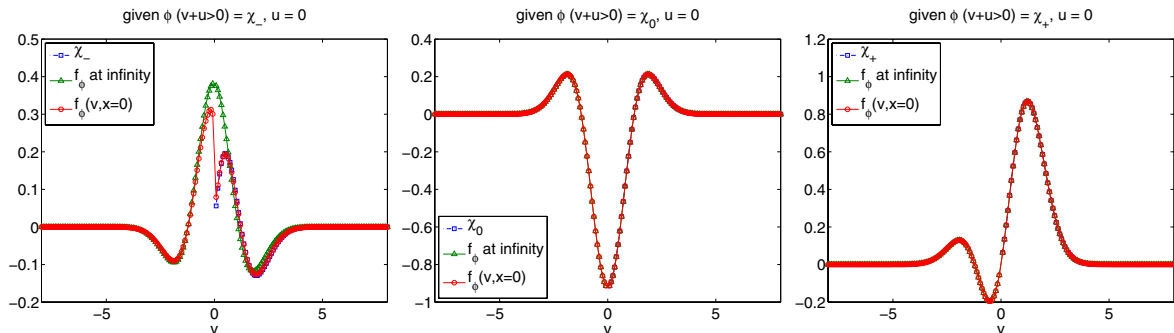


FIGURE 3. $u = 0$. In this case $\chi_+ \in H^+$, $\chi_0 \in H^0$ and $\chi_- \in H^-$.

We calculate this using Gaussian quadrature with the weight $e^{-(v+u)^2}$. The error of the quadrature depends on the number of quadrature points and the regularity of the incoming data ϕ .

We now present some numerical results for the linearized BGK equation. In the first set of examples, we compare our numerical results with analytical solutions, when the specified boundary data ϕ is given by the restriction of some $f \in H^0 \oplus H^+$ on $v > -u$. In this case, the solution to the undamped equation (1.1) is simply f on the whole velocity space. As discussed in (2.10), the dimension of the space $H^0 \oplus H^+$ depends on the bulk velocity u and the sound speed, which is $c = \sqrt{3/2}$ in our case as $T = 1/2$. We will choose $\chi_{+/-/0}$ defined in (2.8) as the incoming data. By the uniqueness of the half-space equation, the solution will simply be

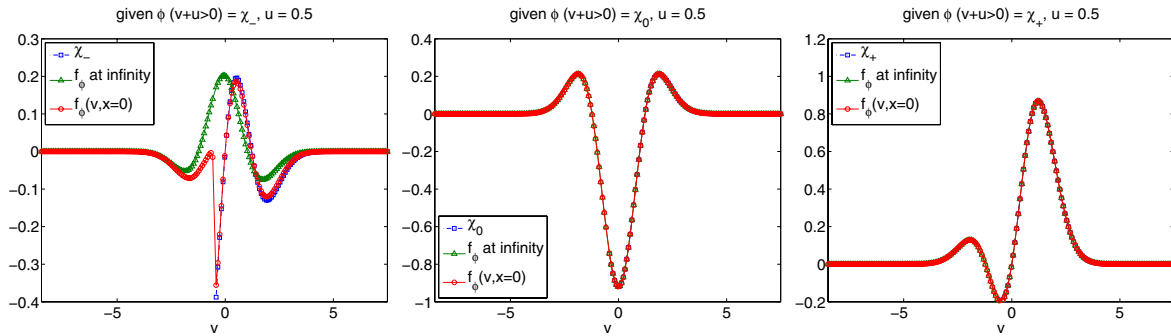


FIGURE 4. $0 < u = 0.5 < c$. In this case χ_+ and χ_0 are in H^+ , and $\chi_- \in H^-$.

