Study of Discrete Scattering Operators for Some Linear Kinetic Models

Yanping Chen, Zheng Chen, Yingda Cheng, Adrianna Gillman, Fengyan Li

Abstract In this paper, we consider spatially homogeneous linear kinetic models arising from semiconductor device simulations and investigate how various deterministic numerical methods approximate their scattering operators. In particular, methods including first and second order discontinuous Galerkin methods, a first order collocation method, a Fourier-collocation spectral method, and a Nyström method are examined when they are applied to one-dimensional models with singular or continuous scattering kernels. Mathematical properties are discussed for the corresponding discrete scattering operators. We also present numerical experiments to demonstrate the performance of these methods. Understanding how the scattering operators are approximated can provide insights into designing efficient algorithms for simulating kinetic models and for the implicit discretizations of the problems in the presence of multiple scales.

Yanping Chen
Department of Mathematical Sciences, The University of Texas at Dallas, Dallas, TX 75252. e-mail: yxc110030@utdallas.edu

Zheng Chen
Department of Mathematics, Iowa State University, Ames, IA 50011. e-mail: zchen@iastate.edu

Yingda Cheng
Department of Mathematics, Michigan State University, East Lansing, MI 48824. e-mail: ycheng@math.msu.edu

Adrianna Gillman
Computational and Applied Mathematics Department, Rice University, Houston, TX 77005. e-mail: adrianna.gillman@rice.edu

Fengyan Li
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180. e-mail: lif@rpi.edu

1 Introduction

Kinetic models arise in many applications such as rarefied gas dynamics, plasma physics, nuclear engineering, semiconductor device design, traffic networking, and swarming. Such models evolve the probability distribution function of one or multiple species of particles, with or without forces from external or self-consistent fields. They can describe mesoscopic phenomena lying in between the microscopic particle dynamics governed by fundamental laws such as the Newton's laws of motion, and macroscopic dynamics described by continuum models. The numerical challenges often come from high dimensionality, various collision (or scattering) operators which can be multi-fold integrals or singular, and multiple scales in time or phase space.

In this work, we consider some simple one-dimensional linear kinetic models with either singular or continuous scattering operators, and investigate mathematically and/or computationally the properties of several deterministic numerical discretizations. They include first and second order discontinuous Galerkin methods, a first order collocation method, a Fourier-collocation spectral method, and a Nyström method. Lots of efforts have been put in the literature for simulations of semiconductor Boltzmann equations from algorithm and application points of view, for example, computations by spectral methods [6], finite difference methods [4, 2], and discontinuous Galerkin method [3]. In this paper, we are particularly concerned with characterizing and examining how various numerical methods capture the equilibriums. Since only spatially homogeneous models are considered, what we examine here is essentially on how the scattering operators are approximated numerically. Such study is important for understanding numerical approximations for scattering operators which are a key part in any collisional kinetic model, and can provide insights into designing efficient algorithms for numerical simulations and also for implicit discretizations of the problems in the presence of multiple scales.

Let's start with the models. Consider a one-dimensional electron-phonon scattering model [8, 10, 5]

$$\frac{\partial f(k,t)}{\partial t} = \hat{S}[f](k,t) = \int_{-\infty}^{\infty} \Big(S(k',k) f(k',t) - S(k,k') f(k,t) \Big) dk', \tag{1}$$

which arises from semiconductor device design. Here $f(k,t)$ is the probability distribution function of electrons with wavenumber k at time t, $\hat{S}$ is the scattering operator, and $S(k,k')$ is the scattering kernel which gives the transfer rate of electrons scattering from state k to k'. Note that the space variable x is omitted and the equation (1) is space homogeneous.

The first problem we will focus on is the governing equation (1) that models both the inelastic and elastic scattering, and the scattering kernel is defined as

$$S(k,k') = \sum_{\nu \in \{-1,0,1\}} s_\nu(E(k),E(k')) \delta\Big(E(k) - E(k') + \nu \varepsilon_p \Big). \tag{2}$$

Here $E(k)$ is the energy of the electron with wavenumber k, $s_\nu(E(k),E(k'))$ is the transfer rate from k to k' by absorbing ($\nu=1$) or emitting ($\nu=-1$) a phonon with an energy $\varepsilon_p>0$, or by keeping the energy unchanged ($\nu=0$). And $\delta(\cdot)$ is the Dirac-δ function. It is assumed $s_\nu(\cdot,\cdot)>0$ with $\nu=\pm1$, and $s_0(\cdot,\cdot)\geq 0$. We consider the Kane energy band, with the energy function $E(k)$ satisfying

$$E(k)(1+\alpha E(k))=k^2/2 \tag{3}$$

and the non-parabolicity factor $\alpha\geq 0$ is some constant parameter. We also define

$$\mathscr{K}_\alpha(E)=\sqrt{2E(1+\alpha E)}.$$

The energy function $E(k)$ is non-negative and it is an even function of k. When $\alpha=0$, it corresponds to the quadratic energy band.

With $T>0$ being any given lattice temperature, the following distribution function

$$f^G(k)=\exp\left(-\frac{E(k)}{T}\right) \tag{4}$$

defines an equilibrium of our model, under the assumption

$$s_1(E(k)-\varepsilon_p,E(k))=s_{-1}(E(k),E(k)-\varepsilon_p)\exp\left(-\frac{\varepsilon_p}{T}\right). \tag{5}$$

This assumption is made throughout this paper, and it ensures the detailed balance principle $S(k',k)f^G(k')=S(k,k')f^G(k)$. For the quadratic energy (3) with $\alpha=0$, the equilibrium (4) after normalization is a Gaussian distribution. Following a similar analysis as in [9], any equilibrium of our model is given by

$$f^e(k)=f^G(k)h(E(k)), \tag{6}$$

where $h(E)$ is some periodic function of period ε_p. The inclusion of an ε_p-periodic function factor $h(E)$ in an equilibrium is due to the δ-type scattering rule in (2).

The model we have described so far, defined in (1), (2), (3) with the assumption (5), involves a scattering kernel with δ-type singularity. In this work, we will also examine a model which is defined by (1) with a continuous scattering kernel

$$S(k,k')=\sigma(k,k')M(k') \tag{7}$$

where $\sigma(k,k')=\sigma(k',k)\geq 0$ and $M(k)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{k^2}{2T})$. For any given temperature $T>0$, this model has a unique Gaussian-type equilibrium $M(k)$ (up to a constant factor).

In [7], a first order finite volume method was introduced for the linear kinetic model (1) with (2), (3) and the assumption (5), when the energy band is quadratic ($\alpha=0$) without the elastic collision ($s_0=0$). A detailed study of the scattering matrix which approximates the scattering operator was performed. In particular, the eigenvalues of the scattering matrix were proven to be non-positive, showing the stability of the numerical scheme. The dependence of the geometric multiplicity of the

zero eigenvalue on the choice of the mesh grids was established based on linear algebra tools. Such theory was extended in [11] to more general models, such as those with general energy band (including Kane energy), with anisotropic scattering, and in higher dimensions. Our aim in this paper is to perform a thorough numerical study of the model with either a singular or continuous scattering kernel by considering more general methods, including higher order Galerkin-type method, collocation methods of low or high order accuracy. We are particularly concerned with the scattering matrix resulted from different types of discretization, and the interpretation of the numerical results when compared with their continuous counterparts in the models.

The rest of the paper is organized as follows. In Section 2, the kinetic model (1) with a singular scattering kernel (2)-(3) is considered. More specifically, a first order finite volume method as in [7, 11], which is also a first order discontinuous Galerkin (DG) method, is formulated in Section 2.2.1. Mathematical properties of the numerical scheme as well as the scattering matrix are reviewed, followed with some discussions. More general numerical methods, including a second order DG method, a first order collocation method, and a Fourier-collocation spectral method are formulated in Sections 2.2.2, 2.2.3 and 2.2.4, respectively. It turns out it is nontrivial to extend the algebraic analysis in [7] to more general numerical discretizations. Instead, we rely on extensive numerical experiments to understand these methods, see Section 2.3. In Section 3, the kinetic model (1) with a continuous scattering kernel (7) is considered, for which a Nyström discretization is introduced and tested numerically. A detailed summary and concluding remarks are made in Section 4.

2 Numerical methods for singular scattering kernels

In this section, the kinetic model (1) will be considered with a singular scattering kernel (2) and (3). We will start with rewriting the equation. We then formulate a first order discontinuous Galerkin (DG) method, which is also a first order finite volume method, a second order DG method, a first order collocation method, and a Fourier-collocation spectral method. For the first order DG method, we will also discuss the mathematical properties of the discrete scattering operator.

2.1 Reformulation of the model

Before introducing numerical methods, we first reformulate the scattering terms in our model to formally remove the δ-type singularity. Details will be given only for one term associated with the inelastic scattering, and the remaining terms can be treated similarly. Recall the definition of the composition of the δ-function with a differentiable function $z(\cdot)$,

$$\int_{-\infty}^{\infty}\delta(z(x))v(x)dx=\sum_{x_\star\in\{y:z(y)=0\}}\frac{v(x_\star)}{|z'(x_\star)|},$$

and using this, one gets

$$\begin{aligned}\mathscr{R}_1[f](k,t)&=\int_{-\infty}^{\infty}s_1(E(k'),E(k))\delta(E(k')-E(k)+\varepsilon_p)f(k',t)dk'\\&=\sum_{k_\star\in\{k_\star:E(k_\star)=E(k)-\varepsilon_p\}}\frac{s_1(E(k_\star),E(k))f(k_\star,t)}{|E'(k_\star)|}.\end{aligned}\tag{8}$$

Notation wise, one should understand that for any k with $E(k)-\varepsilon_p<0$, the corresponding term in (8) is excluded.

With $E(k_\star)=E(k)-\varepsilon_p$, one can easily verify

$$k_\star=\pm\sqrt{2E(k_\star)(1+\alpha E(k_\star))}=\pm\mathscr{K}_\alpha(E(k)-\varepsilon_p),\tag{9}$$

$$E'(k_\star)=\frac{k_\star}{1+2\alpha E(k_\star)}=\frac{\pm\mathscr{K}_\alpha(E(k)-\varepsilon_p)}{1+2\alpha(E(k)-\varepsilon_p)}.\tag{10}$$

Combining (8)-(10), we have

$$\mathscr{R}_1[f](k,t)=\frac{s_1(E,E+\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(f(\mathscr{K}_\alpha(E),t)+f(-\mathscr{K}_\alpha(E),t)\Big)|_{E=E(k)-\varepsilon_p}.\tag{11}$$

Following the similar derivation for other terms, our model (1)-(3) with the singular scattering kernel is reformulated as below,

$$\frac{\partial f(k,t)}{\partial t}=\hat{S}[f](k,t)=\sum_{m=1}^{4}\mathscr{R}_m[f](k,t),\tag{12}$$

where

$$\begin{aligned}\mathscr{R}_2[f](k,t)=&\frac{s_{-1}(E,E-\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(f(\mathscr{K}_\alpha(E),t)+f(-\mathscr{K}_\alpha(E),t)\Big)|_{E=E(k)+\varepsilon_p}\\\mathscr{R}_3[f](k,t)=&\frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(f(\mathscr{K}_\alpha(E),t)+f(-\mathscr{K}_\alpha(E),t)\Big)|_{E=E(k)}\\\mathscr{R}_4[f](k,t)=&-2f(k,t)\Bigg(\frac{s_1(E-\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k)+\varepsilon_p}\\&+\frac{s_{-1}(E+\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k)-\varepsilon_p}\\&+\frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k)}\Bigg).\end{aligned}$$

2.2 Numerical methods

2.2.1 First order discontinuous Galerkin method

In this subsection, we will describe a discontinuous Galerkin (DG) method using piecewise constant discrete space to numerically approximate the reformulated model (12). The method is also the first order finite volume method studied in [7, 11] in the absence of the elastic scattering term, namely when $s_0 = 0$.

We start with introducing some notation. Let $[-K_{\max}, K_{\max}]$ be the computational domain, with the assumption that the exact solution is zero in the machine accuracy level outside this domain. Let $0 = k_{1/2} < k_{3/2} < ... < k_{N+1/2} = K_{\max}$ be a partition of $[0, K_{\max}]$, and define $I_i = [k_{i-1/2}, k_{i+1/2}]$, $\Delta k_i = k_{i+1/2} - k_{i-1/2}$, $\forall i \in \mathscr{N}^+ = \{1, 2, \cdots, N\}$, and $\Delta k = \max_{1\le i\le N} \Delta k_i$. For the left half domain $[-K_{\max}, 0]$, a "symmetric" mesh is introduced with $I_{-i} = [k_{-i-1/2}, k_{-i+1/2}]$, and $k_{-i-1/2} = -k_{i+1/2}$, $i \in \mathscr{N}^+$. In terms of the energy variable, we define $E_{\max} = E(K_{\max})$, $E_{i-1/2} = E(k_{i-1/2}), i = 1, \cdots, N+1$, $\Omega_i = [E_{i-1/2}, E_{i+1/2}]$, $\Delta E_i = E_{i+1/2} - E_{i-1/2}, i \in \mathscr{N}^+$, and $\Delta E = \max_{1\le i\le N} \Delta E_i$. We also use $\Omega_i \pm \varepsilon_p = \{E \pm \varepsilon_p : E \in \Omega_i\}$ and $\mathscr{N} = \{-N, \cdots, -2, -1, 1, 2, \cdots, N\}$.

