Universitat Autònoma de Barcelona
Facultat de Ciències, Departament de Matemàtiques

A CONTRIBUTION TO THE SIMULATION OF VLASOV-BASED MODELS

Ph.D. Thesis

supervised by

Prof. José Antonio Carrillo de la Plata — ICREA - Universitat Autònoma de Barcelona
Prof. Naoufel Ben Abdallah — Université Paul Sabatier (Toulouse)

written by

Francesco Vecil.

Thesis Committee.

Presidente	Rosa Donat Beneito	Universitat de València
Secretario	Josep Maria Mondelo González	Universitat Autònoma de Barcelona
Vocal 1	María José Cáceres	Universidad de Granada
Vocal 2	Stéphane Cordier	Université d'Orléans
Vocal 3	Armando Majorana	Università di Catania
Suplente 1	Pauline Godillon-Lafitte	Université de Lille
Suplente 2	Yolanda Vidal i Seguí	Universitat Politècnica de Catalunya

Contents

Introduction

The goal of this work is a contribution to the numerical simulation of kinetic models in electronic engineering and plasma physics [32, 84]. The MOSFET is an example of *electronic device*. It is a transistor, the building block for all the commonly used electronic objects, from CD players to laptop computers. Its particular interest is due to the effort being made to reduce it to nanoscale. During the last fifty years electronic components have been made smaller and smaller, which allows better performances for processors and saving of energy.
Electronic devices are physical solid state devices; they have a fixed electronic lattice, in which impurities are injected in order to modify electronic properties. A standard semiconductor is made of Silicon, which is a tetravalent atom and might be doped by injecting Phosphoros (P) or Arsenic (As) to obtain a negative doping (there is an excess of free electrons), or by injecting Boron (B), which is electron-deficient (it possesses a vacant p-orbital), to obtain a positive doping: a sort of excess of positive charges is produced, which is in fact an excess of electron holes. To give an idea of the dimensions of the doping phenomenon, in 1 cm^3 of Silicon there are order 10^{24} atoms; a low doping means injecting 10^{13} atoms per cm^3, while a high doping 10^{20} atoms per cm^3. The doping of semiconductors is essential in order to create a potential barrier high enough to induce an electron current.

Another physical object on which we shall focus our attention are *plasmas*, i.e. ionized gases: the electrons of the most external orbits are separated from the atom, which happens to gases when warmed up to $10^6 - 10^8$ Celsius degrees (that is why plasma is considered the fourth state of matter). Positive, negative and neutral charges dissociate, like 99% of the matter. Plasmas are a central point in the research in fusion energy.

Transport of charged particles and collisions are the main aspects which we need to describe when simulating electronic devices and plasmas. Basically two categories of models can be used:

- **microscopic**: we expect to have a precise and detailed physical description of the phenomena, even if more numerically costly; at kinetic level the motion is described by a probabilistic magnitude (the dynamical description being unrealizable because of the huge number

of particles involved) defined in the phase space (x, v), (x, p) or (x, k): the choice of the problem may make more suitable the use of the velocity v instead of the impulsion p or the wave vector k. The probability distribution function (pdf) is called $f(t, x, v)$, where $f(t, x, v)dxdv$ represents the number of particles in volume $dxdv$ around point (x, v) at time t;

- **macroscopic**: starting from the kinetic model, hydrodynamics limits [9, 15, 55] or fluid limits give on one side Euler and Navier-Stokes models and on the other side Spherical Harmonic Expansion [16, 39], Energy Transport [2] or Drift-Diffusion [3] models. Refer to [101] for more detail.

For the scope of this work we are interested in transient-state microscopic models, therefore the Vlasov operator is the main tool describing the motion of carriers under the effects of a positional force field $F(t, x)$ and the free motion. Collisions are taken into account by the Boltzmann operator: sometimes we shall neglect them, sometimes we shall just use a linear BGK, sometimes we shall take into account acoustic phonon scattering and optical phonon scattering (phonons are pseudo-particles describing the vibration of the lattice).

The force field $F(t, x)$ can be of very different natures. Usually it has the contribution of the self-consistent electrostatic field given by the Poisson equation. In the nanoMOSFET model, the Poisson equation is replaced by the Schrödinger-Poisson equation, in order to take into account the discrete energy levels of the confined dimensions. In plasmas simulations magnetic events cannot be neglected, so that we have to use Lorentz force (computed solving Maxwell equations) instead of just the electrostatic force.