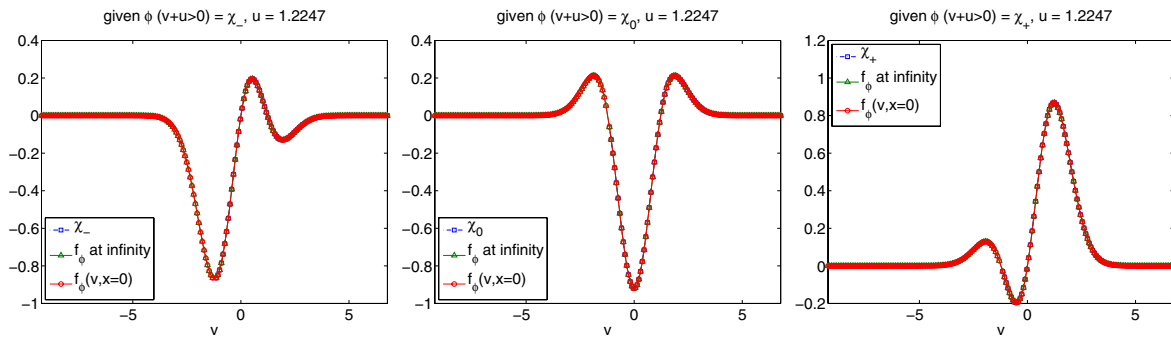


FIGURE 5. $u = \sqrt{1.5} = c$. In this case χ_+ and χ_0 are in H^+ , and $\chi_- \in H^0$.

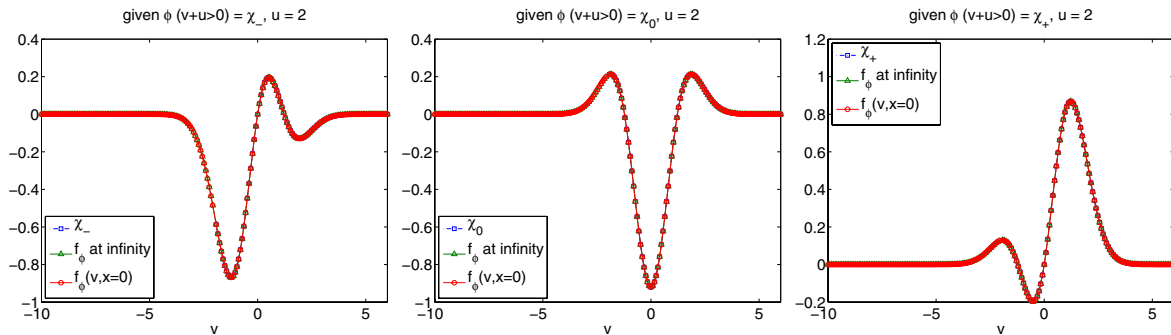


FIGURE 6. $u = 2 > c$. In this case all χ are in H^+ .

$\chi_{+ / 0}$ when the incoming data is chosen as $\chi_{+ / 0}$. We take six choices of u corresponding to the six cases listed in (2.10) (the case $u < -c$ gives an empty $H^0 \oplus H^+$ hence not included). The results are shown in Figures 1–6 below. In all these figures, the blue squared line is the incoming data, given by χ_- , χ_0 and χ_+ respectively. The green triangle line is the solution at $x = \infty$, and the red dotted line is the solution at $x = 0$.

Several remarks are in order: First, when the χ modes lie in $H^0 \oplus H^+$ for the given bulk background velocity u , we observe in Figure 1–6 that the solution at $x = 0$ gives a perfect match. We thus recover the exact solution from the numerical scheme. Second, we note that in general, the solution exhibits a jump at $v = -u$, as clearly seen for example in Figure 1 (left). This justifies our choice of the even-odd formulation and basis functions from

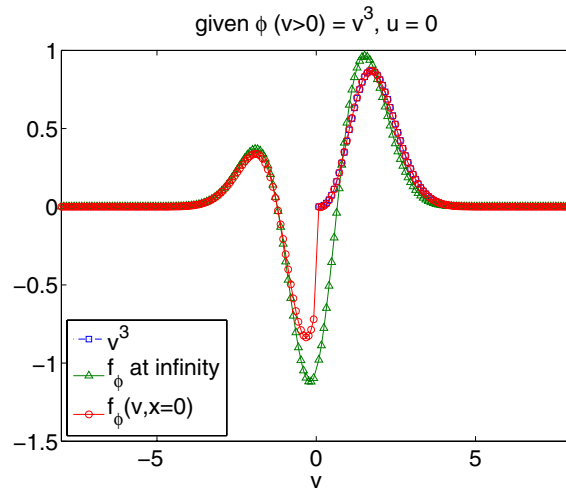


FIGURE 7. Blue boxed line is the input data $\phi = v^3(v > 0)$. Green triangle line is the solution at infinity and the red circled line is the solution at the boundary. $N = 36$ here. (Color online)

the half-space Hermite polynomials. Finally, we remark that we have used a filtering (with 2nd order cosine filter) to reduce the Gibbs oscillations caused by the large derivatives in some cases (for instance Fig. 2(left)).