To formulate the method, we approximate $f(k,t)$ by a piecewise constant function $f_h(k,t)$, namely, $f_h(\cdot,t) \in V_h = V_h^0 = \{g : g|_{I_i} \in P^0(I_i), \forall i \in \mathscr{N}\}$, satisfying

$$\int_{I_i} \frac{\partial f_h(k,t)}{\partial t} \phi(k) dk = \int_{I_i} \hat{S}[f_h](k,t)\phi(k)dk = \sum_{m=1}^{4} \int_{I_i} \mathscr{R}_m[f_h](k,t)\phi(k)dk \qquad (13)$$

for any $\phi \in V_h$ and any $i \in \mathscr{N}$. Here and below $P^r(I_i)$ is the set of polynomials on I_i of degree r. This scheme, in its finite volume form, is also given as (14) with $\phi = 1$,

$$\int_{I_i} \frac{\partial f_h(k,t)}{\partial t} dk = \int_{I_i} \hat{S}[f_h](k,t)dk = \sum_{m=1}^{4} \int_{I_i} \mathscr{R}_m[f_h](k,t)dk. \qquad (14)$$

1.) The scheme in its algebraic form

Next, we will convert the scheme (14) into its algebraic form. To do so, we represent the numerical solution as $f_h(k,t)|_{I_i} = f_i(t)$, with $f_i(t) = \frac{1}{\Delta k_i} \int_{I_i} f_h(k,t)dk$ which approximates the cell average of the exact solution $f(k,t)$ over I_i, $\forall i \in \mathscr{N}$. It's straightforward to get,

$$\int_{I_i} \frac{\partial f_h(k,t)}{\partial t} dk = \frac{d}{dt}(\Delta k_i f_i(t)). \qquad (15)$$

To proceed with the remaining terms related to the scattering operator, we will take a change of variable from k to E. With the relation between the velocity k and the energy E in (3), we have $dk = (1 + 2\alpha E)/\mathscr{K}_\alpha(E)\, dE$ and

$$\int_{I_i} z(E(k))dk = \int_{\Omega_{|i|}} z(E)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}dE$$

for any given function $z(\cdot)$.

For the first term on the right-hand side of (14), we have

$$\begin{aligned}
&\int_{I_i} \mathscr{R}_1[f_h](k,t)dk = \int_{\Omega_{|i|}} \mathscr{R}_1[f_h](\mathscr{K}_\alpha(E),t)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}dE \\
&= \int_{\Omega_{|i|}-\varepsilon_p} s_1(E,E+\varepsilon_p)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\cdot\frac{1+2\alpha(E+\varepsilon_p)}{\mathscr{K}_\alpha(E+\varepsilon_p)}\Big(f_h(\mathscr{K}_\alpha(E),t)+f_h(-\mathscr{K}_\alpha(E),t)\Big)dE \\
&= \int_{\Omega_{|i|}-\varepsilon_p} s_1(E,E+\varepsilon_p)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\cdot\frac{1+2\alpha(E+\varepsilon_p)}{\mathscr{K}_\alpha(E+\varepsilon_p)}\sum_{j\in\mathscr{N}}\chi_{\Omega_{|j|}}(E)f_j(t)dE \\
&= \sum_{j\in\mathscr{N}} f_j(t)r_{i,j}^{(1)}, \qquad (16)
\end{aligned}$$

with

$$r_{i,j}^{(1)} = \int_{(\Omega_{|i|}-\varepsilon_p)\cap\Omega_{|j|}} s_1(E,E+\varepsilon_p)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\cdot\frac{1+2\alpha(E+\varepsilon_p)}{\mathscr{K}_\alpha(E+\varepsilon_p)}dE. \qquad (17)$$

Following similar derivation, we can further get $\int_{I_i}\mathscr{R}_m[f_h](k,t)dk = \sum_{j\in\mathscr{N}} f_j(t)r_{i,j}^{(m)}$, $m=2,3$ with

$$\begin{aligned}
r_{i,j}^{(2)} &= \int_{(\Omega_{|i|}+\varepsilon_p)\cap\Omega_{|j|}} s_{-1}(E,E-\varepsilon_p)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\cdot\frac{1+2\alpha(E-\varepsilon_p)}{\mathscr{K}_\alpha(E-\varepsilon_p)}dE \\
r_{i,j}^{(3)} &= \int_{\Omega_{|i|}\cap\Omega_{|j|}} s_0(E,E)\left(\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\right)^2 dE = \delta_{|i|,|j|}\int_{\Omega_{|i|}} s_0(E,E)\left(\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\right)^2 dE,
\end{aligned}$$

and $\int_{I_i}\mathscr{R}_4[f_h](k,t)dk = -2f_i(t)\hat{\lambda}_i$, with

$$\begin{aligned}
\hat{\lambda}_i = \int_{\Omega_{|i|}} &\left(\frac{s_1(E,E+\varepsilon_p)(1+2\alpha(E+\varepsilon_p))}{\mathscr{K}_\alpha(E+\varepsilon_p)} + \frac{s_{-1}(E,E-\varepsilon_p)(1+2\alpha(E-\varepsilon_p))}{\mathscr{K}_\alpha(E-\varepsilon_p)}\right) \\
&\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}dE + \int_{\Omega_{|i|}} s_0(E,E)\left(\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\right)^2 dE. \qquad (18)
\end{aligned}$$

Here δ_{ij} is the Kronecker-δ function.

By combining what we have so far in (14)-(18), the proposed first order DG scheme for the model (12) with the singular scattering kernel is converted to its algebraic form,

$$\frac{d}{dt}(\Delta k_i f_i) = -2\lambda_i(\Delta k_i f_i) + \sum_{j\in\mathscr{N}} s_{i,j}(\Delta k_j f_j), \quad \forall i\in\mathscr{N}, \qquad (19)$$

where

$$s_{i,j} = \frac{1}{\Delta k_j}(r_{i,j}^{(1)} + r_{i,j}^{(2)} + r_{i,j}^{(3)}) \qquad \text{and} \qquad \lambda_i = \frac{1}{\Delta k_i}\hat{\lambda}_i. \tag{20}$$

2.) Properties of the scheme

With $s_v(\cdot,\cdot) > 0, v = \pm 1$ and $s_0(\cdot,\cdot) \geq 0$, one can easily see that all the coefficients in the linear algebraic system (19) are non-negative, more specifically,

$$s_{i,j} \geq 0, \quad \lambda_i > 0, \qquad \forall i, j \in \mathscr{N}. \tag{21}$$

They also have some symmetry property, namely,

$$\lambda_i = \lambda_{-i}, \quad s_{i,j} = s_{-i,j} = s_{i,-j} = s_{-i,-j}, \qquad \forall i, j \in \mathscr{N} \tag{22}$$

due to that the energy $E(k)$ is an even function in k and the mesh is "symmetrically" defined.

Now we introduce

$$\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_N\}, \quad S = (s_{i,j})_{i,j\in\mathscr{N}^+}, \tag{23}$$

and $\mathbf{f}_- = [\Delta k_1 f_{-1}, \cdots, \Delta k_N f_{-N}]^T$, $\mathbf{f}_+ = [\Delta k_1 f_1, \cdots, \Delta k_N f_N]^T$, then the proposed scheme in (19) can be written as

$$\frac{d}{dt}\begin{bmatrix}\mathbf{f}_- \\ \mathbf{f}_+\end{bmatrix} = \mathbb{S}\begin{bmatrix}\mathbf{f}_- \\ \mathbf{f}_+\end{bmatrix} = \left(-2\begin{bmatrix}\Lambda & 0 \\ 0 & \Lambda\end{bmatrix} + \begin{bmatrix}S & S \\ S & S\end{bmatrix}\right)\begin{bmatrix}\mathbf{f}_- \\ \mathbf{f}_+\end{bmatrix}. \tag{24}$$

The matrix $\mathbb{S}$ is the discrete matrix corresponding to the discrete scattering operator. If we further define $\mathbf{g} = (\mathbf{f}_- + \mathbf{f}_+)/2 \in \mathbb{R}^N$, $\mathbf{h} = (\mathbf{f}_+ - \mathbf{f}_-)/2 \in \mathbb{R}^N$, and $M = 2(S - \Lambda)$, the linear system (24) can be decoupled into two systems of halved size,

$$\frac{d}{dt}\mathbf{g} = M\mathbf{g}, \qquad \frac{d}{dt}\mathbf{h} = -2\Lambda\mathbf{h}. \tag{25}$$

It is easy to see solving the proposed scheme (19) (or (24)) is equivalent to solving (25). In next lemma, we will summarize more properties of $\mathbb{S}$, M and Λ.

Lemma 2.1 *1.) Λ is non-singular.*
2.) The eigenvalues of $\mathbb{S}$ consist of all eigenvalues of M and of -2Λ. Hence the dimensions of $ker(\mathbb{S})$ and $ker(M)$ are the same.
3.) $\mathbf{g} \in ker(M) \Leftrightarrow [\mathbf{g}^\top, \mathbf{g}^\top]^\top \in ker(\mathbb{S})$.

The proof is straightforward, and it is omitted here. Based on the properties in this lemma, we can see that to address the types of questions as in [7] for the scattering matrix $\mathbb{S}$, such as the dimension of the null space of $\mathbb{S}$, the sign of the real part of the eigenvalues of $\mathbb{S}$, it is equivalent to ask similar questions to the reduced scattering matrix M. On the other hand, to get numerical solution $f_h(k,t)$ at any time t, one would have to work with both equations in (25) or with equation (24).

Next we will verify directly that the scheme given above has mass conservation property. An important consequence is that the column sum of M is zero. This property ensures zero is an eigenvalue of M, and it was also extensively used in analyzing M in [7]. Such property is usually not possessed by collocation-type methods. Instead with collocation methods, zero eigenvalue of the scattering operator can be approximated by nonzero numerical eigenvalues (see numerical results in Section 2.3).

Lemma 2.2 *Suppose the numerical solution $f_h(k,t)$ has compact support in $[-K_{max}, K_{max}]$, then the proposed scheme* (14) *satisfies mass conservation, namely,*

$$\frac{d}{dt}\int_{-K_{max}}^{K_{max}} f_h(k,t)dk = \frac{d}{dt}\sum_{i\in\mathscr{N}} \Delta k_i f_i(t) = 0. \tag{26}$$

Moreover $\sum_{i\in\mathscr{N}^+} M_{ij} = 0, \quad \forall j \in \mathscr{N}^+$.

Proof. Based on the formulas for $s_{i,j}$ and λ_j in (20) as well as the symmetry relation in (22), one can verify

$$\frac{1}{2}\sum_{i\in\mathscr{N}} s_{i,j} = \sum_{i\in\mathscr{N}^+} s_{i,j} = \lambda_j, \quad \forall j \in \mathscr{N}^+ \tag{27}$$

and hence $\sum_{i\in\mathscr{N}^+} M_{ij} = 2\,(\sum_{i\in\mathscr{N}^+} s_{i,j} - \lambda_j) = 0$ for all $\forall j \in \mathscr{N}^+$

With this and (19), we have

$$\begin{aligned}\frac{d}{dt}\int_{-K_{\max}}^{K_{\max}} f_h(k,t)dk &= \frac{d}{dt}\sum_{i\in\mathscr{N}} \Delta k_i f_i(t) = -2\sum_{i\in\mathscr{N}} \lambda_i(\Delta k_i f_i) + \sum_{i\in\mathscr{N}}\sum_{j\in\mathscr{N}} s_{i,j}(\Delta k_j f_j)\\ &= -2\sum_{i\in\mathscr{N}} \lambda_i(\Delta k_i f_i) + \sum_{j\in\mathscr{N}}\Big(\sum_{i\in\mathscr{N}} s_{i,j}\Big)(\Delta k_j f_j)\\ &= -2\sum_{i\in\mathscr{N}} \lambda_i(\Delta k_i f_i) + \sum_{j\in\mathscr{N}} 2\lambda_j(\Delta k_j f_j) = 0.\end{aligned} \tag{28}$$

□

In next theorem, we summarize the main results which were proved for the (reduced) scattering matrix M in [7, 11] when $s_0 = 0$. s_0 being nonzero does not pose new difficulty.

Theorem 2.3 *1.) $M_{ij} \geq 0$ for $i \neq j$, $M_{ii} < 0$, and $M^\top$ is weakly diagonally dominant. Each nonzero eigenvalue of M has a negative real part.*

2.) $M_{ij} > 0 \Leftrightarrow M_{ji} > 0$. In addition, there exists a unique positive integer s and a permutation matrix P such that

$$M = P^\top \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_s \end{bmatrix} P, \tag{29}$$

where each $M_i \in \mathbb{R}^{r_i \times r_i}, i = 1, \cdots, s$ is irreducible. Moreover $rank(M_i) = r_i - 1$, and this implies $rank(M) = N - s$ and $dim(ker(M)) = s$. Let $\mathbf{g}_i \in null(M_i)$ be nonzero for any i, all entries of $\mathbf{g}_i$ have the same sign.

