

## SPACE-TIME DISCRETIZATION OF SERIES EXPANSION METHODS FOR THE BOLTZMANN TRANSPORT EQUATION\*

CHRISTIAN RINGHOFER†

**Abstract.** The approximate solution of the Boltzmann transport equation via Galerkin-type series expansion methods leads to a system of first order differential equations in space and time for the expansion coefficients. This system is extremely stiff close to the fluid dynamical regime (for small Knudsen numbers), and exhibits a mildly dispersive behavior, due to the acceleration of waves by the external force (the electric field). In this paper a class of difference methods is presented and analyzed which represent a generalization of the well-known Scharfetter–Gummel exponential fitting approach for the drift-diffusion equations. It is shown that, by using appropriate operator splitting methods for the time discretization, one obtains stability properties which are only mildly dependent on the Knudsen number and essentially independent of the size of the electric field.

**Key words.** Boltzmann equation, Galerkin methods, finite differences

**AMS subject classifications.** 65N35, 65N05

**PII.** S0036142998339921

**1. Introduction.** In this paper we derive and analyze discretization methods for Galerkin-type series expansion methods for the semiconductor Boltzmann equation

$$(1.1) \quad \lambda^2 \partial_t f + \lambda \nabla_k \varepsilon \bullet \nabla_x f + \lambda E \bullet \nabla_k f = Q[f],$$

where the density function  $f(x, k, t)$  is a function of position  $x \in R_x^3$ , wave vector  $k \in R_k^3$ , and time  $t$ . The function  $\varepsilon(k)$  describes the energy band under consideration. So  $\nabla_k \varepsilon$  denotes the velocity with which an electron with wave vector  $k$  travels. (If more than one energy band is considered one density function  $f$  per band would have to be computed.) Equation (1.1) has already been brought into a scaled dimensionless form, where the parameter  $\lambda$  denotes scaled mean free path, i.e., the average distance an electron travels in one time unit before it undergoes a collision event (see [10, Chapter 1] for the details of the scaling). The function  $E(x, t)$  denotes the electric field and satisfies  $E = -\nabla_x V$ , where  $V(x, t)$  is the electric potential. Finally the collision events are modeled by the integral operator  $Q$  on the right hand side of (1.1) which is of the form

$$(1.2) \quad Q[f](x, k, t) = \int_{R_k^3} dk' [S(k, k')f(x, k', t) - S(k', k)f(x, k, t)],$$

where  $S(k, k')$  denotes the scaled scattering cross section. Because of the principle of detailed balance [3] the symmetry relation

$$(1.3) \quad S(k, k')M(k') = S(k', k)M(k), \quad M(k) = \frac{1}{M_0} \exp[-\beta\varepsilon(k)],$$

$$M_0 = \int_{R_k^3} dk \exp[-\beta\varepsilon(k)]$$

---

\*Received by the editors June 3, 1998; accepted for publication (in revised form) November 30, 1999; published electronically July 19, 2000.

<http://www.siam.org/journals/sinum/38-2/33992.html>

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (ringhofer@asu.edu). This work was supported by NSF grants DMS 970-6792 and INT 960-3253.

holds where  $M(k)$  denotes the scaled Maxwell distribution.  $\beta$  in (1.3) is a dimensionless  $O(1)$  parameter. The assumption of a linear collision operator of the form (1.2) implies that we neglect electron-electron interactions. The scattering cross section  $S$  need not necessarily be a classical function. If the emission/absorption of phonons is considered,  $S$  will be of the form

$$(1.4) \quad S(k, k') = \sum_{j=-1,1} \tilde{S}(k, k') \delta(\varepsilon(k) - \varepsilon(k') - j\omega),$$

where  $\omega$  is the energy lost/gained by the emission/absorption of a phonon.

This paper is concerned with the spatial and time discretization of first order hyperbolic systems which arise from employing a series expansion method for the Boltzmann equation (1.1) in the wave vector direction. If we expand the density function  $f$  according to

$$(1.5) \quad f(x, k, t) \approx \sum_{n=1}^N f_n(x, t) \phi_n(k),$$

where the  $\phi_n$  form an orthonormal system (ONS) with respect to a certain weight function, then

$$(1.6) \quad \int_{R_k^3} dk \quad w(k) \phi_m(k) \phi_n(k) = \delta_{mn}$$

holds, the resulting Galerkin discretization giving a first order hyperbolic system of the form

$$(1.7) \quad \lambda^2 \partial_t F + \lambda \sum_{s=1}^3 [A_s \partial_{x_s} F + E_s B_s F] = CF.$$

$F(x, t)$  in (1.7) denotes the coefficient vector  $(f_1, \dots, f_N)^T$  and the matrices  $A_s, B_s$  and  $C$  are given by

$$(1.8) \quad (a) \quad A_s(m, n) = \int_{R_k^3} dk [w(k) \phi_m(k) (\partial_{k_s} \varepsilon(k)) \phi_n(k)],$$

$$(b) \quad B_s(m, n) = \int_{R_k^3} dk [w(k) \phi_m(k) (\partial_{k_s} \phi_n(k))],$$

$$(c) \quad C(m, n) = \int_{R_k^3} dk [w(k) \phi_m(k) Q[\phi_n(k)]].$$

The quality of this approximation procedure will of course crucially depend on the choice of the ONS  $\{\phi_n\}$  and the weight function  $w$ . We will assume that the weight function is chosen as

$$(1.9) \quad w(k) = \frac{1}{M(k)} = \exp[\beta \varepsilon(k)],$$

where  $M$  denotes again the Maxwellian. This choice has the following advantage. A straightforward calculation (see [11]) gives that

$$(1.10) \quad \int_{R_k^3} dk \frac{1}{M(k)} g(x, k, t) Q[f](x, k, t) = \\ -\frac{1}{2} \int_{R_k^3} dk \int_{R_k^3} dk' \left\{ S(k, k') M(k') \left[ \frac{f(x, k, t)}{M(k)} - \frac{f(x, k', t)}{M(k')} \right] \left[ \frac{g(x, k, t)}{M(k)} - \frac{g(x, k', t)}{M(k')} \right] \right\}$$

holds. With the choice  $w = \frac{1}{M}$  for the weight function (1.10) immediately implies that

$$(1.11) \quad (a) \quad \int_{R_k^3} dk \quad wgQ[f] = \int_{R_k^3} dk \quad wfQ[g] \\ (b) \quad \int_{R_k^3} dk \quad wfQ[f] < 0 \quad (c) \quad \int_{R_k^3} dk \quad Q[f] = 0$$

holds. (1.11)(a)–(b) implies that the matrix  $C$  in (1.8) is symmetric and negative semidefinite. The choice (1.9) for the weight function is essentially a linear version of the entropy closures in [8], [9], where the entropy functional can be taken as a bilinear form (i.e., a scalar product) since we are dealing with a linear collision operator. (1.11)(c) means that the collision operator  $Q$  preserves charge. If we assume that the first basis function  $\phi_1$  is given by the Maxwellian itself, this means that the first row of the matrix  $C$  vanishes. In this case, because of (1.8)(b), the first rows of the matrices  $B_s$  vanish as well. Furthermore, we have the relation

$$(1.12) \quad 0 = \int_{R_k^3} dk \{ \partial_{k_s} [\exp(\beta\varepsilon) \phi_m \phi_n] \} = \int_{R_k^3} dk \{ w [\beta \partial_{k_s} \varepsilon \phi_m \phi_n + \partial_{k_s} \phi_m \phi_n + \phi_m \partial_{k_s} \phi_n] \} \\ = \beta A_s(m, n) + B_s(n, m) + B_s(m, n)$$

which implies for the matrices  $A_s$  and  $B_s$

$$(1.13) \quad B_s + B_s^T + \beta A_s = 0, \quad s = 1, 2, 3.$$

There is considerable freedom in the choice of basis functions  $\phi_n$ . A collision term of the form (1.4) suggests the use of spherical harmonic basis functions (split appropriately into their real and imaginary parts) together with a suitable variable transformation, which transforms the equipotential surfaces  $\varepsilon = \text{const}$  into spheres. This basis has been successfully employed in [15], [16], and [4] and subsequent papers. In [12], [13] Hermite polynomials are used for a relaxation time approximation ( $S(k, k') = \text{const} M(k)$ ). Generally speaking, the advantage of these types of methods lies in the fact that they allow for a relatively easy coupling of kinetic to fluid dynamic equations by varying the number of terms used in the expansion (see [5], [6], [7], and references therein.)