Next, we consider an example where the exact solution is not known. We solve the equation (1.1) for $u = 0$ with boundary data $\phi = v^3, v > 0$. The numerical solution is shown in Figure 7.

5.2. Isotropic neutron transport equation

We further consider the isotropic neutron transport equation. The construction of the basis functions is similar to the linearized BGK case. However, instead of using half-space Hermite polynomials, we start with Legendre polynomials on the interval $[0, 1]$ and carry out the even-odd extensions. The Legendre polynomials, which are orthogonal polynomials for constant weight function, are used since the equilibrium states for the neutron transport equation are simply constants. We then apply Gauss–Legendre quadrature to assemble A and B for the generalized eigenvalue problem. The rest of the details are skipped here since the construction is relatively straightforward compared with the linearized BGK case.

To validate our methods in this case, we compare the numerical solution with the analytical solution with boundary data given by $\phi = v$ for $v \in [0, 1]$. The analytical solution is known as

$$f_\phi(-v) = \frac{1}{\sqrt{3}}H(v) - v, \quad v > 0, \quad (5.9)$$

where H is the Chandrasekhar H-function. In Figure 8 we plot both analytical and numerical solutions, where a second order cosine filter is used. The plot shows good agreement of the numerical solution with the exact one. Using the knowledge of the singularity of the solution at $v = 0$, more sophisticated techniques can be used to post-process the Galerkin solution. For example, Figure 9 shows the result of using Gegenbauer reprojection method (with end-point singularity) [10, 11, 17]. Excellent agreement with the exact solution is observed.

Furthermore, the limit at $x = \infty$ of the solution to the half-space isotropic NTE is a constant, whose amplitude agrees with the extrapolation length. In Table 1 we compare our numerical approximation of the extrapolation length with the exact result, which is again in good agreement. In comparison, we note that the approximate value for the extrapolation length obtained in [9] is 0.71040377 with 70 modes, while we achieve better results with piecewise polynomial of orders up to 12.

TABLE 1. Numerical approximations of the extrapolation length.

4	0.709324539775964	24	0.710445373807707	44	0.710446026371328	64	0.710446075479882
8	0.710386430787361	28	0.710445703544666	48	0.710446044962143	68	0.710446078520678
12	0.710434523809144	32	0.710445863417934	52	0.710446057194912	72	0.710446080785171
16	0.710442451548528	36	0.710445948444682	56	0.710446065509628	76	0.710446082499459
20	0.710444603305304	40	0.710445997010591	60	0.710446071320336	exact	0.710446089598763

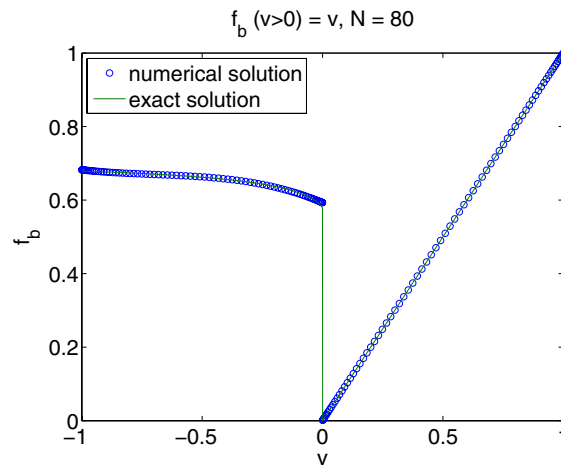


FIGURE 8. Analytical solution and numerical solution to the isotropic neutron transport equation at $x = 0$.

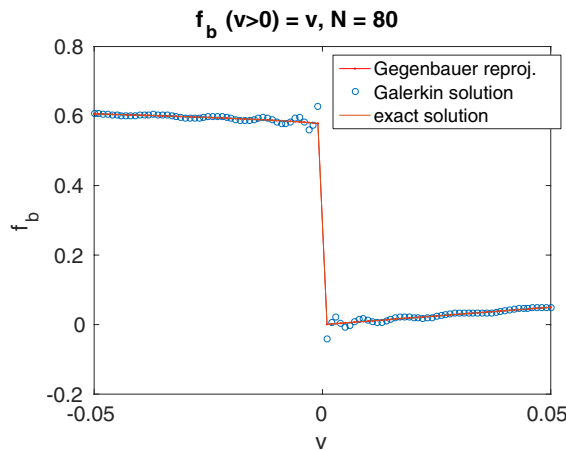


FIGURE 9. Analytical solution, numerical Galerkin solution, and the Gegenbauer reprojected solution to the isotropic neutron transport equation at $x = 0$ (zoomed in around $v = 0$).