3.) *The fact that $dim(ker(M)) = s$ can be equivalently characterized by the following property of the mesh: there exist $E_1^\star, \cdots E_s^\star \in [0, \varepsilon_p)$ with $E_1 < \cdots < E_s^\star$, such that*

$$\{E_i^\star + \tau\varepsilon_p : E_i^\star + \tau\varepsilon_p \le E_{max}, \tau \in \mathbb{N}\} \subseteq \{E_{j-1/2} : j = 1, 2, \cdots N+1\}, \quad i = 1, \cdots, s. \tag{30}$$

From the theorem, one can see that there is no nonzero eigenvalue of M with real positive part, and this implies the stability of the scheme and ensures the correct decay behavior of the numerical solution over long time period. One can always find a set of basis for the null space of M such that each basis vector in nonnegative. In addition, the geometric multiplicity of zero eigenvalue being s, hence $\dim(\ker(M)) = s$, can be fully characterized by the choice of the mesh grids. To further understand the mesh condition in (30), recall our model admits infinitely many equilibriums (6), and the presence of an ε_p-periodic function factor $h(E(k))$ is due to the δ-type scattering rule in the model. With this, the behaviors of an equilibrium $f^e(k)$ at k and k' are related only when $E(k) = E(k') + v\varepsilon_p$, with $v = -1, 0, 1$. The statement in 3.) implies that the dimension s of the null space of M is the same as the total number of decoupled subregions of the energy domain under the scattering rule on the *numerical* level. (This is best illustrated by Figure 1 in [7].) Such result is not hard to get intuitively, and it is mathematically justified by the Theorem above for the first order DG method. It turns out similar analysis is non-trivial to establish for other numerical discretizations considered in Section 2.2. Without any analysis available, in order to understand how the scattering rule determined by each numerical discretization of the model decouples the energy domain, to what extent the numerical discretization captures the equilibriums of the scattering operator, we will numerically examine the null space of M or the steady-state of the discretized system, see Section 2.3.

Remark 2.4 *In practice, uniform meshes in the energy variable E are often used with $\Delta E_i = \Delta E, \forall i \in \mathscr{N}^+$. In such situation, if $\varepsilon_p/\Delta E = n \in \mathbb{Z}^+$, we have $dim(ker(M)) = n$; if $\varepsilon_p/\Delta E$ is not an integer, then $dim(ker(M)) = 1$.*

Remark 2.5 *The mass conservation property is one of the keys for the results in the above theorem. It is ensured by the relation* (27)*. To implement the proposed scheme, if $s_{i,j}$ and $\lambda_i, \forall i, j \in \mathscr{N}^+$ are computed independently using numerical quadrature, this relation will hold only up to the accuracy of the quadrature formulas. In our actual implementation, $\{s_{i,j}\}_{i,j\in\mathscr{N}^+}$ are computed first, then λ_j is obtained based on* (27)*, hence the mass conservation is enforced.*

2.2.2 Second order discontinuous Galerkin method

Following the same notation for the computational domain and the mesh as in Section 2.2.1, we introduce the discrete space

$$V_h = V_h^1 = \{g : g|_{I_i} \in P^1(I_i), \forall i \in \mathscr{N}\} \tag{31}$$

which consists of piecewise linear polynomials with respect to the mesh. We then approximate the solution $f(k,t)$ by $f_h(\cdot,t) \in V_h$, satisfying

$$\int_{I_i} \frac{\partial f_h(k,t)}{\partial t}\phi(k)dk = \int_{I_i} \hat{S}[f_h](k,t)\phi(k)dk = \sum_{m=1}^{4}\int_{I_i} \mathscr{R}_m[f_h](k,t)\phi(k)dk \tag{32}$$

for any $\phi \in V_h$ and $i \in \mathscr{N}$. This results in a (formally) second order DG method.

To convert our scheme into its algebraic form, suppose $\phi_i^0(k)$ and $\phi_i^1(k)$ are the basis functions of $P^1(I_i)$, and the numerical solution is represented as $f_h(k,t)|_{I_i} = f_i^0(t)\phi_i^0(k) + f_i^1(t)\phi_i^1(k)$, with $f_i^0(t)$ and $f_i^1(t)$ to be determined by the scheme (32). With the test function $\phi \in V_h$ in (32) taken to be $\phi|_{I_i} = g_i^0\phi_i^0(k) + g_i^1\phi_i^1(k)$, the term on the left-hand side becomes

$$\int_{I_i} \frac{\partial f_h(k,t)}{\partial t}\phi(k)dk = [g_i^0, g_i^1]A_i\frac{d}{dt}\begin{bmatrix} f_i^0 \\ f_i^1 \end{bmatrix},$$

with

$$A_i = \int_{I_i} \begin{bmatrix} (\phi_i^0)^2 & \phi_i^1\phi_i^0 \\ \phi_i^0\phi_i^1 & (\phi_i^1)^2 \end{bmatrix} dk.$$

For the first term on the right-hand side of (32), we have

$$\begin{aligned}
&\int_{I_i} \mathscr{R}_1[f_h](k,t)\phi(k)dk = \int_{\Omega_{|i|}} \mathscr{R}_1[f_h](\mathscr{K}_\alpha(E),t)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\phi\Big(\text{sign}(i)\ \mathscr{K}_\alpha(E)\Big)dE \\
&= \int_{\Omega_{|i|}-\varepsilon_p} \mathscr{R}_1[f_h](\mathscr{K}_\alpha(E+\varepsilon_p),t)\frac{1+2\alpha(E+\varepsilon_p)}{\mathscr{K}_\alpha(E+\varepsilon_p)}\phi\Big(\text{sign}(i)\ \mathscr{K}_\alpha(E+\varepsilon_p)\Big)dE \\
&= \int_{\Omega_{|i|}-\varepsilon_p} s_1(E,E+\varepsilon_p)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\cdot\frac{1+2\alpha(E+\varepsilon_p)}{\mathscr{K}_\alpha(E+\varepsilon_p)}\Big(f_h(\mathscr{K}_\alpha(E),t) \\
&\qquad + f_h(-\mathscr{K}_\alpha(E),t)\Big)\phi\Big(\text{sign}(i)\ \mathscr{K}_\alpha(E+\varepsilon_p)\Big)dE.
\end{aligned} \tag{33}$$

Note that

$$\begin{aligned}
&f_h(\mathscr{K}_\alpha(E),t) + f_h(-\mathscr{K}_\alpha(E),t) \\
&= \sum_{j\in\mathscr{N}} \chi_{\Omega_{|j|}}(E)\left(f_j^0\phi_j^0(k) + f_j^1\phi_j^1(k)\right)|_{k=\text{sign}(j)\ \mathscr{K}_\alpha(E)},
\end{aligned} \tag{34}$$

then

$$\int_{I_i} \mathscr{R}_1[f_h](k,t)\phi(k)dk = \sum_{j\in\mathscr{N}} [g_i^0, g_i^1]S_{i,j}^1\begin{bmatrix} f_j^0 \\ f_j^1 \end{bmatrix}, \tag{35}$$

with

$$S_{i,j}^1 = \int_{(\Omega_{|i|}-\varepsilon_p)\cap\Omega_{|j|}} s_1(E,E+\varepsilon_p)\frac{(1+2\alpha E)}{\mathscr{K}_\alpha(E)}\cdot\frac{(1+2\alpha(E+\varepsilon_p))}{\mathscr{K}_\alpha(E+\varepsilon_p)}\begin{bmatrix} \phi_j^0(\Delta)\phi_i^0(\Delta_1) & \phi_j^1(\Delta)\phi_i^0(\Delta_1) \\ \phi_j^0(\Delta)\phi_i^1(\Delta_1) & \phi_j^1(\Delta)\phi_i^1(\Delta_1) \end{bmatrix} dE$$

and $\Delta = \text{sign}(j)\ \mathscr{K}_\alpha(E)$, $\Delta_1 = \text{sign}(i)\ \mathscr{K}_\alpha(E+\varepsilon_p)$. Similarly,

$$\int_{I_i} \mathscr{R}_m[f_h](k,t)\phi(k)dk = \sum_{j\in\mathscr{N}} [g_i^0, g_i^1] S_{i,j}^m \begin{bmatrix} f_j^0 \\ f_j^1 \end{bmatrix} \tag{36}$$

for $m = 2,3$, with

$$S_{i,j}^2 = \int_{(\Omega_{|i|}+\varepsilon_p)\cap\Omega_{|j|}} s_{-1}(E,E-\varepsilon_p)\frac{(1+2\alpha E)}{\mathscr{K}_\alpha(E)}\cdot\frac{(1+2\alpha(E-\varepsilon_p))}{\mathscr{K}_\alpha(E-\varepsilon_p)} \begin{bmatrix} \phi_j^0(\Delta)\phi_i^0(\Delta_2) & \phi_j^1(\Delta)\phi_i^0(\Delta_2) \\ \phi_j^0(\Delta)\phi_i^1(\Delta_2) & \phi_j^1(\Delta)\phi_i^1(\Delta_2) \end{bmatrix} dE,$$

$$S_{i,j}^3 = \delta_{|i|,|j|}\int_{\Omega_{|i|}} s_0(E,E)\Big(\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\Big)^2 \begin{bmatrix} \phi_j^0(\Delta)\phi_i^0(\Delta_3) & \phi_j^1(\Delta)\phi_i^0(\Delta_3) \\ \phi_j^0(\Delta)\phi_i^1(\Delta_3) & \phi_j^1(\Delta)\phi_i^1(\Delta_3) \end{bmatrix} dE,$$

and $\Delta_2 = \text{sign}(i)\ \mathscr{K}_\alpha(E-\varepsilon_p)$, $\Delta_3 = \text{sign}(i)\ \mathscr{K}_\alpha(E)$. Moreover

$$\int_{I_i} \mathscr{R}_4[f_h](k,t)\phi(k)dk = -2[g_i^0, g_i^1]\Lambda_i \begin{bmatrix} f_i^0 \\ f_i^1 \end{bmatrix}, \tag{37}$$

with

$$\Lambda_i = \int_{\Omega_{|i|}} \Theta(E) \begin{bmatrix} (\phi_i^0(\Delta_3))^2 & \phi_i^1(\Delta_3)\phi_i^0(\Delta_3) \\ \phi_i^0(\Delta_3)\phi_i^1(\Delta_3) & (\phi_i^1(\Delta_3))^2 \end{bmatrix} dE.$$

Here $\Theta(E) = \Big(\frac{s_1(E,E+\varepsilon_p)(1+2\alpha(E+\varepsilon_p))}{\mathscr{K}_\alpha(E+\varepsilon_p)} + \frac{s_{-1}(E,E-\varepsilon_p)(1+2\alpha(E-\varepsilon_p))}{\mathscr{K}_\alpha(E-\varepsilon_p)}\Big)\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)} + s_0(E,E)\Big(\frac{1+2\alpha E}{\mathscr{K}_\alpha(E)}\Big)^2$.

Now with $S_{i,j} = S_{i,j}^1 + S_{i,j}^2 + S_{i,j}^3$, and the test function ϕ being arbitrary, the scheme becomes

$$A_i\frac{d}{dt}\begin{bmatrix} f_i^0 \\ f_i^1 \end{bmatrix} = -2\Lambda_i \begin{bmatrix} f_i^0 \\ f_i^1 \end{bmatrix} + \sum_{j\in\mathscr{N}} S_{i,j}\begin{bmatrix} f_j^0 \\ f_j^1 \end{bmatrix}, \qquad i\in\mathscr{N}. \tag{38}$$

Next we specify the local basis functions $\{\phi_i^r\}_{i\in\mathscr{N},r=1,2}$ as Lagrangian basis, given as

$$\phi_i^0(k) = \frac{1}{\Delta k_i}(k_{i+\frac{1}{2}} - k), \quad \phi_i^1(k) = \frac{1}{\Delta k_i}(k - k_{i-\frac{1}{2}}), \quad \text{if } i > 0, \tag{39}$$

$$\phi_i^0(k) = \frac{1}{\Delta k_{|i|}}(k - k_{i-\frac{1}{2}}), \quad \phi_i^1(k) = \frac{1}{\Delta k_{|i|}}(k_{i+\frac{1}{2}} - k), \quad \text{if } i < 0. \tag{40}$$

With such choice, the local basis functions have certain symmetry,

$$\phi_i^r(k) = \phi_{-i}^r(-k), \qquad r = 0,1, \quad i\in\mathscr{N}, \tag{41}$$

and so are the element-wise matrices

$$S_{i,j} = S_{i,-j} = S_{-i,j} = S_{-i,-j}, \qquad \Lambda_i = \Lambda_{-i}, \qquad A_i = A_{-i}. \tag{42}$$

If we introduce $\mathbf{f}_- = [f_{-1}^0, f_{-1}^1, \cdots, f_{-N}^0, f_{-N}^1]^T$, $\mathbf{f}_+ = [f_1^0, f_1^1, \cdots, f_N^0, f_N^1]^T$, the scheme (38) can be written more compactly,

$$\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \frac{d}{dt} \begin{bmatrix} \mathbf{f}_- \\ \mathbf{f}_+ \end{bmatrix} = \mathbb{S} \begin{bmatrix} \mathbf{f}_- \\ \mathbf{f}_+ \end{bmatrix} = \left(-2 \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda \end{bmatrix} + \begin{bmatrix} S & S \\ S & S \end{bmatrix} \right) \begin{bmatrix} \mathbf{f}_- \\ \mathbf{f}_+ \end{bmatrix}. \tag{43}$$

The matrix $A \in \mathbb{R}^{2N\times 2N}$ (resp. $\Lambda \in \mathbb{R}^{2N\times 2N}$) is a $N \times N$ block-diagonal matrix, with its (i,i)-th block being A_i (resp. Λ_i). The matrix $S \in \mathbb{R}^{2N\times 2N}$ is a $N \times N$ block-structured matrix, with its (i,j)-th block being $S_{i,j}$. And the scheme (43) can be further decoupled into two systems of halved size,

$$A\frac{d}{dt}\mathbf{g} = M\mathbf{g}, \qquad A\frac{d}{dt}\mathbf{h} = -2\Lambda\mathbf{h}. \tag{44}$$

Here $\mathbf{g} = (\mathbf{f}_- + \mathbf{f}_+)/2$, $\mathbf{h} = (\mathbf{f}_+ - \mathbf{f}_-)/2$, and $M = 2(S - \Lambda)$. We can verify directly from the definition that both Λ and the mass matrix A are invertible.