However, in this paper we are not so much concerned with the Galerkin procedure itself but with the discretization of the first order hyperbolic system (1.7) in space and time. We will therefore not make any further use of the structure of the underlying

Boltzmann equation but only assume that the coefficients in the system (1.7) have the properties outlined above. In summary they are

$$\begin{aligned}
 (1.14) \quad & \text{(a) } A_s^T = A_s, \quad A_s(1,1) = 0, \quad s = 1, 2, 3, \\
 & \text{(b) } B_s + B_s^T = -\beta A_s, \quad B_s(1,n) = 0, \quad n = 1, \dots, N, \quad s = 1, 2, 3, \\
 & \text{(c) } C = C^T, \quad F^T C F \leq 0 \quad \forall F \in R^N, \quad C(1,n) = 0, \quad n = 1, \dots, N.
 \end{aligned}$$

**Boundary conditions.** For practical applications the boundary  $\partial\Omega$  of the simulation domain  $\Omega$  will consist of a part  $\partial\Omega_c$  representing Ohmic contacts and a part  $\partial\Omega_i$  representing insulating surfaces. The correct physical treatment of the boundary conditions for the Boltzmann equation (1.1) is not at all trivial, and it could be argued that a correct description is not possible at all at the level of the Boltzmann equation, but has to involve the wave nature of the electrons. In this paper we will ignore this controversy and instead simply mimic the boundary conditions used in Monte Carlo simulations. There, the general principle is that insulating surfaces simply reflect the electrons, and at Ohmic contacts electrons are injected according to a Maxwell distribution. The formulation of this principle in the case of general band structures involves the precise geometry of the surfaces of equal energy. For simplicity we will restrict ourselves to parabolic bands ( $\varepsilon = \frac{|k|^2}{2}, \nabla_k \varepsilon = k$ ) for the rest of this paper. In this case the boundary conditions are given by

$$\begin{aligned}
 (1.15) \quad & \text{(a) } f(x, k, t) = f(x, -k, t), \quad x \in \partial\Omega_i, \quad k \bullet r < 0, \\
 & \text{(b) } f(x, k, t) = \rho(x, t)M(k), \quad x \in \partial\Omega_c, \quad k \bullet r < 0, \\
 & \text{(c) } \partial\Omega = \partial\Omega_c \cup \partial\Omega_i
 \end{aligned}$$

holds, where  $r$  denotes the unit outward normal vector on the boundary  $\partial\Omega$ . In order to translate the conditions (1.15) into the framework of the Galerkin method, we will need to use a little bit of the theory of collocation methods to see that the coefficient matrices  $A_s$  in (1.7) are diagonalizable simultaneously. We assume the existence of a Gaussian integration rule of the form

$$(1.16) \quad \int_{R_k^3} dk \left[ \frac{1}{M} k_s \phi_\mu \phi_\nu \right] = \sum_{\alpha=1}^N w_\alpha k_s(\alpha) \phi_\mu(k(\alpha)) \phi_\nu(k(\alpha)), \quad \mu, \nu = 1, \dots, N$$

with nodal vectors  $k(\alpha) \in R^3$  and integrations weights  $w_\alpha$ . Such integration rules exist if the basis functions are products of the Maxwellian and polynomials or spherical harmonics. If we change the basis from the  $\{\phi_\mu\}$  to the collocation basis  $\{\psi_\mu\}$  according to

$$(1.17) \quad \psi_m(k) = \sum_{\nu=1}^N \phi_\nu(k) R(\nu, \mu), \quad \psi_\mu(k(\nu)) = \delta_{\mu\nu},$$

the resulting system matrices  $\tilde{A}_s = R^T A_s R$  become diagonal and  $\tilde{A}_s = \text{diag}\{k_s(1), \dots, k_s(N)\}$  holds. Thus, all of the matrices  $A_s$  are diagonalizable simultaneously and their eigenvalues  $k_s(\mu)$  correspond to discrete velocities. Moreover, since the Maxwellian  $M$  is symmetric in all variables the set of eigenvalues will be symmetric around the origin. Therefore, if  $k(\mu)$  is a Gaussian node, so is  $-k(\mu)$ . Imposing

conditions for  $k \bullet r < 0$  as in (1.15) corresponds to imposing conditions on the solution component in the subspace corresponding to the eigenvalues for which  $k(\mu) \bullet r < 0$  holds. In the Galerkin system (1.7) the flux through the boundary is given by the term  $\sum_{s=1}^3 r_s A_s F$ , and the matrices  $\sum_{s=1}^3 r_s A_s$  can be diagonalized by

$$(1.18) \quad \sum_{s=1}^3 r_s A_s = R \Lambda(x) R^T, \quad \Lambda(x) = \sum_{s=1}^3 r_s(x) \tilde{A}_s = \text{diag}(r \bullet k(1), \dots, r \bullet k(N)).$$

Because of the symmetry of the nodal vectors  $k(\mu)$  we can partition  $\Lambda$  into

$$(1.19) \quad \Lambda(x) = \begin{pmatrix} -\Lambda_1 & 0 & 0 \\ 0 & \Lambda_1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Lambda_1 > 0,$$

where the matrix  $\Lambda_1$  consists of the positive eigenvalues of  $\sum_{s=1}^3 r_s A_s$ . The components of the vector  $F$  corresponding to  $-\Lambda_1$  represent the inflow part of the solution, which has to be prescribed by the boundary conditions. We therefore define the projection matrix  $P(x)$  by

$$(1.20) \quad P(x) = R \begin{pmatrix} I_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} R^T \quad \text{for } x \in \partial\Omega_c,$$

$$P(x) = R \begin{pmatrix} I_1 & -I_1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} R^T \quad \text{for } x \in \partial\Omega_i,$$

with  $I_1$  and the identity matrix of the same dimension as  $\Lambda_1$ . The resulting boundary conditions are of the form

$$(1.21) \quad P(x)F(x, t) = 0 \quad \text{for } x \in \partial\Omega_i, \quad P(x)[F(x, t) - \rho e_1] = 0 \quad \text{for } x \in \partial\Omega_c,$$

where  $e_1$  denotes the first unit vector  $(1, 0, \dots, 0)^T$ .

It is important to note that the boundary conditions (1.21) are purely phenomenological and one should rather pay attention to the underlying principle that the inflow is given by a Maxwellian on  $\partial\Omega_c$  and by the outflow on  $\partial\Omega_i$ . This distinction becomes relevant in section 3, where we will employ operator splitting techniques for the time discretization of (1.7). Rather than worrying about a precise discretization of the boundary conditions (1.21) we will employ this principle in each partial step of the method.

On the surface, since (1.7) is a linear hyperbolic system, any standard hyperbolic method would seem to be appropriate. However, there are two features which make the choice of discretization nontrivial. Small values of the scaled mean free path  $\lambda$  in (1.7) correspond to the regime of the Hilbert expansion (c.f. [10]) which yields the parabolic drift-diffusion equation. Therefore the system (1.7) can be expected to behave like a parabolic rather than a hyperbolic system for small  $\lambda$ . Heuristically, this behavior can be explained in the following way: For small  $\lambda$  signals (waves) will propagate in all directions with wave speeds of order  $O(\frac{1}{\lambda})$ . Most of the signal will be damped by the matrix  $C$  on a time scale of order  $O(\frac{1}{\lambda^2})$ . However, the component of

the signal lying in the nontrivial null space of the matrix  $C$  will remain undamped. Thus, in the limit  $\lambda \rightarrow 0$  finite amplitude signals will propagate in all directions with infinite velocity, producing the parabolic behavior. Second, the electric field  $E$  can become extremely large locally putting an unrealistically small restriction on the Courant number of an explicit scheme. This problem of locally large fields has to be dealt with in a similar way as for the parabolic drift-diffusion equation.

This paper is organized as follows. In section 2 we discuss general aspects of the spatial discretization. For quite general meshes, we derive a discretization which retains the skew self-adjoint property of the free streaming operator, and therefore immediately guarantees the stability of the ordinary differential equation (ODE) system arising from the method of lines. In section 3 we consider the full space-time discretization of the system (1.7). It turns out that, in order to obtain a meaningful stability result for explicit time discretizations, which is not dependent on the size of the electric field  $E$ , and in order to capture the acceleration of waves properly, quite a bit more work has to be done. The chosen approach for the time discretization is an operator splitting scheme. At the end of section 3 we discuss the limiting behavior of the discrete system for small Knudsen numbers  $\lambda$ , and show that in the limit of the Hilbert expansion the system reduces to the well-known Scharfetter–Gummel discretization for the drift-diffusion equations. Section 4 presents some numerical experiments. The more technically involved proofs are collected in the appendix.

**2. The spatial discretization.** The topic of this section is a finite difference discretization of the operator

$$(2.1) \quad L[f] = \sum_{s=1}^3 (A_s \partial_{x_s} f + E_s B_s f)$$

in the spatial direction, which is appropriate for large electric fields  $E$ . The result of this section, summarized in Lemmas 2.1 and 2.2, will be a spatial discretization whose stability properties are independent of the size of the electric field  $E$ . Therefore the system of ODEs arising from the method of lines will be stable for any choice of mesh. Moreover, the resulting scheme reduces to the well-known Scharfetter–Gummel discretization for the drift-diffusion equation in the Hilbert expansion limit, that is, for  $\lambda \rightarrow 0$ .

We start by giving a brief review of the salient features of the stability analysis given in [11], [12], and [13] for the continuous Boltzmann equation and the series expansion system (1.7). The original convection operator  $\nabla_k \varepsilon \bullet \nabla_x + E \bullet \nabla_k$  is antisymmetric with respect to the usual  $L^2$  inner product. So

$$(2.2) \quad \int_{\Omega} dx \int_{R_k^3} dk \quad g[\nabla_k \varepsilon \bullet \nabla_x + E \bullet \nabla_k]f \\ = \int_{\partial\Omega} d\sigma [(\nabla_k \varepsilon \bullet r)fg] - \int_{\Omega} dx \int_{R_k^3} dk \quad f[\nabla_k \varepsilon \bullet \nabla_x + E \bullet \nabla_k]g$$

holds. This antisymmetry lies at the heart of all stability considerations, since it implies a purely imaginary spectrum. It has been lost in the Galerkin approximation due to the use of the weight function  $\exp(\beta\varepsilon)$ , which was needed to make the collision operator negative semidefinite. The antisymmetry can be restored by using a weighted scalar product in the x-direction as well. If we define for vector functions  $F$  and  $G$

the scalar product

$$(2.3) \quad \langle G, F \rangle = \int_{\Omega} dx \quad \kappa(x) G^T(x) F(x), \quad \kappa := e^{\beta V},$$

it is a straightforward exercise to show that, because of (1.14)(b), the discretized operator  $L$  in (2.1) is antisymmetric with respect to this scalar product, i.e., that

$$(2.4) \quad \int_{\Omega} dx \quad \kappa G^T \sum_{s=1}^3 [A_s \partial_{x_s} F + E_s B_s F] \\ = \int_{\partial\Omega} d\sigma \left[ \kappa \sum_{s=1}^3 r_s G^T A_s F \right] - \int_{\Omega} dx \quad \kappa \left[ \sum_{s=1}^3 A_s \partial_{x_s} G + E_s B_s G \right]^T F$$

holds. The use of this scalar product lies at the heart of the stability analysis in [12], [13] for the Galerkin system (1.7), and will be used for the derivation of the difference scheme as well. Given the relation (2.4), the most obvious thing to do would be to simply perform the variable transformation  $\tilde{F} = \sqrt{\kappa} F$  and discretize the transformed variable  $\tilde{F}$ . However, for applications it is crucial to preserve the charge conservation property

$$(2.5) \quad \partial_t \int_{\Omega} dx \quad F(x, t) = \text{boundary terms},$$

which holds because the first rows of the matrices  $B_s$  and  $C$  vanish, exactly. Just discretizing the transformed variable  $\tilde{F}$  would violate this conservation property.