APPENDIX A. HALF-HERMITE POLYNOMIAL

Here we derive the half-space orthogonal polynomial with weight $\exp(-(v-u)^2)$ with u a real number. The zeroth order half space Hermite polynomial is:

$$B_0 = \frac{1}{\sqrt{m_0}} \quad \text{with} \quad m_0 = \frac{\sqrt{\pi}}{2} (1 + \operatorname{erf}(u)). \quad (\text{A.1})$$

The higher order polynomials are defined through recurrence relation:

$$\sqrt{\beta_{n+1}} B_{n+1} = (v - \alpha_n) B_n - \sqrt{\beta_n} B_{n-1}, \quad (\text{A.2})$$

where α and β are defined by

$$\begin{cases} \beta_{n+1} = n + \frac{1}{2} + u\alpha_n - \alpha_n^2 - \beta_n; \\ \alpha_{n+1} = u - \alpha_n + \frac{1}{2\beta_{n+1}} \sum_{k=0}^n \alpha_k \end{cases} \quad (\text{A.3})$$

with $\alpha_0 = m_1/m_0$ and $\sqrt{\beta_1} = \sqrt{m_0 m_2 - m_1^2}/m_0$, where m_i , $i = 0, 1, 2$ are moments of the Gaussian:

$$m_i = \int_0^\infty v^i e^{-(v-u)^2} dv, \quad i = 0, 1, 2. \quad (\text{A.4})$$

The deduction formula are derived from the Christoffel–Darboux identity

$$\sum_{k=0}^n B_k^2 = \sqrt{\beta_{n+1}} (B'_{n+1} B_n - B_{n+1} B'_n) \quad (\text{A.5})$$

as follows. By orthogonality of $\{B_n\}$, we get

$$\alpha_n = \int_0^\infty v B_n^2 e^{-(v-u)^2} dv, \quad \text{and} \quad \sqrt{\beta_{n+1}} = \int_0^\infty v B_n B_{n+1} e^{-(v-u)^2} dv.$$

Integrate the identity (A.5) over v with the weight, we get

$$\begin{aligned} n+1 &= \sqrt{\beta_{n+1}} \int_0^\infty B'_{n+1} B_n e^{-(v-u)^2} dv = \int_0^\infty v B_{n+1} B'_{n+1} e^{-(v-u)^2} dv \\ &= -\frac{1}{2} + \int_0^\infty v^2 B_{n+1}^2 e^{-(v-u)^2} dv - u\alpha_n, \end{aligned}$$

where the second equality is obtained by taking the inner product with B'_{n+1} of recursion equation (A.2), and the third comes from integration by parts. From this we get the first deduction relation in (A.3). Next multiply (A.5) with v and then integrate, we obtain

$$\begin{aligned} \sum_{k=0}^n \alpha_k &= \sqrt{\beta_{n+1}} \int_0^\infty v B'_{n+1} B_n e^{-(v-u)^2} dv \\ &= \sqrt{\beta_{n+1}} \left(2 \int_0^\infty v^2 B_{n+1} B_n e^{-(v-u)^2} dv - 2u \int_0^\infty v B_{n+1} B_n e^{-(v-u)^2} dv \right) \\ &= 2\beta_{n+1} (\alpha_n + \alpha_{n+1} - u), \end{aligned}$$

where the first equality comes from the fact that $\int_0^\infty v B_{n+1} B'_n e^{-(v-u)^2} dv = 0$, the second is due to integration by parts, and the third comes from integrating the recursion equation (A.2) multiplied by $v B_{n+1}$. This gives the other deduction relation in (A.3).