Similar as for the first order DG method, if we are only concerned with the discrete equilibrium such as the dimension of the null space of the scattering matrix $\mathbb{S}$ in (43), it is sufficient to simply consider $A\frac{d}{dt}\mathbf{g} = M\mathbf{g}$ for the same question. For the time evolving numerical solution $f_h(k,t)$, one needs to work with (43), or equivalently the two equations in (44). On the other hand, it is non-trivial to extend most of the algebraic analysis in [7, 11] to this second order method, for which the involved matrices are of block-structure.

What we do know is that the column sum of M is zero, and this again is closely related to the mass conservation of the method, as stated in next lemma.

Lemma 2.6 *Suppose the numerical solution $f_h(k,t)$ has compact support in $[-K_{max}, K_{max}]$, then the proposed scheme* (32) *satisfies mass conservation, namely,*

$$\frac{d}{dt}\int_{-K_{max}}^{K_{max}} f_h(k,t)dk = 0. \tag{45}$$

In addition, the sum of each column of M is zero.

Proof. Based on the formulas for Λ_i and $S_{i,j}$, the symmetry in (42), as well as the equality $\phi_i^0 + \phi_i^1 = 1$ on I_i, one can verify

$$[1,1]\left(-2\Lambda_i + \sum_{j\in\mathcal{N}} S_{j,i} \right) = 0 \tag{46}$$

and the sum of M being 0.

Using (46) as well as (32) with $\phi(k) \equiv 1$, we have

$$\frac{d}{dt}\int_{-K_{\max}}^{K_{\max}} f_h(k,t)dk = \sum_{i\in\mathscr{N}}\int_{I_i}\frac{\partial f_h(k,t)}{\partial t}dk = \sum_{i\in\mathscr{N}}\int_{I_i}\hat{S}[f_h](k,t)dk$$
$$= \sum_{i\in\mathscr{N}}[1,1]\left(-2\Lambda_i\begin{bmatrix} f_i^0 \\ f_i^1\end{bmatrix} + \sum_{j\in\mathscr{N}} S_{j,i}\begin{bmatrix} f_j^0 \\ f_j^1\end{bmatrix}\right)$$
$$= \sum_{i\in\mathscr{N}}[1,1]\left(-2\Lambda_i + \sum_{j\in\mathscr{N}} S_{j,i}\right)\begin{bmatrix} f_i^0 \\ f_i^1\end{bmatrix} = 0.$$

□

2.2.3 First order collocation method

So far, Galerkin-type methods are considered. In next two sections, our attention will be turned to collocation methods. In this subsection, we will construct a first order collocation scheme for (12). We start with introducing one collocation point $\xi_i \in I_i$ from each cell, and the actual choices will be specified later. A collocation method of the first order is then defined by requiring the piecewise constant numerical solution $f_h(k,t) \in V_h = V_h^0$ satisfy

$$\frac{\partial f_h(\xi_i,t)}{\partial t} = \hat{S}[f_h](\xi_i,t), \qquad \forall i \in \mathscr{N}. \tag{47}$$

We define $f_h(\xi_i,t) = f_i(t)$. Recall from Section 2.2.1,

$$f_h(\mathscr{K}_\alpha(E),t) + f_h(-\mathscr{K}_\alpha(E),t) = \sum_{j\in\mathscr{N}} \chi_{\Omega_{|j|}}(E) f_j(t) \tag{48}$$

then the scheme becomes

$$\frac{d}{dt} f_i(t) = -2\lambda_i f_i(t) + \sum_{j\in\mathscr{N}} s_{i,j} f_j(t). \tag{49}$$

Here

$$s_{i,j} = \left(\frac{s_1(E,E+\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\chi_{\Omega_{|j|}}(E)\right)|_{E=E(\xi_i)-\varepsilon_p}$$
$$+ \left(\frac{s_{-1}(E,E-\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\chi_{\Omega_{|j|}}(E)\right)|_{E=E(\xi_i)+\varepsilon_p}$$
$$+ \left(\frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\chi_{\Omega_{|j|}}(E)\right)|_{E=E(\xi_i)}, \tag{50}$$

and

$$\lambda_i = \frac{s_1(E-\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(\xi_i)+\varepsilon_p} + \frac{s_{-1}(E+\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(\xi_i)-\varepsilon_p} + \frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(\xi_i)}. \tag{51}$$

Again, the terms involving $E = E(\xi_i) - \varepsilon_p < 0$ are excluded.

Note that all the coefficients in the linear algebraic equation (49) are non-negative, more specifically,

$$s_{i,j} \geq 0, \quad \lambda_i > 0, \qquad \forall i, j \in \mathscr{N}. \tag{52}$$

If we further require the collocation points are chosen to satisfy

$$\xi_i = -\xi_{-i}, \quad i \in \mathscr{N}^+, \tag{53}$$

then the energy function $E(k)$ being an even function implies $E(\xi_i) = E(\xi_{-i})$, and the following symmetries hold

$$\lambda_i = \lambda_{-i}, \quad s_{i,j} = s_{-i,j} = s_{i,-j} = s_{-i,-j}, \qquad \forall i, j \in \mathscr{N}.$$

Now we let

$$\Lambda = \mathrm{diag}\{\lambda_1, \cdots, \lambda_N\}, \quad S = (s_{i,j})_{i,j\in\mathscr{N}^+}, \tag{54}$$

and $\mathbf{f}_- = [f_{-1}, \cdots, f_{-N}]^T$, $\mathbf{f}_+ = [f_1, \cdots, f_N]^T$, then the proposed scheme in (49) can be written as

$$\frac{d}{dt}\begin{bmatrix}\mathbf{f}_-\\ \mathbf{f}_+\end{bmatrix} = \mathbb{S}\begin{bmatrix}\mathbf{f}_-\\ \mathbf{f}_+\end{bmatrix} = \left(-2\begin{bmatrix}\Lambda & 0\\ 0 & \Lambda\end{bmatrix} + \begin{bmatrix}S & S\\ S & S\end{bmatrix}\right)\begin{bmatrix}\mathbf{f}_-\\ \mathbf{f}_+\end{bmatrix}. \tag{55}$$

Note $\mathbf{f}_+$ and $\mathbf{f}_-$ are defined differently from those in section 2.2.1 and they do not contain the mesh parameter $\{\Delta k_i\}_i$.

If we further define $\mathbf{g} = (\mathbf{f}_- + \mathbf{f}_+)/2$, $\mathbf{h} = (\mathbf{f}_+ - \mathbf{f}_-)/2$, and $M = 2(S - \Lambda)$, then the proposed scheme (55) can be decoupled into two systems of halved size

$$\frac{d}{dt}\mathbf{g} = M\mathbf{g}, \qquad \frac{d}{dt}\mathbf{h} = -2\Lambda\mathbf{h}. \tag{56}$$

Just as for the DG methods in Sections 2.2.1 and 2.2.2, if we are only concerned with the properties of the scattering matrix $\mathbb{S}$ regarding the discrete equilibrium, it is sufficient to simply consider $\frac{d}{dt}\mathbf{g} = M\mathbf{g}$.

Remark 2.7 *Compared with Galerkin methods in Sections 2.2.1 and 2.2.2, collocation methods proposed here and in next subsection are much simpler to formulated and to implement. On the other hand, collocation methods in general do not preserve mass conservation property.*

2.2.4 Fourier-collocation spectral method

In this subsection, we will formulate a Fourier-collocation spectral method for the linear kinetic model with a singular scattering kernel, which is now reformulated into (12). It is assumed that the solution $f(k,t)$ is zero outside the interval $[-K_{\max}, K_{\max}]$, thus can be extended periodically. For simplicity, we use $K = K_{\max}$ throughout this subsection.

We seek an approximating solution $f_N(k,t)$ in the space $\hat{B}_N[-K,K] = \text{span}\{e^{i\frac{\pi}{K}nk}\}_{|n|\le N}$, i.e.

$$f_N(k,t) = \sum_{n=-N}^{N} \hat{f}_n(t) e^{i\frac{\pi}{K}nk}, \tag{57}$$

with the unknown coefficients $\hat{f}_n(t), n = -N, \cdots, N$, to be determined. For any function $g \in \hat{B}_N[-K,K]$, one can define its residual associated with the equation (1)

$$R_N(k,t;g) = \frac{\partial g(k,t)}{\partial t} - \hat{S}[g](k,t) = \frac{\partial g(k,t)}{\partial t} - \sum_{m=1}^{4} \mathscr{R}_m[g](k,t).$$

In the Fourier-collocation method, we require that the residual of the numerical solution $f_N(k,t)$ vanishes at a set of collocation grid points $\{k_j\}_{-N\le j\le N}$, defined as

$$k_j = K\frac{2j}{2N+1}, \qquad -N \le j \le N.$$

Having this choice of the collocation points, the Fourier coefficients $\hat{f}_n(t)$ of the numerical solution $f_N(k,t)$ can be approximated by the discrete Fourier coefficients $\tilde{f}_n(t)$ based on the trapezoidal rule,

$$\tilde{f}_n(t) = \frac{1}{2N+1} \sum_{j=-N}^{N} f_N(k_j,t) e^{-i\frac{\pi}{K}nk_j}. \tag{58}$$

Thus the numerical solution $f_N(k,t)$, as a trigonometric polynomial, can also be expressed as

$$f_N(k,t) = \sum_{j=-N}^{N} f_N(k_j,t) g_j(k), \tag{59}$$

where $g_j(k)$ $(-N \le j \le N)$ is the Lagrange interpolation polynomial, given as

$$g_j(k) = \frac{\sin\left(\frac{2N+1}{2}\frac{\pi}{K}(k-k_j)\right)}{(2N+1)\sin\left(\frac{\pi}{2K}(k-k_j)\right)} \tag{60}$$

and satisfying $g_j(k_n) = \delta_{jn}$. Now the Fourier-collocation method can be stated as follows. Look for $f_N(k,t)$ in the form of (59), such that

$$R_N(k_j,t;f_N) = \frac{\partial f_N(k_j,t)}{\partial t} - \hat{S}[f_N](k_j,t)$$
$$= \frac{\partial f_N(k_j,t)}{\partial t} - \sum_{m=1}^{4} \mathscr{R}_m[f_N](k_j,t) = 0, \qquad -N \le j \le N. \tag{61}$$

This yields $2N+1$ equations to determine the $2N+1$ point values $f_N(k_j,t), j = -N,\cdots,N$, of the numerical solution.

Next we will convert the scheme to its algebraic form. From (12),

$$\mathscr{R}_1[f_N](k_j,t) = \frac{s_1(E,E+\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(f(\mathscr{K}_\alpha(E),t) + f(-\mathscr{K}_\alpha(E),t)\Big)|_{E=E(k_j)-\varepsilon_p}$$
$$= \frac{s_1(E,E+\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)} \sum_{j=-N}^{N} f_N(k_j,t)\Big(g_j(\mathscr{K}_\alpha(E)) + g_j(-\mathscr{K}_\alpha(E))\Big)|_{E=E(k_j)-\varepsilon_p}. \tag{62}$$

The remaining terms in (61) can be treated similarly. We define the solution vector $\mathbf{f}$ by collecting all the unknown coefficients in (59),

$$\mathbf{f} = [f_N(k_{-N},t),\ldots,f_N(k_{-1},t),f_N(k_0,t),f_N(k_1,t),\ldots,f_N(k_N,t)]^\top \in \mathbb{R}^{2N+1},$$

then the proposed Fourier-collocation method becomes a linear system

$$\frac{d\mathbf{f}}{dt} = \mathbb{S}\mathbf{f}, \tag{63}$$

where $\mathbb{S} = -2\Lambda + S \in \mathbb{R}^{(2N+1)\times(2N+1)}$, with

$$\Lambda = \mathrm{diag}\{\lambda_{-N},\cdots,\lambda_N\}, \qquad S = (s_{n,j})_{n,j\in\{-N,\cdots,N\}}, \tag{64}$$

and

$$\lambda_n = \frac{s_1(E-\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k_n)+\varepsilon_p} + \frac{s_{-1}(E+\varepsilon_p,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k_n)-\varepsilon_p}$$
$$+ \frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}|_{E=E(k_n)}, \tag{65}$$

$$s_{n,j} = \frac{s_1(E,E+\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(g_j(\mathscr{K}_\alpha(E)) + g_j(-\mathscr{K}_\alpha(E))\Big)|_{E=E(k_n)-\varepsilon_p}$$
$$+ \frac{s_{-1}(E,E-\varepsilon_p)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(g_j(\mathscr{K}_\alpha(E)) + g_j(-\mathscr{K}_\alpha(E))\Big)|_{E=E(k_n)+\varepsilon_p}$$
$$+ \frac{s_0(E,E)}{\mathscr{K}_\alpha(E)/(1+2\alpha E)}\Big(g_j(\mathscr{K}_\alpha(E)) + g_j(-\mathscr{K}_\alpha(E))\Big)|_{E=E(k_n)}, \tag{66}$$

and $n, j = -N, \cdots, N$. Given the notation is self-explained, the negative sub-indices are used for the entry of S for simplicity.

Similar to other collocation methods, our Fourier-collocation scheme does not satisfy mass conservation property. In terms of approximating the equilibrium of the scattering operator, this spectral method performs quite differently from the other methods in previous sections, see numerical examples in Section 2.3.4.

2.3 Numerical experiments

In this section, we will demonstrate the performance of the numerical schemes when they are applied to two examples with the following parameter choices.