We will derive the discretization on a rather general level, i.e., for general order and a general unstructured mesh. We assume a general unstructured mesh of the form

$$(2.6) \quad \mathbf{M} = \{\mathbf{x}_j, j = 1, 2, \dots, J, \mathbf{x}_j \in R^3, \mathbf{x}_j = (x_{j1}, x_{j2}, x_{j3})\}$$

and a mesh parameter  $h := \max_j \min_k |\mathbf{x}_j - \mathbf{x}_k|$  measuring the coarseness of the mesh. We assume that the integrals of a function  $f$  with respect to  $x$  will be approximated by

$$(2.7) \quad \int_{\Omega} dx \quad f \approx \sum_{j=1}^J \gamma_j f(\mathbf{x}_j)$$

with some integration weights  $\gamma_j$ . Integrals over the boundary  $\partial\Omega$  will be approximated by

$$(2.8) \quad \int_{\partial\Omega} d\sigma \quad f \approx \sum_{j=1}^J \omega_j f(\mathbf{x}_j)$$

with some boundary integration weights  $\omega_j$ , which are of course only nonzero if  $\mathbf{x}_j \in \partial\Omega$  holds. Furthermore we will assume some general discretization of the partial derivatives  $\partial_{x_s}$  of the form

$$(2.9) \quad \partial_{x_s} f(\mathbf{x}_j) \approx \sum_{k=1}^J a_s(j, k) f(\mathbf{x}_k).$$

For a difference method the coefficient  $a_s(j, k)$  will of course only be nonzero if the grid points  $\mathbf{x}_j$  and  $\mathbf{x}_k$  are in some sense neighbors. There are two properties which we require of the approximation of the partial derivative. The first is a conservation property which means that

$$(2.10) \quad \int_{\Omega} dx \quad \partial_{x_s} f \approx \sum_{j=1}^J \sum_{k=1}^J \gamma_j a_s(j, k) f(\mathbf{x}_k) = \sum_{j=1}^J \omega_j r_s(\mathbf{x}_j) f(\mathbf{x}_j) \quad \forall f \quad \Rightarrow$$

$$\sum_{j=1}^J \gamma_j a_s(j, k) = \omega_k r_s(\mathbf{x}_k) \quad \forall k, \quad s = 1, 2, 3$$

holds. The second is consistency of the discretization and its dual. Since the term

$$(2.11) \quad \sum_{j=1}^J \sum_{k=1}^J \gamma_j g(\mathbf{x}_j) a_s(j, k) f(\mathbf{x}_k)$$

approximates (via integration by parts)  $\int_{\partial\Omega} d\sigma (fg) - \int_{\Omega} dx (f \partial_{x_s} g)$  we write

$$(2.12) \quad \sum_{j=1}^J \sum_{k=1}^J \gamma_j g(\mathbf{x}_j) a_s(j, k) f(\mathbf{x}_k)$$

$$= \sum_{j=1}^J r_s(\mathbf{x}_j) \omega_j g(\mathbf{x}_j) f(\mathbf{x}_k) - \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \gamma_k f(\mathbf{x}_k) \tilde{a}_s(k, j) g(\mathbf{x}_j),$$

$$\tilde{a}_s(k, j) := r_s(\mathbf{x}_j) \frac{\omega_j}{\gamma_k} \delta_{jk} - \frac{\gamma_j}{\gamma_k} a_s(j, k),$$

and the coefficients  $\tilde{a}_s$  are the dual discretization of the partial derivative  $\partial_{x_s}$ . Here  $r = (r_1, r_2, r_3)$  denotes again the unit outward normal vector on the boundary  $\partial\Omega$ . For most difference methods  $a_s = \tilde{a}_s$  will hold with an appropriate choice of integration weights  $\gamma_j$ . We assume that the coefficients  $a_s$  as well as their dual  $\tilde{a}_s$  represent order  $p$  consistent discretizations of the partial derivative, so

$$(2.13) \quad \sum_{k=1}^{\infty} a_s(j, k) f(\mathbf{x}_k) = \partial_{x_s} f(\mathbf{x}_j) + O(h^p),$$

$$\sum_{k=1}^{\infty} \tilde{a}_s(j, k) f(\mathbf{x}_k) = \partial_{x_s} f(\mathbf{x}_j) + O(h^p), \quad s = 1, 2, 3$$

holds for sufficiently smooth functions  $f$ . Next we discretize the term  $E_s f$  by

$$(2.14) \quad E_s(\mathbf{x}_j, t) f(\mathbf{x}_j) \approx \sum_{k=1}^{\infty} b_s(j, k, t) f(\mathbf{x}_k).$$

Of course,  $b_s(j, k, t) = \delta_{jk} E_s(\mathbf{x}_j, t)$  would be an obvious choice. However, as will be seen, it pays to be a little more sophisticated in the choice of the  $b_s$ . However, regardless of the choice of the coefficients  $b_s$ , the discretized system

$$(2.15) \quad \lambda^2 \partial_t F(j, t) + \lambda \sum_{s=1}^3 \sum_{k=1}^J [a_s(j, k) A_s + b_s(j, k, t) B_s] F(k, t) = CF(j, t)$$

will now satisfy the discretized charge conservation property

$$(2.16) \quad \partial_t \sum_{j=1}^J \gamma_j F_1(j, t) = \text{boundary terms}$$

because of the conservative property (2.10) of the discretization coefficients  $a_s$ , and because the first rows of the matrices  $B_s$  and  $C$  vanish. The trick is now to choose the coefficients  $b_s$  in a way which preserves the discrete equivalent of (2.4). We introduce the corresponding discrete scalar product

$$(2.17) \quad \langle F, G \rangle_d := \sum_{j=1}^J \kappa_j \gamma_j F^T(j) G(j), \quad \kappa_j := \exp(\beta V(\mathbf{x}_j, t))$$

for vector valued grid functions  $F$  and  $G$ , and discretize the term  $E_s f$  in the form  $E_s f = \frac{1}{\beta} [\partial_{x_s} f - e^{-\beta V} \partial_{x_s} (e^{\beta V} f)]$ , using the coefficients  $a_s$  and their duals  $\tilde{a}_s$  for the discretization of the derivatives in this identity. This gives for the coefficients  $b_s$  the formula

$$(2.18) \quad b_s(j, k, t) = \frac{1}{\beta} [a_s(j, k) - \exp(\beta V(\mathbf{x}_k, t) - \beta V(\mathbf{x}_j, t)) \tilde{a}_s(j, k)],$$

with the dual coefficients  $\tilde{a}_s$  defined by (2.12). With this choice of  $b_s$  a straightforward calculation gives that the free streaming operator on the left hand side of (2.15) is skew self-adjoint with respect to the scalar product (2.17) and we can prove the following lemma.

LEMMA 2.1. *Let  $a_s(j, k)$  be the coefficients of the difference discretization of the partial derivative  $\partial_{x_s}$  on the mesh  $\mathbf{M}$  satisfying the consistency conditions (2.13). Let the coefficients  $b_s(j, k, t)$  be given by*

$$(2.19) \quad b_s(j, k, t) = \frac{1}{\beta} [a_s(j, k) - \exp(\beta V(\mathbf{x}_k, t) - \beta V(\mathbf{x}_j, t)) \tilde{a}_s(j, k)],$$

$$\tilde{a}_s(j, k) := r_s(\mathbf{x}_j) \frac{\omega_j}{\gamma_j} \delta_{jk} - \frac{\gamma_k}{\gamma_j} a_s(k, j), \quad j, k = 1, 2, \dots, s = 1, 2, 3.$$

Then the discrete operator  $L_d$  given by

$$(2.20) \quad L_d[F](j, t) = \sum_{s=1}^3 \sum_{k=1}^{\infty} [a_s(j, k) A_s + b_s(j, k, t) B_s] F(k, t)$$

is an order  $p$  consistent discretization of the operator  $L$  in (2.1). Moreover, for fixed time  $t$ , the operator  $L_d$  is antisymmetric with respect to the scalar product  $\langle \cdot, \cdot \rangle_d$  given by

$$(2.21) \quad \langle F, G \rangle_d := \sum_{j=1}^{\infty} \exp(\beta V(\mathbf{x}_j, t)) \gamma_j F^T(j) G(j),$$

so

$$(2.22) \quad \langle G, L_d[F] \rangle_d = \left[ \sum_{j=1}^J \omega_j \kappa_j \sum_{s=1}^3 r_s(\mathbf{x}_j) G(j)^T A_s F(j) \right] - \langle L_d[G], F \rangle_d$$

holds for all vector valued grid functions  $F$  and  $G$ .