REFERENCES

- [1] I. Babuška and A.K. Aziz, Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical foundation of the Finite Element method with Applications to Partial Differential Equations*, edited by A.K. Aziz. Academic Press, New York (1972) 1–359.
- [2] C. Besse, S. Borghol, T. Goudon, I. Lacroix-Violet and J.-P. Dudon, Hydrodynamic regimes, Knudsen layer, numerical schemes: definition of boundary fluxes. *Adv. Appl. Math. Mech.* **3** (2011) 519–561.
- [3] A. Bensoussan, J.L. Lions and G.C. Papanicolaou, Boundary-layers and homogenization of transport processes. *J. Publ. RIMS Kyoto Univ.* **15** (1979) 53–157.
- [4] C. Bardos, R. Santos and R. Sentis, Diffusion approximation and computation of the critical size. *Trans. Amer. Math. Soc.* **284** (1984) 617–649.
- [5] C. Bardos and X. Yang, The classification of well-posed kinetic boundary layer for hard sphere gas mixtures. *Commun. Partial Differ. Equ.* **37** (2012) 1286–1314.
- [6] Y. Cheng, I. Gamba and J. Proft, Positivity-preserving discontinuous Galerkin schemes for linear Vlasov-Boltzmann transport equations. *Math. Comput.* **81** (2012) 153–190.
- [7] F. Coron, F. Golse and C. Sulem, A classification of well-posed kinetic layer problems. *Commun. Pure Appl. Math.* **41** (1988) 409–435.
- [8] C.-C. Chen, T.-P. Liu and T. Yang, Existence of boundary layer solutions to the Boltzmann equation. *Anal. Appl.* **2** (2004) 337–363.
- [9] F. Coron, Computation of the asymptotic states for linear half space kinetic problems. *Transport Theory Statist. Phys.* **19** (1990) 89–114.
- [10] Z. Chen and C.-W. Shu, Recovering exponential accuracy from collocation point values of smooth functions with end-point singularities. *J. Comput. Appl. Math.* **265** (2014) 83–95.
- [11] Z. Chen and C.-W. Shu, Recovering exponential accuracy in Fourier spectral methods involving piecewise smooth functions with unbounded derivative singularities. *J. Sci. Comput.* (2015) 121.
- [12] S. Dellacherie, Coupling of the Wang Chang-Uhlenbeck equations with the multispecies Euler system. *J. Comput. Phys.* (2003) 189.
- [13] P. Degond and S. Mas-Gallic, Existence of solutions and diffusion approximation for a model Fokker-Planck equation. *Transport Theory Statist. Phys.* **16** (1987) 589–636.
- [14] H. Egger and M. Schlottbom, A mixed variational framework for the radiative transfer equation. *Math. Models Methods Appl. Sci.* **22** (2012) 1150014.
- [15] F. Golse and A. Klar, A numerical method for computing asymptotic states and outgoing distributions for kinetic linear half-space problems. *J. Stat. Phys.* **80** (1995) 1033–1061.
- [16] F. Golse, Analysis of the boundary layer equation in the kinetic theory of gases. *Bull. Inst. Math. Acad. Sin. (N.S.)* **3** (2008) 211–242.
- [17] D. Gottlieb and C.-W. Shu, On the Gibbs phenomenon and its resolution. *SIAM Rev.* **30** (1997) 644–668.
- [18] S. Jin, L. Pareschi and G. Toscani, Uniformly accurate diffusive relaxation schemes for multiscale transport equations. *SIAM J. Numer. Anal.* **38** (2001) 913–936.
- [19] Q. Li, J. Lu and W. Sun, Half-space kinetic equations with general boundary conditions. *Math. Comp.* **86** (2017) 1269–1301
- [20] R.E. Marshak, Note on the spherical harmonic method as applied to the Milne problem for a sphere. *Phys. Rev.* **71** (1947) 443–446.
- [21] B. Shizgal, A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems. *J. Comput. Phys.* **41** (1981) 309–328.
- [22] S. Ukai, T. Yang and S.-H. Yu, Nonlinear boundary layers of the Boltzmann equation. I. Existence. *Commun. Math. Phys.* **236** (2003) 373–393.
- [23] W. Wang, T. Yang and X. Yang, Nonlinear stability of boundary layers of the Boltzmann equation for cutoff hard potentials. *J. Math. Phys.* **47** (2006) 083301.
- [24] W. Wang, T. Yang and X. Yang, Existence of boundary layers to the Boltzmann equation with cutoff soft potentials. *J. Math. Phys.* **48** (2007) 073304.