- **Parameter choice 1**. We consider the parabolic energy band model with $\alpha = 0$ in (3). There is no elastic collision, that is, $s_0 = 0$. In addition, we take the phonon energy $\varepsilon_p = 0.1$, lattice temperature $T = 0.0883$, transfer rate parameter $s_{-1} = 1$, and the maximum energy $E_{max} = 8$.
- **Parameter choice 2**. We use the non-dimensionalized parameters for silicon, and this involves $\alpha = 0.01292$ in the energy model (3), phonon energy $\varepsilon_p = 2.43723$, lattice temperature $T = 1$, transfer rates $s_0 = 0.26531$ and $s_{-1} = 0.04432$, and the maximum energy $E_{max} = 16$.

Throughout, μ_j is the the eigenvalue of M which has the j-th largest real part, $j = 1, 2, \cdots$.

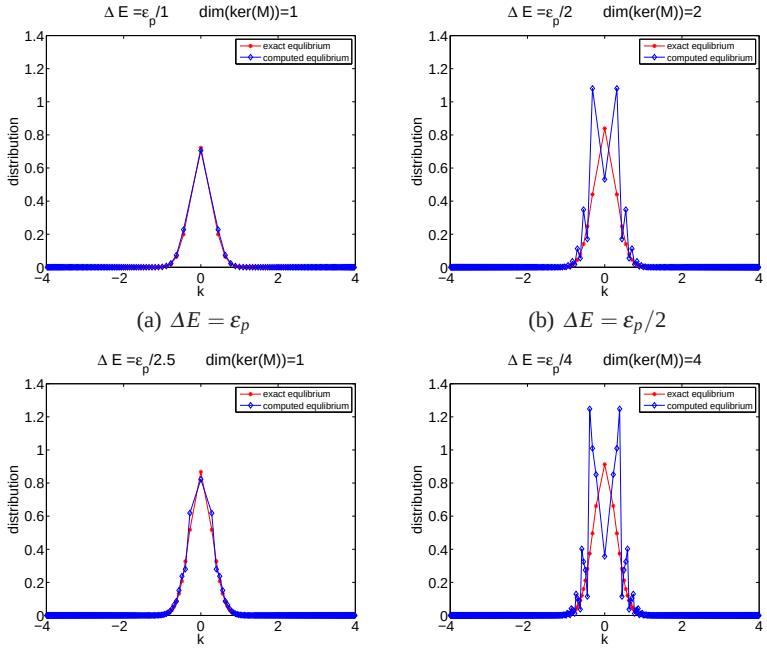
2.3.1 First order discontinuous Galerkin method

In this subsection, we shall verify the results of the first order DG method. Notice that this method has also been studied numerically in [7] for the parabolic energy band model without the elastic term.

We use a uniform grid in the energy space with cell size ΔE. The method (25) is implemented with backward Euler method applied in time and $\Delta t = \Delta E$. The initial data is randomly generated, and it is non-negative and normalized to have the same total mass as the exact equilibrium in (3). The criteria for stopping the time evolution is set to be $||\mathbf{g}^{\text{old}} - \mathbf{g}^{\text{new}}||_2, ||\mathbf{h}^{\text{old}} - \mathbf{h}^{\text{new}}||_2 \leq 10^{-7}$. The entries of S are computed using a mid-point rule quadrature, while the entries of Λ are obtained based on the column sum of M being zero.

Figures 1 and 2 contain comparison of one exact equilibrium and the computed equilibrium based on parameter choices 1 and 2. Here and in all the figures throughout Section 2.3, the exact equilibrium is taken as $f^e(k) = cf^G(k)$, where the normalized constant c is chosen so as to achieve the same total mass as the numerical solution. Figure 1 agrees well with the theory obtained in [7] (also see Section 2.2.1). When $\Delta E = \varepsilon_p/n$, and n is not an integer or $n = 1$, the matrix is irreducible, making the computed equilibrium closer to $f^G(k)$ qualitatively. When $n > 1$ is an integer, the

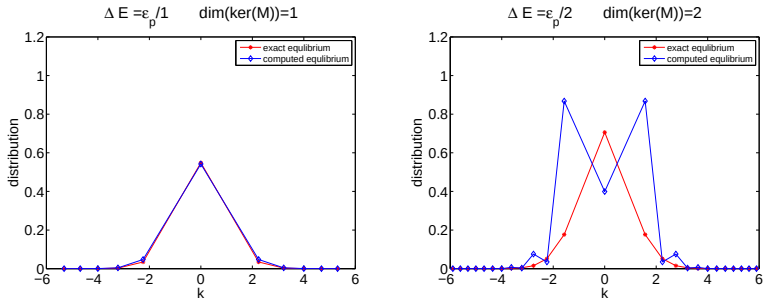
scattering matrix M (hence $\mathbb{S}$) is reducible, and the dimension of its null space, also called kernel space, is bigger than one, specifically, $dim(ker(M)) = n$. In this case, the computed equilibrium distribution is no longer monotone in each half of the domain. Just as observed in Figure 2 of [7], each monotone subregion of the computed equilibrium involves n points on the grid, and this implies that the computed equilibrium is approximately in the form of $\hat{h}(E(k))\widehat{f^G}(k)$, where $\hat{h}(E)$ is approximating a ε_p-periodic grid-based function defined on the mesh grid of the energy domain, and $\widehat{f^G}(k)$ is an approximation for $f^G(k)$. In other words, the computed equilibrium captures the characteristics of the exact equilibriums. Computations based on parameter choice 2, which uses the Kane energy band model and has elastic scattering, as demonstrated in Figure 2 give a similar conclusion, verifying our claim in Section 2.2.1 and the results in [11].



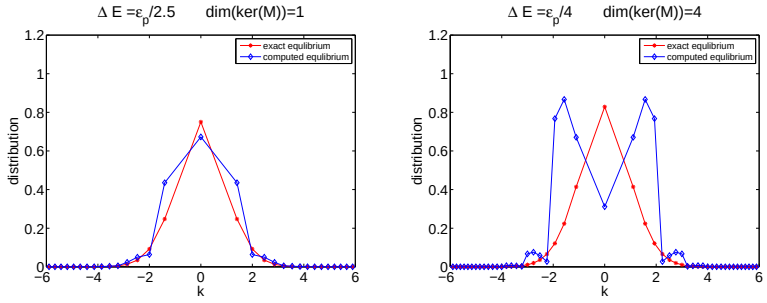
(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$

(c) $\Delta E = \varepsilon_p/2.5$ (d) $\Delta E = \varepsilon_p/4$

Fig. 1 The comparison of the exact equilibrium and the computed equilibrium by DG method with P^0 discrete space. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in E with the indicated mesh size. Here and in all the figures throughout Section 2.3, the exact equilibrium is taken as $f^e(k) = cf^G(k)$ with some normalized constant c. Parameter choice 1.



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$



(c) $\Delta E = \varepsilon_p/2.5$ (d) $\Delta E = \varepsilon_p/4$

Fig. 2 The comparison of the exact equilibrium and the computed equilibrium by DG method with P^0 discrete space. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in E with the indicated mesh size. Parameter choice 2.

2.3.2 Second order discontinuous Galerkin method

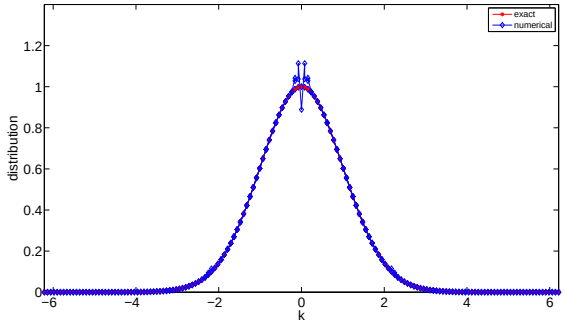
In this subsection, we will present numerical experiments with the DG method using the P^1 discrete space introduced in Section 2.2.2. Particularly, we will investigate the importance of sufficiently accurate numerical quadratures, the dimension of $ker(M)$, and the accuracy of the scheme.

A close examination reveals that the integrals for computing the entries of Λ and S involve $E^{-1/2}$-type singularity near $E = 0$. In our implementation, the following strategy is adopted to compute $S_{i,j}, \Lambda_j$: When $j \leq n_{\text{singular}}$, we apply a special 6th order quadrature, obtained from the Trapzoidal rule with Alpert correction to the left end of the reference element [1]; When $j > n_{\text{singular}}$, the standard 5-point Gauss quadrature is applied. To illustrate the effect of numerical quadratures, we consider the method implemented on a uniform mesh in k and $\Delta k = K_{\max}/N$. The first 3 eigenvalues $\mu_{1,2,3}$ with the largest real part are reported in Table 1 for $N = 80$, and $n_{\text{singular}} = N/8, 2N/8, 3N/8$ and $4N/8$ with parameter choice 2. One can see that numerical quadratures with sufficiently large n_{singular} ensures that μ_1 is an accurate approximation for the zero eigenvalue, instead of contributing to a non-trival growing mode. We further march the scheme with the equilibrium in (4) as the initial data

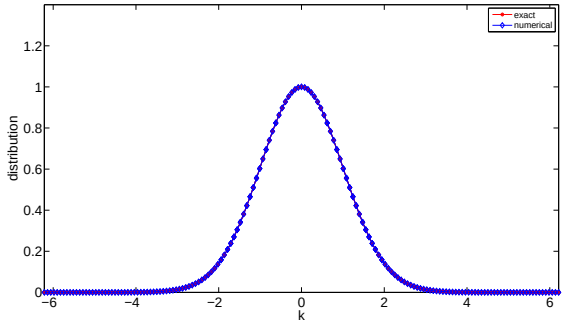
and Trapezoidal method in time with $\Delta t = \Delta k$, and plot in Figure 3 the numerical equilibriums compared with the exact one (again given by (4)) at time $t = 7$. The results confirm again the importance of using accurate enough numerical quadratures. In fact with n_{singular} being large enough, the numerical eigenvalues, other than those approximating zero, always have negative real part.

Table 1 Effect of numerical quadrature by taking different values of n_{singular}. Uniform mesh in k with $\Delta k = K_{\max}/N$ and $N = 80$. Parameter choice 2.

n_{singular}	μ_1	$\mu_{2,3}$
$N/8$, $2N/8$	1.31e-04	-1.02e-08 $\pm$ 9.39e-09i
$3N/8$	2.27e-12	-1.69e-08 $\pm$ 1.30e-08i
$4N/8$	4.33e-13	-1.69e-08 $\pm$ 1.30e-08i



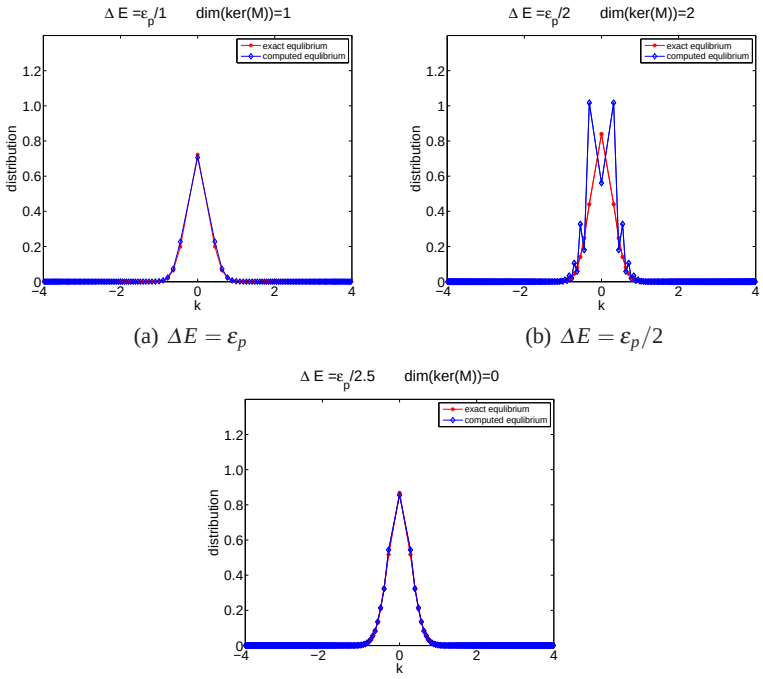
(a) $n_{\text{singular}} = N/4$



(b) $n_{\text{singular}} = 3N/8$

Fig. 3 Effect of numerical integration. Uniform mesh in k with $\Delta k = K_{\max}/N$ and $N = 80$. Trapezoidal method in time with $\Delta t = \Delta k$. Initial condition is the exact equilibrium in (4) with parameter choice 2.

Next we examine how well our scheme approximates the dimension of $ker(M)$. Motivated by the P^0 results, we implement our DG method with the P^1 space on uniform meshes in E and $\Delta E = \varepsilon_p/n$. Both parameter choices are examined, with $n_{\text{singular}} = N/8$ for parameter choice 1 and $n_{\text{singular}} = 3N/8$ for parameter choice 2. When $n = m$ is an integer, the dimension of $ker(M)$ is m; and when $n = m + \frac{1}{2}$, the *numerical* dimension of $ker(M)$ is 1 in the sense that $\mu_1 = O(10^{-12,-13})$ and $\mu_2 = O(10^{-3,-5})$. This has been tested for $m = 1, \cdots, 10$. In Figure 4, we also plot the numerical equilibrium computed from marching the scheme in time with backward Euler method and $\Delta t = \Delta E$, $n = 1, 2, 2.5$. (When $n = 2.5$, the numerical dimension of $ker(M)$ is 1.) Though there is no mathematical analysis available, our numerical results seem to imply that the dependence of the (numerical) dimension of $ker(M)$ on the choice of the mesh grid in E for the DG method with the P^1 space is similar to that with the P^0 space. The computed equilibrium also shows the characteristics in (6) of the exact equilibriums. The setup for initialization and the stoping criteria is taken the same as in Section 2.3.1.