As a consequence we obtain the stability of the infinite system of ODEs arising from the method of lines for the hyperbolic system (1.7). The significance of the following lemma lies in the fact that the stability is independent of the mean free path  $\lambda$  as well as of the electric field  $E$ . This has been achieved by the special construction of the spatial difference operator  $L_d$ .

LEMMA 2.2. *The system of ODEs given by*

(2.23)

$$(a) \quad \lambda^2 \partial_t F(j, t) + \lambda L_d[F](j, t) - CF(j, t) = \lambda^2 H(j, t), \quad \mathbf{x}_j \in \Omega - \partial\Omega,$$

$$(b) \quad (I - P_j) \{ \lambda^2 \partial_t F(j, t) + \lambda L_d[F](j, t) - CF(j, t) \} = (I - P_j) \lambda^2 H(j, t), \quad \mathbf{x}_j \in \partial\Omega,$$

$$(c) \quad P_j F(j, t) = 0, \quad \mathbf{x}_j \in \partial\Omega$$

satisfies

(2.24)

$$\partial_t \|F\|_d \leq \|F\|_d \frac{\beta}{2} \max\{|\partial_t V(\mathbf{x}_j, t)|, j = 1, 2, \dots\} + \|H\|_d, \quad \|\cdot\|_d := \sqrt{\langle \cdot, \cdot \rangle_d},$$

The proof of Lemma 2.2 is deferred to the appendix.

**3. Time discretization.** The topic of this section is the derivation of an appropriate time discretization for the hyperbolic system (1.7). Essentially, we have to deal with two problems:

1. *Stability.* Although the spatial discretization from the previous sections yields a stable ODE-system for the method of lines, this implies unconditional stability for the fully discretized system only if backward differences are used in time. Besides being computationally expensive, backward differencing is not appropriate for the discretization of wave propagation problems. If, on the other hand, the system (2.23) is discretized explicitly in time, adding an appropriate artificial diffusion term, the resulting CFL condition will again depend on the electric field  $E$ , yielding unacceptably small time steps.
2. *Dispersivity.* As will be seen, the system (1.7) is mildly dispersive in that the solution can be constructed as a superposition of modulated plane waves whose velocities depend on their spatial frequencies. This arises from the asymmetry of the coefficient matrices  $B_s$  and is a consequence of the fact that electrons (and therefore also waves) are accelerated in the electric field  $E$ .

The methodology in this section will consist of dealing with these problems separately in the context of an operator splitting approach. The result, summarized in (3.12), will be a scheme where each time step consists of three substeps simulating the three mechanisms present in the equations, namely wave propagation, wave acceleration due to the electric field, and diffusion, due to the collision term  $C$ .

As can be expected from a hyperbolic system, the method of lines theory from section 2 does by no means tell the whole story. For instance, one could conclude from Lemma 2.2 that a completely implicit discretization, such as backward Euler, would be stable. However, such a scheme would not reflect the hyperbolic nature of the problem and essentially eliminate all wave propagation in practice. As will be seen this hyperbolic structure can be quite complex due to the asymmetry of the coefficient matrices. To demonstrate this behavior, let's for the moment neglect the

collision matrix  $C$  in (1.7) and assume that the electric field  $E$  is constant in space and time. (So  $V(x) = -E \bullet x$  holds for the potential  $V$ .) We seek plane wave solutions of (1.7) of the form

$$(3.1) \quad F(x, t) = \exp \left( -\frac{\beta}{2} V(x) + i\omega t - i\xi \bullet x \right) z$$

for some constant vector  $z$ . Inserting (3.1) into (1.7) gives the eigenvalue problem

$$(3.2) \quad \omega z = \frac{1}{\lambda} \left[ \sum_{s=1}^3 A_s \xi_s - iE_s \left( B_s + \frac{\beta}{2} A_s \right) \right] z$$

for  $\omega(\xi)$  and  $z(\xi)$ . Since the matrices  $B_s + \frac{\beta}{2} A_s$  are skew symmetric and the matrices  $A_s$  are symmetric (see (1.14)) the complex matrix on the right hand side of (3.2) is hermitian, and so the eigenvalue  $\omega$  is real. Thus the solution can be constructed of undamped plane waves of the form (3.1). Moreover, the problem is dispersive, since the velocities  $\frac{|\omega(\xi)|}{|\xi|}$  depend on the frequency  $\xi$ . This is not surprising since the term  $E \bullet \nabla_k f$  is an acceleration term and therefore waves cannot be expected to travel with a fixed set of velocities. Particularly, low frequency signals can travel at high speeds for large values of the electric field  $E$ . Therefore any standard explicit discretization will suffer from a CFL condition much more restrictive than the  $\Delta t = O(\lambda \Delta x)$  one would expect. On the other hand, care has to be taken when using an implicit method since the original wave solution (3.1) is, in the absence of collisions, undamped. We will deal with this problem in a way which can be formulated best in the context of an operator splitting method (cf. [1], [2] for a reference). In this framework, to advance the solution from a time  $t_m$  to the next step  $t_{m+1} = t_m + \Delta t$  we solve the two problems

$$(3.3) \quad \begin{aligned} \text{(a)} \quad & \lambda \partial_t \tilde{G} + \sum_{s=1}^3 (A_s \partial_{x_s} \tilde{G} - \frac{\beta}{2} E_s(x, t_m) A_s \tilde{G}) = 0, \quad \tilde{G}(x, t_m) = F(x, t_m), \quad t_m \leq t \leq t_{m+1} \\ \text{(b)} \quad & \lambda \partial_t \tilde{H} + \sum_{s=1}^3 E_s(x, t_m) (B_s + \frac{\beta}{2} A_s) \tilde{H} = 0, \quad \tilde{H}(x, t_m) = \tilde{G}(x, t_{m+1}), \quad t_m \leq t \leq t_{m+1} \\ \text{(c)} \quad & F(x, t_{m+1}) = \tilde{H}(x, t_{m+1}). \end{aligned}$$

Performing now the variable transformation  $G(x, t) = \exp[\frac{\beta}{2} V(x, t_m)] \tilde{G}(x, t)$ ,  $H(x, t) = \exp[\frac{\beta}{2} V(x, t_m)] \tilde{H}(x, t)$  (3.3) becomes

$$(3.4) \quad \begin{aligned} \text{(a)} \quad & \lambda \partial_t G + \sum_{s=1}^3 A_s \partial_{x_s} G = 0, \quad G(x, t_m) = \exp \left[ \frac{\beta}{2} V(x, t_m) \right] F(x, t_m), \quad t_m \leq t \leq t_{m+1}, \\ \text{(b)} \quad & \lambda \partial_t H + \sum_{s=1}^3 E_s(x, t_m) \left( B_s + \frac{\beta}{2} A_s \right) H = 0, \quad H(x, t_m) = G(x, t_{m+1}), \quad t_m \leq t \leq t_{m+1}, \\ \text{(c)} \quad & F(x, t_{m+1}) = \exp \left[ -\frac{\beta}{2} V(x, t_m) \right] H(x, t_{m+1}). \end{aligned}$$

(3.4)(a) is now a constant coefficient hyperbolic problem, independent of the electric field  $E$  which can be solved by any kind of standard method. For constant field  $E$  this means that, after Fourier transformation in  $x$  we have replaced the matrix  $\exp[\Delta t \sum_{s=1}^3 (i\xi_s A_s + E_s(\frac{\beta}{2} A_s + B_s))]$  by the matrix product  $\exp[\Delta t \sum_{s=1}^3 i\xi_s A_s] \exp[\Delta t \sum_{s=1}^3 E_s(\frac{\beta}{2} A_s + B_s)]$ . This gives an  $O(\Delta t^2)$  error in each step, making the overall method formally first order in time, which becomes even smaller for small frequencies  $\xi$  for which the dispersive effect will be felt most.

The choice of discretization scheme for the ODE system (3.4)(b) is not so trivial. Since (3.4)(b) corresponds to the acceleration of waves and leaves amplitudes unchanged, one would like to minimize any artificial damping introduced by implicit methods. On the other hand, one would like to avoid a severe time step restriction for locally large fields. One relatively elegant and cheap way to satisfy both requirements is to apply operator splitting to (3.4)(b) and to solve the equations involved in the individual steps exactly. We split the ODE system (3.4)(b) according to

(3.5)

$$(a) \quad \lambda \partial_t H_s + E_s(x, t_m) \left( B_s + \frac{\beta}{2} A_s \right) H_s = 0, \quad s = 1, 2, 3, \quad t_m \leq t \leq t_{m+1},$$

$$(b) \quad H_1(x, t_m) = G(x, t_{m+1}), \quad H_s(x, t_m) = H_{s-1}(x, t_{m+1}), \quad s = 2, 3.$$

Each of the steps in (3.5) can now be carried out exactly without any time step restriction at relatively little cost. Since the matrices  $B_s + \frac{\beta}{2} A_s$  are skew symmetric there exist complex hermitian matrices  $Z_s$  and purely imaginary diagonal matrices  $\Gamma_s$  such that

$$(3.6) \quad B_s + \frac{\beta}{2} A_s = Z_s \Gamma_s Z_s^H$$

holds, where  $Z_s^H$  denotes complex conjugate transpose of  $Z_s$ . These matrices can be computed once at the beginning of the computation and stored. The solution of (3.5) is then given by

$$(3.7) \quad H_s(x, t_{m+1}) = Z_s \exp \left[ -\frac{\Delta t}{\lambda} E_s(x, t_m) \Gamma_s \right] Z_s^H H_s(x, t_m).$$

The step (3.5) is then computed as

(3.8)

$$F(x, t_{m+1}) = \exp \left[ -\frac{\beta}{2} V(x, t_m) \right] \prod_{s=1}^3 \left\{ Z_s \exp \left[ -\frac{\Delta t}{\lambda} E_s(x, t_m) \Gamma_s \right] Z_s^H \right\} G_s(x, t_{m+1}).$$

In combining the operator splitting approach with the spatial discretization from section 2 we encounter one major problem: For practical applications it is absolutely essential to discretize the conservation of charge property exactly. So, we want the relation

$$(3.9) \quad \sum_{j=1}^J \gamma_j F_1(j, t_{m+1}) = \sum_{j=1}^J \gamma_j F_1(j, t_m) + \text{boundary terms}$$

to hold exactly. Experimentally, only the fluxes through the boundary can be observed, and it is important that the change in time of the total charge is preserved exactly. Unfortunately, the hyperbolic scheme employed for (3.4)(a) will preserve  $G_1 = \exp[\frac{\beta}{2}V]F_1$  rather than the density  $F_1$ , which, for not so small values of the time step and rapidly varying potentials can make quite a difference. We circumvent this problem by applying the operator splitting technique once more. We partition the matrices  $A_s$  and  $B_s$  according to

$$(3.10) \quad \begin{aligned} \text{(a)} \quad & A_s = \tilde{A}_s + \hat{A}_s, \quad B_s = \tilde{B}_s + \hat{B}_s, \quad s = 1, 2, 3, \\ \text{(b)} \quad & \tilde{A}_s = \begin{pmatrix} 0 & A_s^{12} \\ A_s^{21} & 0 \end{pmatrix}, \quad \hat{A}_s = \begin{pmatrix} 0 & 0 \\ 0 & A_s^{22} \end{pmatrix}, \\ \text{(c)} \quad & \tilde{B}_s = \begin{pmatrix} 0 & 0 \\ B_s^{21} & 0 \end{pmatrix}, \quad \hat{B}_s = \begin{pmatrix} 0 & 0 \\ 0 & B_s^{22} \end{pmatrix}, \end{aligned}$$

where, as before, the first row of  $A_s$  is given by  $(0, A_s^{12})$  and so on. Note that the matrices  $\tilde{A}_s, \tilde{B}_s, \hat{A}_s, \hat{B}_s$  satisfy the same properties as the original matrices  $A_s$  and  $B_s$  in (1.14) and  $B_s^{21} = -\beta A_s^{21}$  holds. We now apply the operator splitting method once more by successively solving

$$(3.11) \quad \text{(a)} \quad \lambda^2 \partial_t F + \lambda \sum_{s=1}^3 [\hat{A}_s \partial_{x_s} F + E_s \hat{B}_s F] = 0,$$

and

$$(b) \quad \lambda^2 \partial_t F + \lambda \sum_{s=1}^3 [\tilde{A}_s \partial_{x_s} F + E_s \tilde{B}_s F] = CF$$

to advance the solution one time step, where we employ hyperbolic method (3.4) to (3.11)(a) and discretize (3.11)(b) implicitly by the backward Euler scheme together with the spatial discretization from section 2. This represents somewhat of a compromise with the hyperbolic nature of the problem. However, step (3.11)(b) will be dominated by the damping due to the matrix  $C$  anyway, and therefore using the implicit method there is not such a tragedy. As mentioned in section 1, we will use the basic idea behind the boundary conditions (1.21), namely that the inflow is either given by a Maxwellian or by the outflow, at each step of the splitting method. This gives the following method: given the solution vector  $F(j, t_m)$  solve

$$(3.12) \quad \begin{aligned} \text{(a)} \quad & \lambda \partial_t G + \sum_{s=1}^3 \hat{A}_s \partial_{x_s} G = 0, \quad t_m \leq t \leq t_{m+1}, \quad G(x, t_m) = \exp \left[ \frac{\beta}{2} V(x, t_m) \right] F(x, t_m) \\ \text{(b)} \quad & \hat{P}(x) [G(x, t) - \exp \left[ \frac{\beta}{2} V(x, t_m) \right] \rho(x) e_1] = 0, \quad x \in \partial\Omega, \\ \text{(c)} \quad & K(j, t_m) = \exp \left[ -\frac{\beta}{2} V(x, t_m) \right] \prod_{s=1}^3 \left\{ \hat{Z}_s \exp \left[ -\frac{\Delta t}{\lambda} E_s(\mathbf{x}_j, t_m) \hat{\Gamma}_s \right] \hat{Z}_s^H \right\} G(\mathbf{x}_j, t_{m+1}), \\ & \hat{B}_s + \frac{\beta}{2} \hat{A}_s = \hat{Z}_s \hat{\Gamma}_s \hat{Z}_s^H, \quad s = 1, 2, 3, \end{aligned}$$

$$(d) \quad \lambda^2 K(j, t_{m+1}) + \lambda \Delta t \tilde{L}_d K(j, t_{m+1}) - \Delta t C K(j, t_{m+1}) = \lambda^2 K(j, t_m),$$

$$\mathbf{x}_j \in \Omega - \partial\Omega,$$

$$(e) \quad (I - \tilde{P}(\mathbf{x}_j))\lambda^2 K(j, t_{m+1}) + (I - \tilde{P}(\mathbf{x}_j))\{\lambda \Delta t \tilde{L}_d K(j, t_{m+1}) - \Delta t C K(j, t_{m+1})\} \\ = (I - \tilde{P}(\mathbf{x}_j))\lambda^2 K(j, t_m), \quad \mathbf{x}_j \in \partial\Omega,$$

$$(f) \quad \tilde{P}(\mathbf{x}_j)K(j, t_{m+1}) = \tilde{P}(\mathbf{x}_j)e_1 \rho(\mathbf{x}_j), \quad \mathbf{x}_j \in \partial\Omega,$$

$$(g) \quad \tilde{L}_d K(j, t_{m+1}) := \sum_{s=1}^3 \sum_{k=1}^J [a_s(j, k)\tilde{A}_s + b_s(j, k, t_{m+1})\tilde{B}_s]K(k, t_{m+1}),$$

$$(h) \quad F(j, t_{m+1}) = K(j, t_{m+1}),$$

where, for notational simplicity, we have set  $\rho(x) = 0, \quad x \in \partial\Omega_i$ . The projection matrices  $\hat{P}$  and  $\tilde{P}$  are defined according to (1.20) for the matrices  $\hat{A}_s$  and  $\tilde{A}_s$ . So

$$(3.13) \quad (a) \quad \hat{P}(x) = \hat{R}(x) \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \hat{R}(x)^T, \quad x \in \partial\Omega_c,$$

$$(b) \quad \hat{P}(x) = \hat{R}(x) \begin{pmatrix} I & -I & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \hat{R}(x)^T, \quad x \in \partial\Omega_i,$$

$$(c) \quad \sum_{s=1}^3 r_s \hat{A}_s = \hat{R} \begin{pmatrix} -\Lambda_1 & 0 & 0 \\ 0 & \Lambda_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \hat{R}^T, \quad \Lambda_1 > 0,$$

$$(d) \quad \tilde{P}(x) = \tilde{R}(x) \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tilde{R}(x)^T, \quad x \in \partial\Omega_c,$$

$$(e) \quad \tilde{P}(x) = \tilde{R}(x) \begin{pmatrix} I & -I & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tilde{R}(x)^T, \quad x \in \partial\Omega_i,$$

$$(f) \quad \sum_{s=1}^3 r_s \tilde{A}_s = \tilde{R} \begin{pmatrix} -\Lambda_1 & 0 & 0 \\ 0 & \Lambda_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tilde{R}^T, \quad \Lambda_1 > 0$$

holds. For the operator splitting scheme (3.12) we now have the following theorem which gives stability with a moderate  $O(\lambda)$  CFL condition.

**THEOREM 3.1.** *Let the time step be such that the discretization of the constant coefficient hyperbolic boundary value problem (3.12)(a) is stable in the discrete  $L^2$  norm, so*

$$(3.14) \quad \sum_{j=1}^J \gamma_j |G(\mathbf{x}_j, t_{m+1})|^2 \leq \sum_{j=1}^J \gamma_j |G(\mathbf{x}_j, t_m)|^2$$

holds. Then, for the boundary influx  $\rho$  in (3.12) equal to zero the stability estimate

$$(3.15) \quad \begin{aligned} \text{(a)} \quad & \|F\|_d(t_{m+1}) \leq \max_j \sqrt{\frac{\kappa_j(t_{m+1})}{\kappa_j(t_m)}} \|F\|_d(t_m), \\ \text{(b)} \quad & \|F\|_d^2(t) := \sum_{j=1}^J \gamma_j \kappa_j(t) |F(j, t)|^2 \end{aligned}$$

holds.

The proof of Theorem 3.1 is deferred to the appendix.