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$

(c) $\Delta E = \varepsilon_p/2.5$

Fig. 4 The comparison of the exact equilibrium and the computed equilibrium by DG method with P^1 discrete space. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform meshes in E with the indicated mesh size. Parameter choice 1 and $n_{\text{singular}} = N/8$.

Finally we turn to the accuracy of the scheme. In Table 2, we report the L_2 errors and convergence orders of the method at a fixed time $t = 7$ for both parameter choices. Uniform meshes in k are considered with $\Delta k = K_{\max}/N$, and the initial condition is taken to be the exact equilibrium in (4). Second order accuracy is confirmed. In addition, the leading eigenvalue μ_1 of M is also reported. Although this eigenvalue is not always negative, it converges to the zero eigenvalue as meshes are refined.

Table 2 Accuracy of the DG method with P^1 discrete space at $t = 7$. Uniform mesh in k with $\Delta k = K_{\max}/N$. Trapezoidal method in time with $\Delta t = \Delta k$. Initial condition is the exact equilibrium in (4). $n_{\text{singular}} = N/8$ for parameter choice 1 and $n_{\text{singular}} = 3N/8$ for parameter choice 2.

	Parameter choice 1			Parameter choice 2		
N	L_2 error	order	μ_1	L_2 error	order	μ_1
40	4.83e-03	-	2.7125e-08	2.67e-03	-	6.0254e-10
80	1.81e-03	1.42	4.4021e-09	7.56e-04	1.82	2.2689e-12
160	4.68e-04	1.95	4.1503e-11	1.91e-04	1.98	-1.3997e-13
320	1.23e-04	1.92	-1.3055e-12	4.73e-05	2.02	-8.9108e-15
640	3.01e-05	2.03	-4.4868e-14	1.22e-05	1.95	-

2.3.3 First order collocation method

In this subsection, we will perform numerical study of the first order collocation method as outlined in Section 2.2.3. We compute the equilibrium using the backward Euler method, random initial data and stopping criteria $||\mathbf{g}^{\text{old}} - \mathbf{g}^{\text{new}}||_2, ||\mathbf{h}^{\text{old}} - \mathbf{h}^{\text{new}}||_2 \leq 10^{-7}$. We consider both parameter choices 1 and 2 on uniform meshes in E or k. Since the collocation method does not achieve mass conservation, all computed equilibrium has been rescaled so that $\sum_i f_i \Delta k_i$ agrees with the exact equilibrium. To investigate the detailed performance of the method, we also obtain the leading eigenvalues of the scattering matrix M.

Figures 5 to 7 contain simulation results with parameter choice 1 on uniform meshes in E. The collocation points $\{\xi_i\}$ are chosen such that they correspond to midpoints in the computational grid for the energy variable. In particular, Figure 5 plots the results when $\Delta E = \varepsilon_p/n$, when n is an integer; while Figure 6 plots the solutions when n is not an integer. When compared with the first order Galerkin method, we can see that the results are similar when n is an integer, i.e. any integer $n > 1$ will yield the dimension of the kernel of the scattering matrix M to be bigger than one, producing oscillatory numerical equilibriums. However, the main difference occurs when n is a non-integer. From Figure 6, we can see when $n = 1.7, 2.2, 2.7, 3.2$, unlike DG method P^0 case, the collocation method still have $dim(ker(M)) > 1$. When $n = 2.5$, we can observe even from Figure 7(d) that the scattering matrix has several positive eigenvalues, which makes the time evolution scheme not converge to a steady state. Preliminary numerical tests show similar conclusions when $n = 1.5, 2.5, 3.5 \ldots 10.5$. From our numerical tests, it seems that

if $n \in (N_n - 0.5, N_n + 0.5)$, where N_n is an integer, then $dim(ker(M)) = N_n$. We believe the different behavior of the collocation method when compared with the Galerkin method is because of the point-based nature of the collocation scheme. However, due to the lack of theoretical studies, we leave the detailed interpretation of this result to future work.

The next set of numerical tests were performed on uniform meshes in E with parameter choice 2. Figures 8 and 9 plot the equilibrium and the leading eigenvalues of the scattering matrix when $n = 1, 2, 1.7, 2.2$. Those selective results show $dim(ker(M)) = 0$ in all cases. However, for $n = 1.7, 2, 2.2$, there are two eigenvalues that are very close to zero, see Figure 8(e) for details. With the numerical dimension of $ker(M)$ being considered, the conclusion for parameter choice 2 are the same as the ones for parameter choice 1.

Finally we plot the results for parameter choice 1 on uniform mesh in k when $N = 40, 80, 120, 160$ in Figure 10. The collocation points $\{\xi_i\}$ are chosen to be the midpoint in each cell in the k variable. Unlike the results for uniform mesh on E, the result for uniform mesh in k is not conclusive, i.e. this mesh choice does not imply the scattering matrix to be reducible/irreducible.

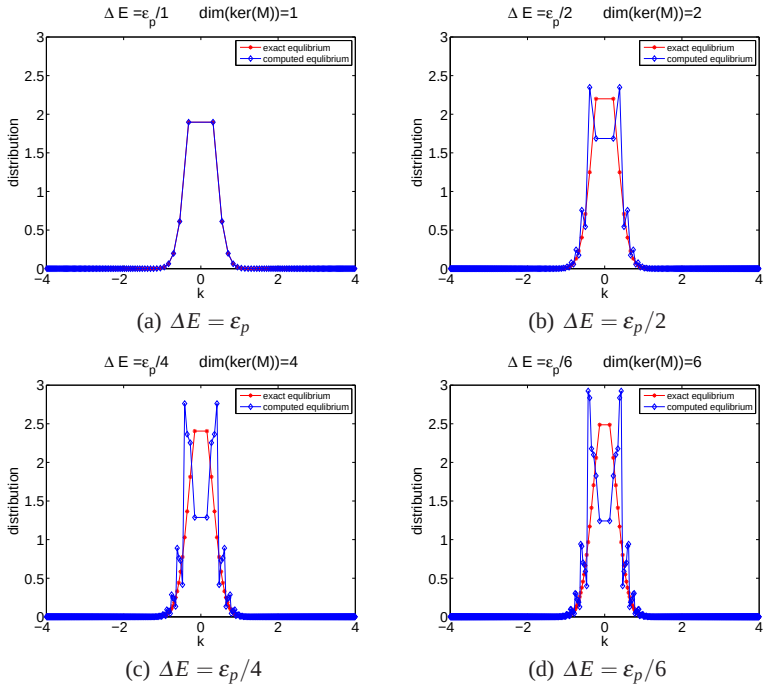
In summary, the first order collocation method does not outperform the first order Galerkin scheme when measuring the qualitative behavior of the computed equilibrium. The collocation method, though being more computationally efficient, does not preserve mass conservation, and the results are highly dependent upon the choice of collocation points.

2.3.4 Fourier-collocation method

In this subsection, we will demonstrate the performance of the Fourier-collocation method defined in Section 2.2.4. This method behavies very differently from those we have examined so far, and it only captures the one-dimensional equilibrium $cf^G(k)$ (with $c > 0$ being a constant), given the computational domain is large enough. We attribute this to the global nature of this spectral method. By looking into the eigenvalues and corresponding eigenvectors of the scattering matrix $\mathbb{S}$ with parameter choice 1 (in Table 3) and parameter choice 2 (in Table 4 and Table 5), the following observations can be made.

- **Parameter choice 1**

  1. With the meshes being refined, the leading eigenvalue μ_1 is approaching 0 exponentially and $\mu_2 = O(10^{-3})$. When $N = 34$, this eigenvalue is zero at the roundoff error level. The numerical dimension of the null space is one. In this case, we take $E_{\max} = 8$ and hence $K_{\max} = \mathscr{K}_\alpha(E_{\max}) = 4$.
  2. The eigenvector corresponding to μ_1 approximates the equilibrium $f^G(k)$. For comparison, the eigenvector is scaled such that the sum of its values at collocation points is the same as that of the exact equilibrium. In Table 3, we present errors of the computed eigenvector in l^∞, l^1 and l^2 vector norms. We



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$

(c) $\Delta E = \varepsilon_p/4$ (d) $\Delta E = \varepsilon_p/6$

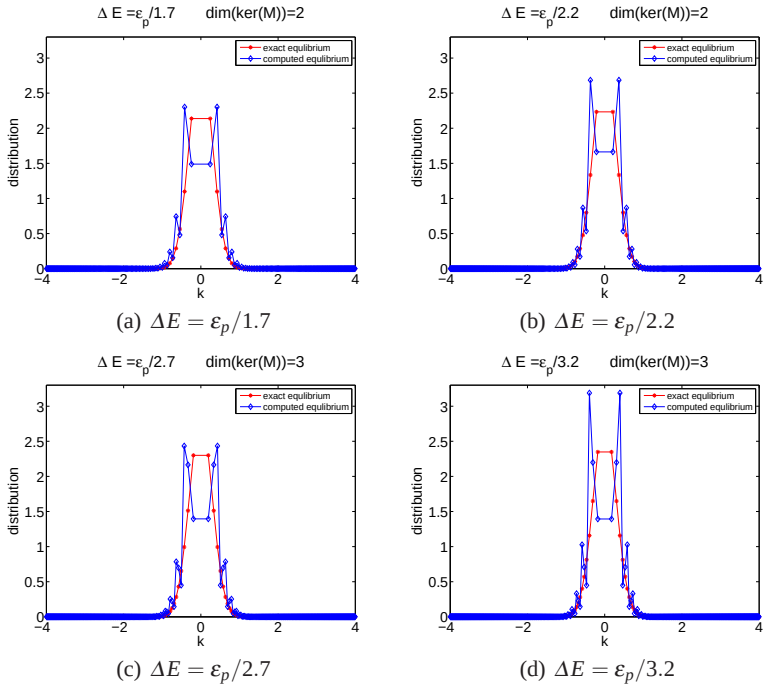
Fig. 5 The comparison of the exact equilibrium and the computed equilibrium by first order collocation method. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in E with the indicated mesh size. Parameter choice 1.

also plot the scaled eigenvector with $N = 34$ in Figure 11, which captures the equilibrium with an error at the level of 10^{-13}.

3. The numerical equilibrium is also obtained by computing the steady state of the ODE system (63). The non-negative initial data is chosen randomly, with the stopping criteria as $\|\mathbf{f}^{old} - \mathbf{f}^{new}\|_\infty \leq 10^{-10}$. In Figure 12, we compare the computed and the exact equilibrium. Though both the computed equilibriums before and after normalization well capture the shape of the equilibrium, the normalized one has a much smaller error at the level of 10^{-9}.

Table 3 Eigenvalues μ_1 and μ_2, together with the errors between the eigenvectors (normalized) corresponding to μ_1 and the exact equilibrium $f^G(k)$. Parameter choice 1.

N	Re(μ_1)	Re(μ_2)	Errors of the eigenvectors		
			l^∞	l^1	l^2
16	5.94e-04	-8.07e-03	9.73e-02	1.80e-02	2.79e-02
20	4.94e-05	-3.19e-03	3.16e-03	5.83e-04	9.03e-04
30	1.77e-11	-8.28e-03	1.78e-09	3.30e-10	5.11e-10
32	3.00e-13	-9.07e-03	4.88e-11	9.02e-12	1.40e-11
34	1.04e-15	-6.80e-03	5.55e-13	1.03e-13	1.60e-13



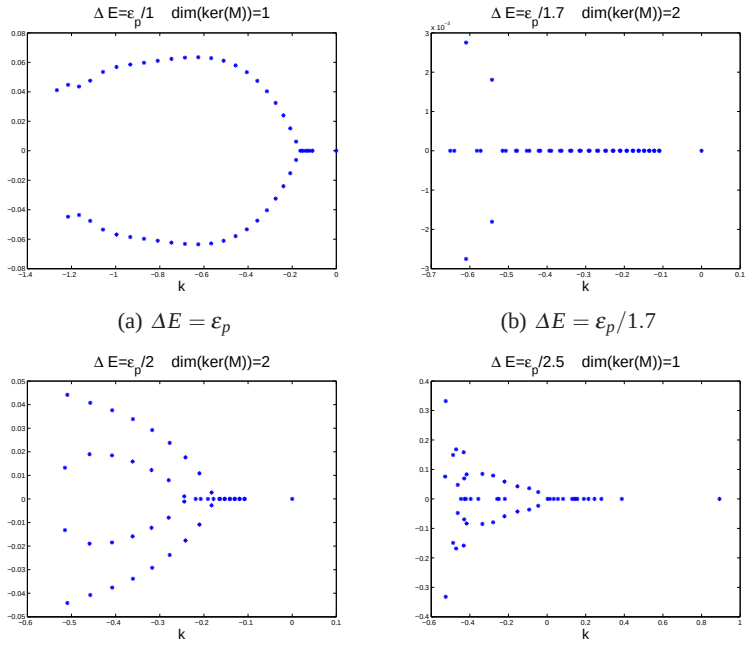
(a) $\Delta E = \varepsilon_p/1.7$ (b) $\Delta E = \varepsilon_p/2.2$

(c) $\Delta E = \varepsilon_p/2.7$ (d) $\Delta E = \varepsilon_p/3.2$

Fig. 6 The comparison of the exact equilibrium and the computed equilibrium by first order collocation method. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in E with the indicated mesh size. Parameter choice 1.

- **Parameter choice 2**
  We start with taking $E_{\max} = 16$ in the computation.