We conclude this section by considering the diffusion limit of the scheme (3.12). In the limit  $\lambda \rightarrow 0$  the Boltzmann equation is, according to the Hilbert expansion, replaced by a diffusion equation, and as we will see, the scheme (3.12) reduces to a difference scheme for this diffusion equation employing backward differences in time and a generalization of the well-known Scharfetter–Gummel scheme [14] in the spatial direction. To this end, we rescale all but the first component of the the solution vector  $F$  by  $F = (F_1, F_2) = (F_1, \lambda \tilde{F}_2)$ , where  $F_1$  denotes the first component of  $F$  and  $F_2$  denotes the rest. In the same way we rescale the intermediate solutions  $G$  and  $K$  in (3.12) by  $F = (G_1, G_2) = (G_1, \lambda \tilde{G}_2)$  and  $K = (K_1, K_2) = (K_1, \lambda \tilde{K}_2)$ . Since the first row and column of the matrices  $\tilde{A}_s$  and  $\tilde{B}_s$  are identically zero the diagonalization matrices  $\tilde{\Gamma}_s$  and  $\tilde{Z}_s$  in (3.12)(c) will be of the form

$$(3.16) \quad \tilde{Z}_s = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{Z}^{22} \end{pmatrix}, \quad \Gamma_s = \begin{pmatrix} 0 & 0 \\ 0 & \Gamma^{22} \end{pmatrix},$$

which implies  $G_1(x, t_{m+1}) = G_1(x, t_m) = \exp[\frac{\beta}{2} V(x, t_m)] F_1(x, t_m)$  and, consequently,  $K_1(x, t_m) = F_1(x, t_m)$ . Splitting (3.12)(d) into the first and second component gives, after rescaling,

$$(3.17) \quad \begin{aligned} \text{(a)} \quad & K_1(j, t_{m+1}) + \Delta t \tilde{L}_d^{12} \tilde{K}_2(j, t_{m+1}) = K_1(j, t_m), \quad \mathbf{x}_j, \\ \text{(b)} \quad & \lambda^2 \tilde{K}_2(j, t_{m+1}) + \Delta t \tilde{L}_d^{21} K_1(j, t_{m+1}) - \Delta t C^{22} \tilde{K}_2(j, t_{m+1}) = \lambda^2 \tilde{K}_2(j, t_m), \quad \mathbf{x}_j, \end{aligned}$$

where the  $\tilde{L}_d^{rs}$  denote the corresponding blocks of the (matrix-) operator  $\tilde{L}_d$ . ( $\tilde{L}_d^{11} = \tilde{L}_d^{22} = 0$  holds because of the structure of the matrices  $\tilde{A}$  and  $\tilde{B}$ .) If we now formally set  $\lambda = 0$  in (3.17) and explicitly write the operator  $\tilde{L}_d$  we obtain

$$(3.18) \quad \begin{aligned} \text{(a)} \quad & K_1(j, t_{m+1}) + \Delta t \sum_{s=1}^3 \sum_k A_s^{12} a_s(j, k) \tilde{K}_2(k, t_{m+1}) = K_1(j, t_m), \quad \mathbf{x}_j, \\ \text{(b)} \quad & \sum_{s=1}^3 \sum_k [A_s^{21} a_s(j, k) + B_s^{21} b_s(j, k)] K_1(k, t_{m+1}) = C^{22} \tilde{K}_2(j, t_{m+1}). \end{aligned}$$

Of course, letting  $\lambda$  go to zero will result in an instability in the hyperbolic step (3.12)(a). However, this is irrelevant since the result of this step is never used. Using the fact that  $B_s^{21} = -A_s^{21}$  holds, (3.18)(b) becomes

$$(3.19) \quad \tilde{K}_2(j, t_{m+1}) = \exp[-V(x_j)] \sum_{s=1}^3 \sum_k (C^{22})^{-1} A_s^{21} \tilde{a}_s(j, k) \exp[V(x_k)] K_1(k, t_{m+1}).$$

If we now rename the variables according to  $\rho = K_1$ ,  $J_s = A_s^{12}K_2$  the difference scheme (3.18) becomes

$$(3.20) \quad \begin{aligned} \text{(a)} \quad & \rho(j, t_{m+1}) + \Delta t \sum_{s=1}^3 \sum_k a_s(j, k) J_s(k, t_{m+1}) = \rho(j, t_m), \quad \mathbf{x}_j, \\ \text{(b)} \quad & J_s(j, t_{m+1}) = -\exp[-\mathbb{V}(x_j)] \sum_{s=1}^3 \sum_k D_s \tilde{a}_s(j, k) \exp[\mathbb{V}(x_k)] K_1(k, t_{m+1}), \end{aligned}$$

with the positive definite diffusion matrices  $D_s$  given by  $D_s = -A_s^{12}(C^{22})^{-1}A_s^{21}$ . (3.20) is now a Scharfetter–Gummel type discretization of the corresponding diffusion equation, since the current relation (3.20)(b) is discretized in a self-adjoint form.

So, in conclusion, the time discretization given by (3.12) takes into account the three essential mechanisms present in the Boltzmann equation, namely convection, acceleration by the force  $E$ , and diffusion due to the collision term. If waves with possible speeds of order  $O(\frac{1}{\lambda})$  shall be resolved a necessary CFL condition of the form  $\Delta t = O(\lambda \Delta x)$  has to be observed. However, in the limit of the Hilbert expansion it reproduces the appropriate discretization of the resulting diffusion equation.

**4. Numerical experiments.** In this section we present some numerical experiments for the discretization derived in sections 2 and 3. The experiments are carried out for the case of one spatial dimension. So, the system

$$(4.1) \quad \lambda^2 \partial_t F + \lambda(A_1 \partial_{x_1} F + E_1 B_1 F) = CF$$

is solved on the interval  $[0, 1]$  together with Dirichlet boundary conditions at both end points. Thus, the boundary  $\partial\Omega_c$  consists of the two end points  $x = 0$  and  $x = 1$  and the boundary conditions (1.20) are imposed there. As a collision term the simple relaxation time approximation is chosen. So  $S(k, k') = M(k)$  in (1.3) holds. This choice of collision operator particularly simplifies the Galerkin procedure in the case of parabolic band structures since in this case the Boltzmann equations allow for solutions of the form

$$(4.2) \quad f(x, k, t) = f(x_1, k_1, t) \exp[-\beta(k_2^2 + k_3^2)],$$

provided the initial conditions are of this form. Accordingly we choose the basis functions  $\phi_n$  as

$$(4.3) \quad \phi_n(k) = \exp[-\beta|k|^2] H_n(k_1),$$

where the  $H_n(k_1)$  are the corresponding orthonormalized polynomials, namely the Hermite polynomials. The collision matrix  $C$  in (1.7) is then of the form

$$(4.4) \quad C = \begin{pmatrix} 0 & 0 \\ 0 & -I \end{pmatrix},$$

where  $I$  denotes the  $n - 1$  dimensional identity matrix. It could be argued that, with so many simplifications, the resulting system is of little physical relevance, but the point of this paper is to investigate the space-time discretization of the system (1.7)

and its salient features will remain the same for more realistic collision models. All the following computational results are obtained by using six terms in the expansion, after verifying that using more terms produces practically identical pictures. For the wave propagation step (3.12)(a)–(b) the Lax–Friedrichs scheme was used. The coefficients  $a_1(j, k)$  for the spatial discretization of the diffusion step (3.12)(e)–(g) were chosen to correspond to central differences in the interior and one sided differences at the boundaries. So,

$$(4.5) \quad a_1(j, k) = \begin{cases} \frac{\delta_{j+1,k} - \delta_{j-1,k}}{x_{j+1} - x_{j-1}} & j = 1, \dots, J - 1, \\ \frac{\delta_{j+1,k} - \delta_{j,k}}{x_{j+1} - x_j} & j = 0, \\ \frac{\delta_{j,k} - \delta_{j-1,k}}{x_j - x_{j-1}} & j = J \end{cases}$$

holds for a mesh  $0 = x_0 < \dots < x_J = 1$ . Integration weights  $\gamma_j$  and  $\omega_j$  for integrals over the interior and the boundary are chosen as

$$(4.6) \quad (a) \quad \gamma_j = \begin{cases} \frac{x_{j+1} - x_{j-1}}{2} & j = 1, \dots, J - 1, \\ \frac{x_{j+1} - x_j}{2} & j = 0, \\ \frac{x_j - x_{j-1}}{2} & j = J, \end{cases}$$

$$(b) \quad \omega_j = \delta_{j,0} + \delta_{j,J}.$$

Direct calculation shows that the adjoint coefficients  $\tilde{a}_1$  in (2.12) coincide with the  $a_1$  for this case, and the term  $A\partial_{x_1}F + EB_1F$  is discretized at  $x_j$  as

$$(4.7) \quad \sum_{k=1}^J a_1(j, k) \left[ A_1 + \frac{1 - \exp(\beta V_k - \beta V_j)}{\beta} B_1 \right] F(x_k, t).$$

The scaled Knudsen number  $\lambda$  was chosen as 0.2, which means that an electron undergoes on average five collision events travelling from  $x = 0$  to  $x = 1$ . We simulate the situation of an electron travelling from right to left through a potential barrier located at  $x = 0.5$ . The corresponding electric field  $E = -\nabla_x V$  is shown in Figure 1. Figure 2 demonstrates the effect of the deceleration of the electron at the barrier. In order to study this effect, we have omitted the diffusion step (3.12)(d)–(g) and only carried out the wave propagation and acceleration parts (3.12)(a)–(c). Figure 2 shows the second component of the solution vector  $F$  which is partially reflected at the barrier due to the deceleration effect. (Omitting the diffusion step, the first component of  $F$ , the electron density, remains of course unchanged in time.) Figures 3–8 show the repetition of the above experiment, now carrying out all three steps in (3.12). As a comparison we also show the solution of the drift-diffusion problem, drawn in dashed lines. The drift-diffusion solution has been computed by rescaling the vector  $F$  according to section 4 and setting  $\lambda = 0$ . In this case, only the diffusion step (3.12)(d)–(g) is carried out. Figure 3 shows the evolution of the electron density  $\rho = F_1$ . Figures 4–8 show the time evolution of the current density  $J = (A_1^2)F$ . While there is quite good agreement for the electron densities between the drift diffusion solution and the solution of the Boltzmann equation, the current densities differ dramatically. We are generally interested in two quantities, namely the size of the currents and the time it takes for the system to reach a steady state. In one spatial dimension a steady state

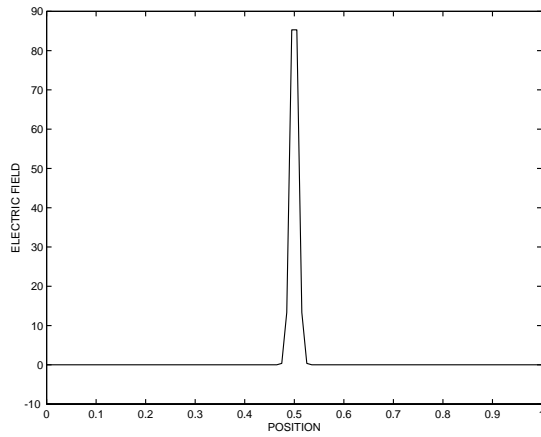


FIG. 1.

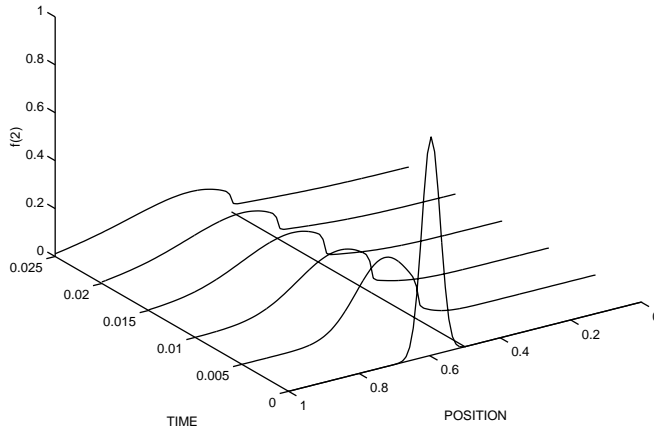


FIG. 2.

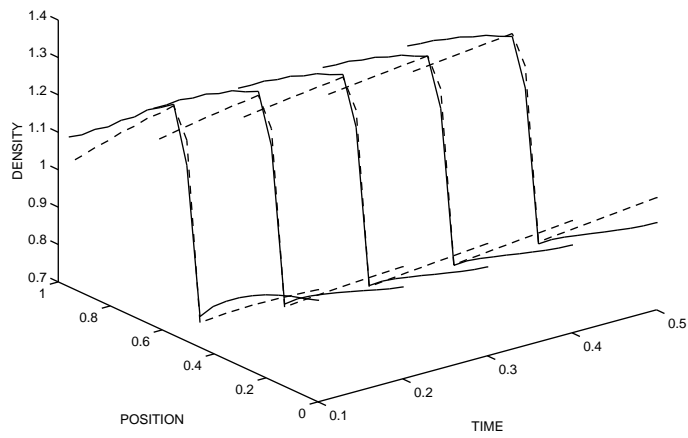


FIG. 3.

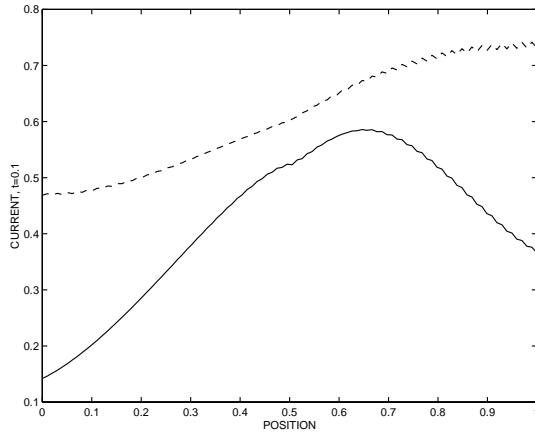


FIG. 4.

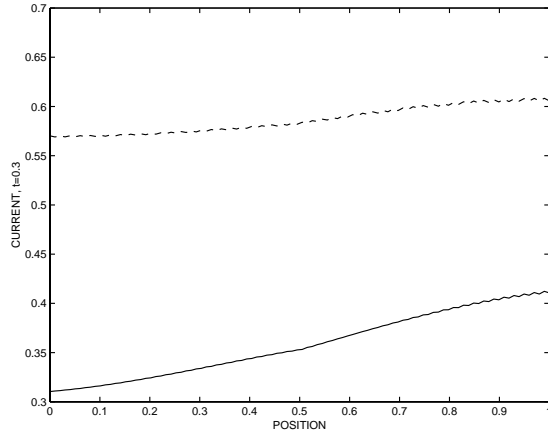


FIG. 5.

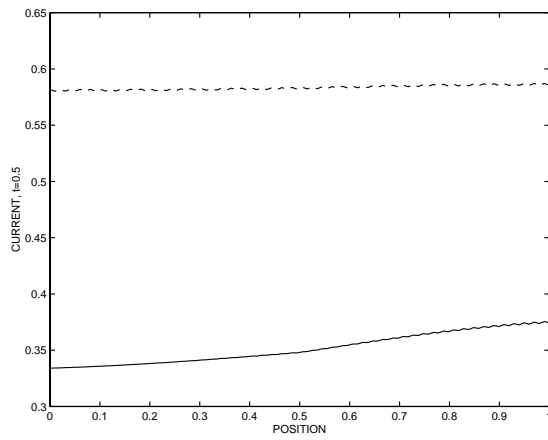


FIG. 6.

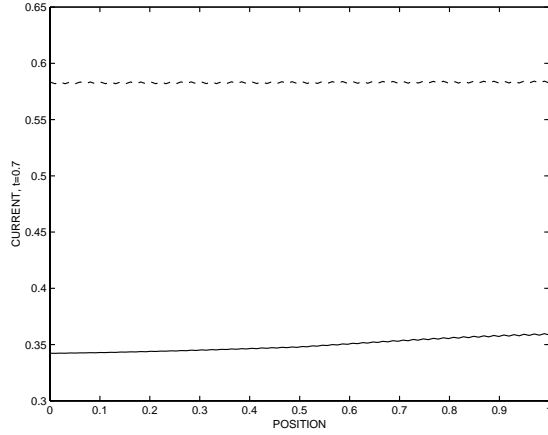


FIG. 7.

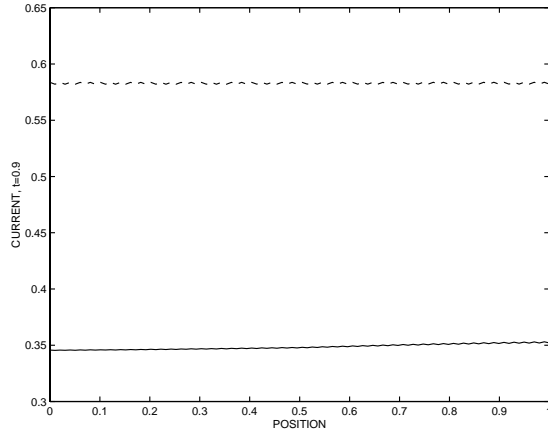


FIG. 8.

is characterized by the current density  $J$  being constant in space. Figures 4–8 show that the Boltzmann solution requires much longer to reach a spatially constant current, and that the value of this current is drastically different from the drift-diffusion solution. This implies that, at least for this simple model problem, the drift diffusion regime is not applicable for geometry dimensions of the order of five times the mean free path.

**Appendix.**

*Proof of Lemma 2.2.* Computing the time derivative of the term  $\langle F, F \rangle_d$  we obtain

$$\begin{aligned}
 \text{(A.1)} \quad \partial_t \langle F, F \rangle_d &= 2 \langle F, \partial_t F \rangle_d + \sum_{j=1}^{\infty} \beta \partial_t V(\mathbf{x}_j, t) \exp(\beta V(\mathbf{x}_j, t)) \gamma_j |F(j, t)|^2 \\
 &\leq 2 \langle F, \partial_t F \rangle_d + \beta \max\{|\partial_t V(\mathbf{x}_j, t)|, j = 1, 2, \dots\} \langle F, F \rangle_d.
 \end{aligned}$$

For the term  $\langle F, \partial_t F \rangle_d$  in (A.1) we obtain

$$(A.2) \quad \begin{aligned} \langle F, \partial_t F \rangle_d &= \sum_{\mathbf{x}_j \in \Omega - \partial\Omega} \gamma_j \kappa_j F(j, t)^T \partial_t F(j, t) \\ &+ \sum_{\mathbf{x}_j \in \partial\Omega} \gamma_j \kappa_j F(j, t)^T (I - P_j) \partial_t F(j, t) + \sum_{\mathbf{x}_j \in \partial\Omega} \gamma_j \kappa_j F(j, t)^T P_j \partial_t F(j, t). \end{aligned}$$