  1. Similar to parameter choice 1, the eigenvalue μ_1 of M is approaching 0 when N increases, with the convergence speed seemingly faster than that for parameter choice 1. At the same time, μ_2 is $O(10^{-3,-4})$. The results in Table 4 are reported for N up to 10.
  2. The eigenvector corresponding to μ_1 approximates the equilibrium $f^G(k)$. In Table 4, we report the errors between the scaled eigenvector and the exact equilibrium. For $N = 10$, the eigenvector approximates the equilibrium with an error at the level of 10^{-6}, as in Figure 13.
  3. The numerical equilibrium is also obtained by computing the steady state of the ODE system (63). The non-negative initial data is chosen randomly, with the stopping criteria as $\|\mathbf{f}^{old} - \mathbf{f}^{new}\|_\infty \leq 10^{-8}$. In Figure 14, we compare the computed and the exact equilibria with parameter choice 2 and $N = 10$. Again, the computed equilibrium after normalization has a smaller error at the level of 10^{-6}.



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/1.7$

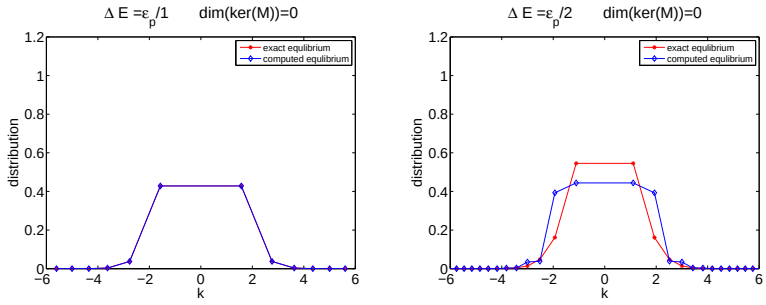
(c) $\Delta E = \varepsilon_p/2.$ (d) $\Delta E = \varepsilon_p/2.5$

Fig. 7 The distribution of the first 50 eigenvalues of the scattering matrix by first order collocation method on uniform mesh in E with the indicated mesh size. Parameter choice 1.

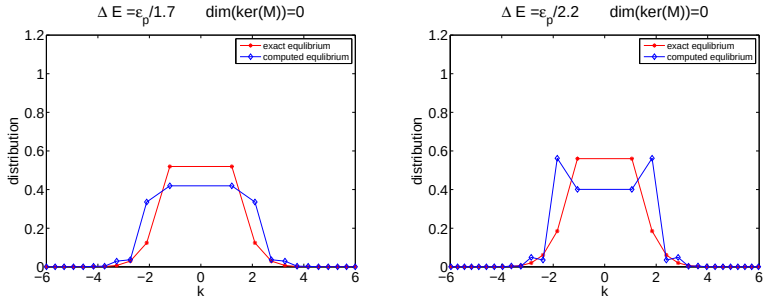
Table 4 Eigenvalues μ_1 and μ_2, together with the errors between the eigenvectors (normalized) corresponding to μ_1 and the exact equilibrium $f^G(k)$. Parameter choice 2. $E_{\max} = 16$.

			Errors of the eigenvectors		
N	Re(μ_1)	Re(μ_2)	l^∞	l^1	l^2
6	6.45e-2	-1.24e-03	6.43e-4	1.82e-2	2.38e-2
7	-1.19e-5	-3.36e-03	2.26e-3	8.89e-4	1.18e-3
8	7.12e-6	-3.89e-03	5.22e-4	1.29e-4	1.69e-4
9	-4.13e-7	-8.58e-04	7.73e-5	1.70e-5	2.47e-5
10	1.76e-8	-7.62e-04	1.54e-5	3.15e-6	5.42e-6

The results we have shown here are for N up to 10. For some larger values of N, it is observed that more than one computed eigenvalues of $\mathbb{S}$ can approach 0. There can also be multiple eigenvalues which have positive real parts. This is because the computational domain is not chosen large enough. To see this, we further test the method on a larger computational domain with $E_{\max} = 32$. Again, the method captures the equilibrium $f^G(k)$ with the eigenvector corresponding to μ_1, as in Table 5 and Figure 15. With $N = 27$, μ_1 is $O(10^{-15})$, and the scaled eigenvector corresponding to μ_1 approximates $f^G(k)$ with an error at the level of 10^{-10}.



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$



(c) $\Delta E = \varepsilon_p/1.7$ (d) $\Delta E = \varepsilon_p/2.2$

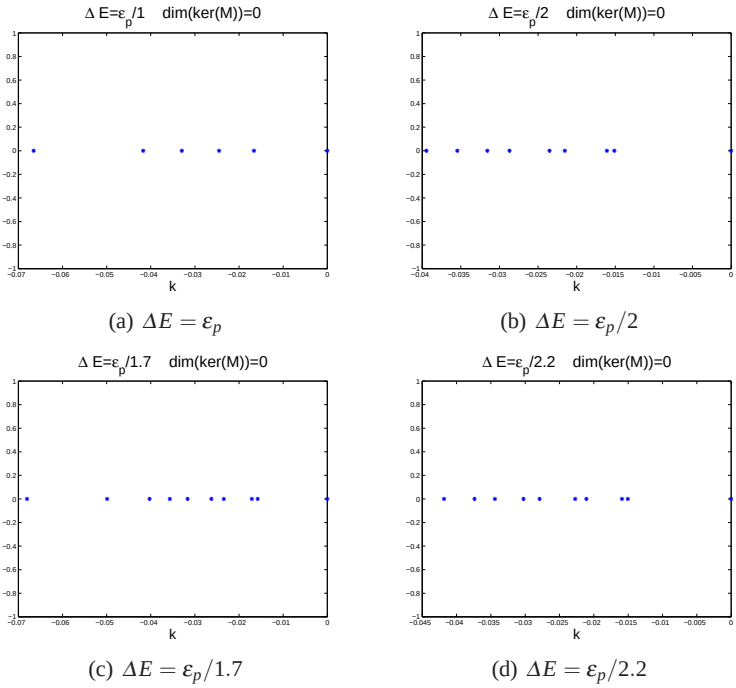
n	μ_1	μ_2	μ_3
1.7	-3.271e-09	-5.305e-08	-0.0158
2	-1.696e-10	-3.377e-09	-0.0151
2.2	-1.136e-10	-2.249e-10	-0.0151

(e) Leading eigenvalues of M: μ_1, μ_2, μ_3 are the eigenvalues with the three largest real part.

Fig. 8 The comparison of the exact equilibrium and the computed equilibrium. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in E with the indicated mesh size. Parameter choice 2.

Table 5 Eigenvalues μ_1 and μ_2, together with the errors between the eigenvectors (normalized) corresponding to μ_1 and the exact equilibrium $f^G(k)$. Parameter choice 2. Larger domain size $E_{\max} = 32$.

N	Re(μ_1)	Re(μ_2)	Errors of the eigenvectors		
			l^∞	l^1	l^2
11	-2.45e-05	-8.43e-04	1.37e-02	3.58e-03	4.91e-03
15	1.47e-07	3.60e-04	8.08e-05	2.30e-05	3.00e-05
20	-3.07e-12	1.49e-04	5.62e-08	6.89e-09	1.55e-08
22	9.58e-13	-1.34e-04	6.38e-09	8.58e-10	1.91e-09
27	1.40e-15	1.86e-05	7.73e-10	9.54e-11	2.22e-10



(a) $\Delta E = \varepsilon_p$ (b) $\Delta E = \varepsilon_p/2$

(c) $\Delta E = \varepsilon_p/1.7$ (d) $\Delta E = \varepsilon_p/2.2$

Fig. 9 The distribution of the first 11 eigenvalues of the scattering matrix by first order collocation method on uniform mesh in E with the indicated mesh size. Parameter choice 2.

3 A numerical method for continuous scattering kernels

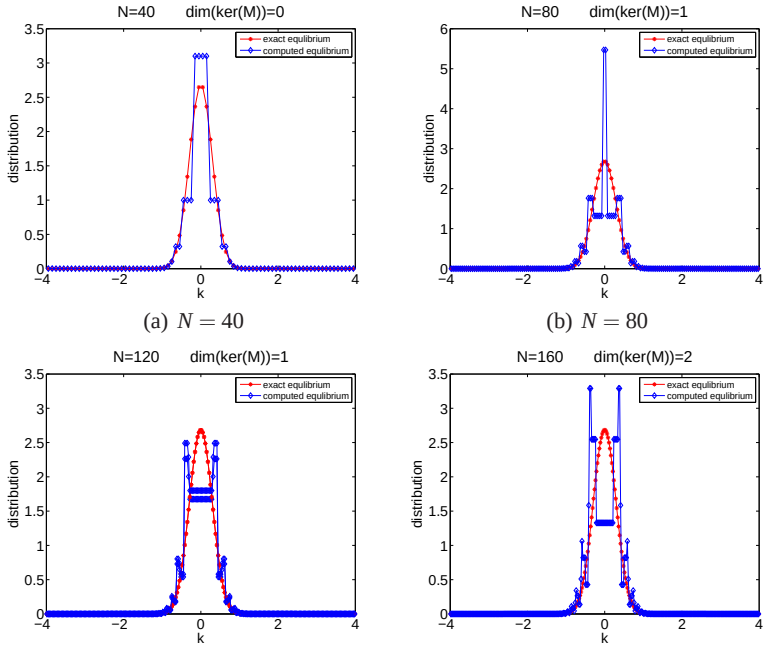
In this section we consider the kinetic model (1) with a continuous scattering kernel (7). If one follows the derivation in Section 2.2.1 to define a first order DG method for this model, it is easy to show that the scattering matrix is always irreducible when $\sigma(k,k') > 0$. Instead, we choose a different discretization which is well-suited for the model with a continuous scattering kernel.

Since the scattering kernel $S(k,k')$ has Gaussian decay and we are concerned with approximating the equilibrium solution, we assume there exists a constant $K_{\max}$ such that

$$\left| \int_{-\infty}^{\infty} \left(S(k',k) f(k',t) - S(k,k') f(k,t) \right) dk' - \int_{-K_{\max}}^{K_{\max}} \left(S(k',k) f(k',t) - S(k,k') f(k,t) \right) dk' \right| < \varepsilon$$

for a user prescribed tolerance ε for all k.

Equation (1) can now be discretized by applying numerical quadrature to the truncated domain. This technique is called *Nyström discretization* of the integral differential equation. Specifically, let $\{k_i\}_{i=1}^N$ denote the set of quadrature nodes in



(a) $N = 40$ (b) $N = 80$

(c) $N = 120$ (d) $N = 160$

Fig. 10 The comparison of the exact equilibrium and the computed equilibrium. The computed equilibrium is obtained by the backward Euler method with random initial data on uniform mesh in k with the indicated mesh size. Parameter choice 1.

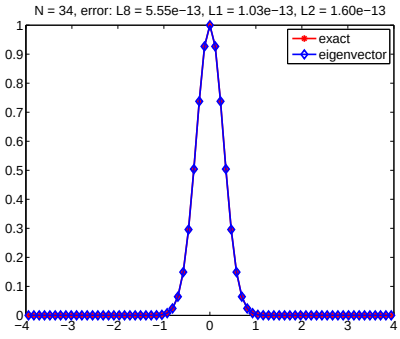
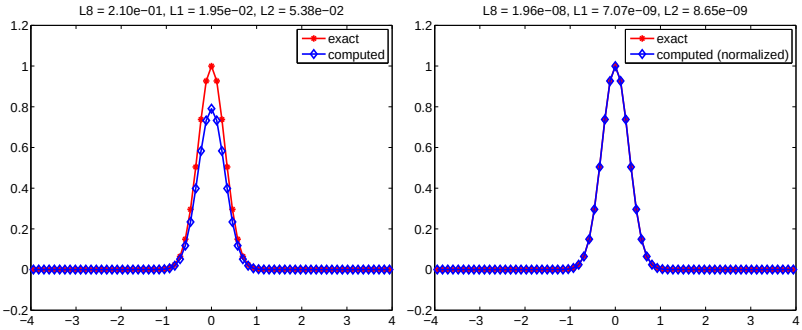


Fig. 11 The normalized eigenvector corresponding to μ_1 for $N = 34$ with exact equilibrium $f^G(k)$. Parameter choice 1.



(a) Computed equilibrium before normalization (b) Computed equilibrium after normalization

Fig. 12 The comparison of the exact equilibrium $f^G(k)$ and the computed equilibrium, obtained by the backward Euler method with random initial data and tolerance being $1.e-10$. $N=34$. Parameter choice 1.

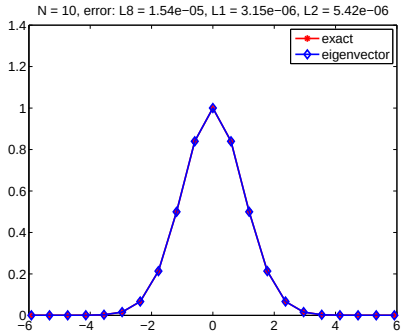


Fig. 13 The normalized eigenvector corresponding to μ_1 for $N=10$ with exact equilibrium $f^G(k)$. Parameter choice 2. $E_{\max}=16$.

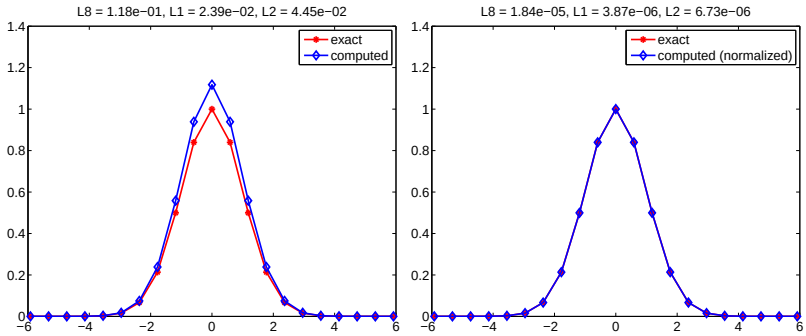
the interval $[-K_{\max},K_{\max}]$ with corresponding weights $\{w_i\}_{i=1}^N$, then (1) is approximated by

$$\frac{\partial \hat{f}(k,t)}{\partial t}=\sum_{i=1}^{N}\left(S(k_i,k)\hat{f}(k_i,t)-S(k,k_i)\hat{f}(k,t)\right)w_i \tag{67}$$

where the solution $\hat{f}$ is an approximation to the exact solution f of (1). The quadrature points $\{k_i\}_{i=1}^N$ will be the discretization points.