The third term in (A.2) vanishes because of the homogeneous boundary conditions (2.23)(c). For the second term in (A.2) we obtain

$$(A.3) \quad \begin{aligned} &\sum_{\mathbf{x}_j \in \partial\Omega} \gamma_j \kappa_j F(j, t)^T (I - P_j) \partial_t F(j, t) \\ &= \lambda^{-2} \sum_{\mathbf{x}_j \in \partial\Omega} \gamma_j \kappa_j F(j, t)^T (I - P_j) \{-\lambda L_d[F](j, t) + CF(j, t) + \lambda^2 H(j, t)\} \\ &= \lambda^{-2} \sum_{\mathbf{x}_j \in \partial\Omega} \gamma_j \kappa_j F(j, t)^T \{-\lambda L_d[F](j, t) + CF(j, t) + \lambda^2 H(j, t)\}, \end{aligned}$$

again because of the homogeneous boundary condition (2.23)(c). For the first term in (A.2) we get

$$(A.4) \quad \begin{aligned} &\sum_{\mathbf{x}_j \in \Omega - \partial\Omega} \gamma_j \kappa_j F(j, t)^T \partial_t F(j, t) \\ &= \lambda^{-2} \sum_{\mathbf{x}_j \in \Omega - \partial\Omega} \gamma_j \kappa_j F(j, t)^T \{-\lambda L_d[F](j, t) + CF(j, t) + \lambda^2 H(j, t)\}. \end{aligned}$$

Adding (A.3) and (A.4) gives

$$(A.5) \quad \langle F, \partial_t F \rangle_d = \lambda^{-2} \langle F, -\lambda L_d[F] + CF \rangle_d + \langle F, H \rangle_d.$$

Now  $\langle F, CF \rangle_d$  is nonpositive and  $L_d$  is antisymmetric. This gives (because of (2.22))

$$(A.6) \quad \begin{aligned} \langle F, \partial_t F \rangle_d &\leq -\frac{1}{2\lambda} \left[ \sum_{j=1}^J \omega_j \kappa_j \sum_{k=1}^3 r_s(\mathbf{x}_j) F(j)^T A_s F(j) \right] + \langle F, H \rangle_d \\ &= -\frac{1}{2\lambda} \left[ \sum_{j=1}^J \omega_j \kappa_j \sum_{s=1}^3 r_s(\mathbf{x}_j) F(j)^T (I - P_j) A_s (I - P_j) F(j) \right] + \langle F, H \rangle_d. \end{aligned}$$

Since the matrices  $\sum_{k=1}^3 r_s(\mathbf{x}_j)(I - P_j)A_s(I - P_j)$  are positive semidefinite, inserting (A.6) into (A.1) gives

$$(A.7) \quad \partial_t \langle F, F \rangle_d \leq 2 \langle F, H \rangle_d + \beta \max\{|\partial_t V(\mathbf{x}_j, t)|, j = 1, 2, \dots\} \langle F, F \rangle_d.$$

Standard application of the Gronwall inequality yields the result (2.24).  $\square$

*Proof of Theorem 3.1.* We start with the stability estimate for (3.12)(c): Multiplying (3.12)(c) by  $\gamma_j \kappa_j(t_{m+1})K(j, t_{m+1})$  and summation over the  $j$  gives

$$(A.8) \quad \lambda^2 \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_{m+1})|^2$$

$$\begin{aligned}
& +\Delta t \sum_{bf x_j \notin \partial\Omega} \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T [\lambda(L_d K)(j, t_{m+1}) - CK(j, t_{m+1})] \\
& +\Delta t \sum_{bf x_j \in \partial\Omega} \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T [I - \tilde{P}(\mathbf{x}_j)] [\lambda(L_d K)(j, t_{m+1}) - CK(j, t_{m+1})] \\
& = \lambda^2 \sum_{bf x_j \notin \partial\Omega} \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T K(j, t_m) \\
& + \lambda^2 \sum_{bf x_j \in \partial\Omega} \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T [I - \tilde{P}(\mathbf{x}_j)] K(j, t_m).
\end{aligned}$$

Since the matrices  $\tilde{P}$  are projections and symmetric, multiplying (3.12)c by  $\tilde{P}(\mathbf{x}_j)$  gives immediately that  $K(j, t_{m+1})^T \tilde{P}(\mathbf{x}_j) = 0$  holds. Therefore (A.8) becomes

$$\begin{aligned}
\text{(A.9)} \quad & \lambda^2 \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_{m+1})|^2 \\
& + \Delta t \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T [\lambda(L_d K)(j, t_{m+1}) - CK(j, t_{m+1})] \\
& = \lambda^2 \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T K(j, t_m).
\end{aligned}$$

Because of (2.21)

$$\begin{aligned}
\text{(A.10)} \quad & \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T (L_d K)(j, t_{m+1}) \\
& = \sum_{j=1}^J \sum_{s=1}^3 \omega_j \kappa_j(t_{m+1}) r_s K(j, t_{m+1})^T \tilde{A}_s K(j, t_{m+1}) \\
& = \sum_{j=1}^J \sum_{s=1}^3 \omega_j \kappa_j(t_{m+1}) r_s K(j, t_{m+1})^T [I - \tilde{P}(\mathbf{x}_j)]^T \tilde{A}_s [I - \tilde{P}(\mathbf{x}_j)] K(j, t_{m+1}) \geq 0
\end{aligned}$$

holds since, by construction, the matrices  $\sum_{s=1}^3 [I - \tilde{P}(\mathbf{x}_j)]^T r_s \tilde{A}_s [I - \tilde{P}(\mathbf{x}_j)]$  are positive semidefinite. Using (A.10) and the fact that the matrix  $C$  is negative semidefinite, we obtain

$$(A.11) \quad \lambda^2 \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_{m+1})|^2 \leq \lambda^2 \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) K(j, t_{m+1})^T K(j, t_m),$$

and, using the Cauchy–Schwartz inequality

$$(A.12) \quad \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_{m+1})|^2 \leq \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_m)|^2.$$

Therefore

$$(A.13) \quad \|F\|_d(t_{m+1}) \leq \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_m)|^2$$

holds. Now, since the hyperbolic scheme in (3.12)(a) is assumed to be stable and the matrices  $\hat{B}_s + \frac{\beta}{2} \hat{A}_s$  are skew symmetric we have

$$(A.14) \quad \begin{aligned} \sum_{j=1}^J \gamma_j \kappa_j(t_{m+1}) |K(j, t_m)|^2 &= \sum_{j=1}^J \gamma_j \frac{\kappa_j(t_{m+1})}{\kappa_j(t_m)} |G(x_j, t_{m+1})|^2 \\ &\leq \max_j \left\{ \frac{\kappa_j(t_{m+1})}{\kappa_j(t_m)} \right\} \sum_{j=1}^J \gamma_j |G(x_j, t_{m+1})|^2 \leq \max_j \left\{ \frac{\kappa_j(t_{m+1})}{\kappa_j(t_m)} \right\} \sum_{j=1}^J \gamma_j |G(x_j, t_m)|^2 \\ &= \max_j \left\{ \frac{\kappa_j(t_{m+1})}{\kappa_j(t_m)} \right\} \sum_{j=1}^J \gamma_j \kappa_j(t_m) |F(x_j, t_m)|^2. \quad \square \end{aligned}$$

#### REFERENCES

- [1] A. ARNOLD AND C. RINGHOFER, *Operator splitting methods applied to spectral discretizations of quantum transport equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1876–1894.
- [2] A. ARNOLD AND C. RINGHOFER, *An operator splitting method for the Wigner–Poisson problem*, SIAM J. Numer. Anal., 33 (1996), pp. 1622–1643.
- [3] N. ASHCROFT AND M. MERMIN, *Solid State Physics*, Holt-Saunders, New York, 1976.
- [4] N. GOLDSMAN, L. HENRICKSON, AND J. FREY, *A physics based analytical-numerical solution to the Boltzmann equation for use in semiconductor device simulation*, Solid State Electr., 34 (1991), p. 389.
- [5] A. KLAR, *A numerical method for kinetic semiconductor equations in the drift diffusion limit*, SIAM J. Sci. Comput., 20 (1999), pp. 1696–1712.
- [6] A. KLAR, *An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit*, SIAM J. Numer. Anal., 35 (1998), pp. 1073–1094.
- [7] A. KLAR, *Asymptotic-induced domain decomposition methods for kinetic and drift diffusion semiconductor equations*, SIAM J. Sci. Comput., 19 (1998), pp. 2032–2050.
- [8] D. LEVERMORE, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 1021–1065.
- [9] C. D. LEVERMORE AND W. J. MOROKOFF, *The Gaussian moment closure for gas dynamics*, SIAM J. Appl. Math., 59 (1998), pp. 72–96.
- [10] P. MARKOWICH, C. RINGHOFER, AND C. SCHMEISER, *Semiconductor Equations*, Springer-Verlag, Vienna, 1990.

- [11] F. POUPAUD, *Diffusion approximation of the linear Boltzmann equation: Analysis of boundary layers*, *Asymptotic Anal.*, 4 (1991), pp. 293–317.
- [12] C. SCHMEISER AND A. ZWIRCHMAYR, *Galerkin methods for the semiconductor Boltzmann equation*, in *Proceedings of the International Congress of Industrial and Applied Mathematics 95*, Hamburg, 1995.
- [13] C. SCHMEISER AND A. ZWIRCHMAYR, *Convergence of moment methods for the semiconductor Boltzmann equation*, *SIAM J. Numer. Anal.*, to appear.
- [14] S. SELBERHERR, *Analysis of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
- [15] D. VENTURA, A. GNUDI, AND G. BACCARANI, *One dimensional simulation of a bipolar transistor by means of spherical harmonics expansions of the Boltzmann equation*, in *Proceedings SISDEP 91 Conference Zurich*, W. Fichtner, ed., 1991, pp. 203–205.
- [16] D. VENTURA, A. GNUDI, G. BACCARANI, AND F. ODEH, *Multidimensional spherical harmonics expansions for the Boltzmann equation for transport in semiconductors*, *Appl. Math. Lett.*, 5 (1992), pp. 85–90.