To arrive at a linear system, the solution $\hat{f}$ is sought at the quadrature points k_j, $j=1,\ldots,N$ for all t. The result is the following discrete ordinary differential equation

$$\frac{\partial \hat{f}(k_j,t)}{\partial t}=\sum_{i=1}^{N}\left(S(k_i,k_j)\hat{f}(k_i,t)-S(k_j,k_i)\hat{f}(k_j,t)\right)w_i \tag{68}$$



(a) Computed equilibrium before normalization (b) Computed equilibrium after normalization

Fig. 14 The comparison of the exact equilibrium $f^G(k)$ and the computed equilibrium, obtained by the backward Euler method with random initial data and tolerance being $1.e-8$. $N = 10$. Parameter choice 2. $E_{\max} = 16$.

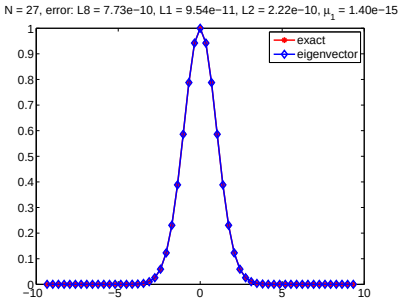


Fig. 15 The normalized eigenvector corresponding to μ_1 for $N = 27$ with exact equilibrium $f^G(k)$. Parameter choice 2. Larger domain size $E_{\max} = 32$.

for each $j = 1,\ldots,N$ or in linear algebraic form

$$\frac{\partial \hat{\boldsymbol{f}}}{\partial t}(t) = (\mathsf{S} - \Lambda)\hat{\boldsymbol{f}}(t) = \mathsf{M}\hat{\boldsymbol{f}}(t) \tag{69}$$

where $\mathsf{S}_{i,j} = S(k_j,k_i)w_j$, $\hat{\boldsymbol{f}}$ denotes the vector of unknowns such that $\hat{\boldsymbol{f}}_j = \hat{f}(k_j,t)$, $\Lambda = \mathrm{diag}\{\boldsymbol{\nu}\}$ and the vector $\boldsymbol{\nu}$ has entries given by $\nu_i = \sum_{j=1}^{N} S(k_i,k_j)w_j$.

Remark 3.1 *Applying the numerical quadrature scheme to (68) (i.e. left multiplying (69) by $\boldsymbol{w}^T$ where $\boldsymbol{w}_j = w_j$) shows that the discretization technique conserves total mass in time.*

3.1 Numerical experiments

The performance of the numerical method is explored in this section with two choices of $\sigma(k,k')$. In Section 3.1.1, the choice of σ results in a problem with a known solution while in Section 3.1.2 the choice of σ yields a problem without a reference solution.

For the numerical experiments, a ten-point composite Gaussian quadrature on equispaced panels is utilized to approximate the solution over the interval $[-K_{\max}, K_{\max}] = [-4,4]$. Thus, the number of discretization points N is ten times the number of panels placed on the interval $[-4,4]$.

3.1.1 An example with a known solution

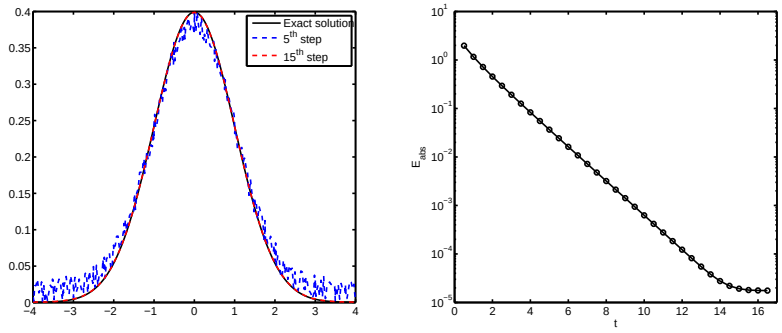
In this subsection, we illustrate the performance of the numerical method when $\sigma(k,k') = 1$. With this choice of σ, the exact solution is known to be $f_{\text{ex}}(k) = \frac{1}{\sqrt{2\pi}} e^{-k^2/2}$. Let $\boldsymbol{f}_{\text{ex}}$ denote the vector whose entries are f_{ex} evaluated at the discretization points.

Table 6 reports the number of discretization points N, the absolute error $E_{\text{abs}} = \|\hat{\boldsymbol{f}} - \boldsymbol{f}_{\text{ex}}\|_2$ and the relative error $E_{\text{rel}} = \frac{\|\hat{\boldsymbol{f}} - \boldsymbol{f}_{\text{ex}}\|_2}{\|\boldsymbol{f}_{\text{ex}}\|_2}$ when computing the equilibrium solution, i.e. approximating solutions to (1) with $\frac{\partial f}{\partial t} = 0$. The numerical approximation is found by computing the null space of M in equation (69).

Table 6 The number of discretization points N, absolute error E_{abs} and the relative error E_{rel} when applying the solution technique to equation (1) with $\sigma(k,k') = 1$.

N	E_{abs}	E_{rel}
10	9.49e-02	1.94e-01
20	1.32e-03	1.36e-03
40	4.13e-04	3.40e-04
80	1.43e-04	8.51e-05
160	5.05e-05	2.13e-05
320	1.79e-05	5.32e-06
640	6.32e-06	1.33e-06
1280	2.23e-06	3.33e-07
2560	7.90e-07	8.31e-08

Next, the backward Euler method was applied to (67) with a fixed $N = 320$ number of discretization points and time step size $h = 0.5$. With this choice of N, Table 6 indicates that the expected converged accuracy should be approximately 1e-05. Thus the iterative process is stopped when the norm of the difference between two iterates is less than 1e-05. Figure 16(a) illustrates the approximate solution at two different times in addition to the exact solution. Figure 16(b) illustrates the absolute error E_{abs} at each time step. At the thirty-third time step, the scheme has converged to the set tolerance.



(a) (b)

Fig. 16 (a) Approximate solutions after 5 and 10 time steps with a step size of $h = 0.5$. (b) Absolute error E_{abs} in approximate solution at time t.

3.1.2 An example with unknown solution

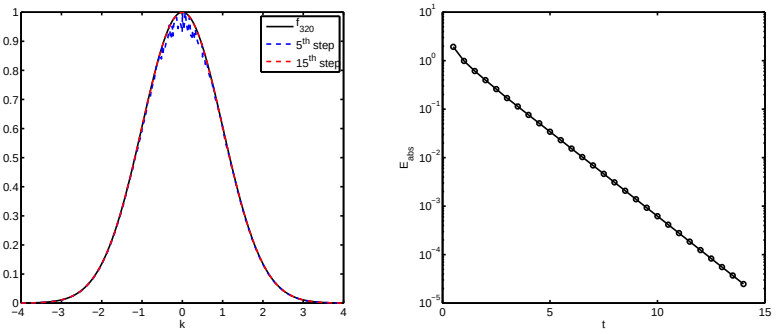
In this subsection, we consider (1) with $\sigma(k,k') = (k-k')^2$. For this choice of σ, the exact solution is unknown. In the first experiment, a convergence study is performed for the equilibrium problem. Let $\hat{\boldsymbol{f}}_N$ denote the approximate solution obtained with N discretization points. Table 7 reports the number of discretization points N, the absolute convergence error $E_{\text{abs}} = \|\hat{\boldsymbol{f}}_N - \mathsf{L}\hat{\boldsymbol{f}}_{2N}\|_2$ where L is a matrix that interpolates $\hat{\boldsymbol{f}}_{2N}$ at the $2N$ discretization points to the N coarse discretization points and the relative convergence error $E_{\text{rel}} = \frac{\|\hat{\boldsymbol{f}}_N - \mathsf{L}\hat{\boldsymbol{f}}_{2N}\|_2}{\|\mathsf{L}\hat{\boldsymbol{f}}_{2N}\|_2}$.

Table 7 The number of discretization points N, absolute error E_{abs} and the relative error E_{rel} when applying the solution technique to equation (1) with $\sigma(k,k') = (k-k')^2$.

N	E_{abs}	E_{rel}
10	2.36e-01	1.92e-01
20	2.47e-03	1.02e-03
40	7.77e-04	2.55e-04
80	2.69e-04	6.38e-05
160	9.50e-05	1.59e-05
320	3.36e-05	3.99e-06
640	1.19e-05	9.97e-07
1280	4.19e-06	2.49e-07

Again backward Euler method is employed with time step size $h = 0.5$ and $N = 320$ discretization points. We define the solution obtained by solving the equilibrium problem with $N = 320$ discretization points to be the reference solution. It takes 28 time steps for the approximate solution to converge to the reference solution. Figure 17(a) illustrates the approximate solution after 5 and 15 time steps. Figure 17(b)

illustrates the absolute approximate error given by $E_{\text{abs}} = \|\hat{\boldsymbol{f}}_{320} - \hat{\boldsymbol{f}}(t)\|_2$ where $\hat{\boldsymbol{f}}_{320}$ is the approximate equilibrium solution when $N = 320$ and $\hat{\boldsymbol{f}}(t)$ is the approximate solution at time t.



(a) (b)

Fig. 17 (a) Approximate solutions after 5 and 10 time steps with a step size of $h = 0.5$. (b) Absolute error E_{abs} in approximate solution at time t.

4 Concluding remarks

In this paper, we consider some one-dimensional space-homogeneous linear kinetic models arising from semiconductor device simulations. The focus of our efforts is to study the qualitative behaviors of the discrete scattering operators and the resulted numerical approximations for steady state equilibrium. We review and discuss the mathematical results in [7, 11] for a first order finite volume method when it is applied to a model with δ-type singularity with the Kane energy band and the additional elastic scattering. Moreover, we investigate the numerical performance of first and higher order Galerkin, a first order collocation method and a Fourier-collocation spectral method for this model, as well as a Nyström method for a kinetic model with a continuous scattering kernel.

It seems to be non-trivial to generalize the analysis developed in [7, 11] to higher order and collocation-type schemes to solve models with δ-type singularity. For second (or higher) order Galerkin methods, the scattering matrix will become block structured, which requires additional tools in algebraic analysis. For collocation schemes, the analysis breaks down because the methods are no longer mass conservative. The numerical study in this paper seems to indicate that similar conclusion as for the discontinuous Galerkin scheme with the P^0 discrete space holds for the discontinuous Galerkin scheme with the P^1 space regarding how the properties of the kernel of the discrete scattering operator depend on the mesh choices. The

first order collocation method computes numerical equilibrium that is highly dependent on the mesh, while the Fourier-collocation method, with its global nature, only captures a one-dimensional equilibrium associated with $f^G(k)$, and the resulting approximation is very accurate with the spectral accuracy of the method. These numerical results motivate our immediate future work on the theoretical analysis of some of the methods. Another interesting future direction consists of generalization to higher dimensions. Real world applications call for attention to models in higher dimensions with transport effect. Such models have different equilibria from the space homogeneous case and the analysis will be more involved.

Acknowledgements The third author was partially supported by NSF grant DMS-1318186, and the fifth author was partially supported by NSF grants DMS-0847241 and DMS-1318409.

References

1. B. K. Alpert. Hybrid Gauss-trapezoidal quadrature rules. *SIAM Journal on Scientific Computing*, 20(5):1551–1584, 1999.
2. J. A. Carrillo, I. M. Gamba, A. Majorana, and C.-W. Shu. A WENO-solver for the transients of Boltzmann-Poisson system for semiconductor devices: performance and comparisons with Monte Carlo methods. *Journal of Computational Physics*, 184(2):498–525, 2003.
3. Y. Cheng, I. M. Gamba, A. Majorana, and C.-W. Shu. A discontinuous Galerkin solver for Boltzmann-Poisson systems in nano devices. *Computer Methods in Applied Mechanics and Engineering*, 198(37):3130–3150, 2009.
4. E. Fatemi and F. Odeh. Upwind finite difference solution of Boltzmann equation applied to electron transport in semiconductor devices. *Journal of Computational Physics*, 108(2):209–217, 1993.
5. D. K. Ferry. *Semiconductors*. IoP Publishing, 2013.
6. Y. L. Le Coz. *Semiconductor device simulation: a spectral method for solution of the Boltzmann transport equation*. PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 1988.
7. R. Li, T. Lu, and W. Yao. Discrete kernel preserving model for 1d electron-optical phonon scattering. *Journal of Scientific Computing*, 62(2):317–335, 2015.
8. M. Lundstrom. *Fundamentals of carrier transport*. Cambridge University Press, 2009.
9. A. Majorana. Equilibrium solutions of the non-linear boltzmann equation for an electron gas in a semiconductor. *Il Nuovo Cimento B*, 108(8):871–877, 1993.
10. P. Markowich, C. Ringhofer, and C. Schmeiser. *Semiconductor equations*. Springer-Verlag, 1990.
11. W. Yao. *Simulation of one-dimensional semiconductors and eigen analysis of discrete electron optical phonon scattering*. PhD thesis, Peking University, School of Mathematical Sciences, 2014.