Numerical schemes solving the transport problems have to deal with the *filamentation* problem: in the phase space strong gradients and oscillations appear, which is a physical phenomenon, essential in the evolution of Vlasov-based models, but it must be numerically treated in a proper way not to add *spurious* oscillations, i.e. numerical non-physical phenomena. In any numerical scheme, interpolations are required at some point; many are the possible choices but some of them do not properly deal with the physics of the model. In the first part of this thesis we develop Pointwise Essentially Non Oscillatory (PWENO) interpolation methods for direct reconstruction, and involve them in the solvers for transport problems. Moreover, we explore the possibilities of splitting algorithms between the dimensions of the phase space (dimensional splitting) and between transport and collisions in the collisional problems.

More precisely, for the time discretization of the models, two techniques are used: either we use the third order TVD (Total Variation Diminishing) Runge-Kutta algorithm in [28], or we split linear operators (like in the

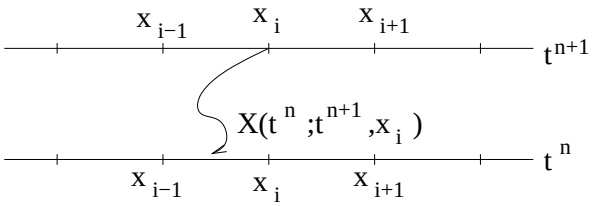


Figure 1: The direct Semi Lagrangian reconstruction, based on following backward the characteristics.

Boltzmann Transport Equation) by means of Strang splittings, a scheme introduced in 1976 by Cheng and Knorr [35] based on a previous 1968 article of Strang [98]. For example, for solving the BTE

$$\frac{\partial f}{\partial t} + a(v) \cdot \nabla_x f + F(x) \cdot \nabla_v f = \mathcal{Q}[f]$$

we can split it into two blocks

$$\begin{aligned} \text{transport:} \quad & \frac{\partial f}{\partial t} + a(v) \cdot \nabla_x f + F(x) \cdot \nabla_v f = 0 \\ \text{collisions:} \quad & \frac{\partial f}{\partial t} = \mathcal{Q}[f], \end{aligned}$$

solve them in separate steps and then recombine them: first perform a $\frac{\Delta t}{2}$-time step in transport, then a Δt-time step in collisions and then a $\frac{\Delta t}{2}$-time step in transport; this gives a second order scheme in time. Details are given in section 1.2.
Runge-Kutta schemes have the drawback of being time consuming (their explicit character constraints the time step to the CFL condition) and may be more difficult to implement, while in splitting schemes, being the blocks solved for separate, easier problems are solved. In principle, no constraints other than the physical consistency appear on the time stepping, which permits solving the problems in less time steps.

In section 1.3 two methods for the transport step are proposed: the direct Semi Lagrangian reconstruction in Figure 1 and the Flux Balance Method [42] in Figure 2. They must be completed by some interpolation technique, for which we use PWENO methods, which are the object of an accurate description in Section 1.1.

In Chapter 2 time splitting schemes (in Section 1.2) with Semi Lagrangian methods (in Section 1.3) for the advection parts and PWENO methods (in Section 1.1) for the reconstruction are tested and then applied to some 1D models: a Vlasov system with given potential with either a

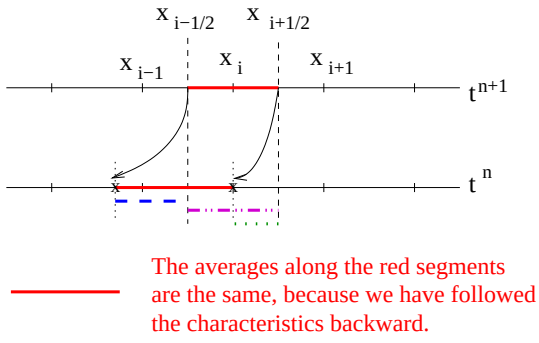


Figure 2: The Flux Balance Method, based on following backward the characteristics, but mass conservation is imposed.

linear BGK or a Fokker-Planck operator for collisions

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - x\frac{\partial f}{\partial v} = \mathcal{Q}[f]$$
$$\mathcal{Q}[f] = \begin{cases} \frac{1}{\tau}\left[\rho(x)M(v) - f\right] \\ \frac{1}{\tau}\frac{\partial}{\partial v}\left[vf + \Theta\frac{\partial f}{\partial v}\right] \end{cases},$$

a collisionless Vlasov-Poisson, where the force field is self-consisiently computed through the Poisson equation (which can be solved for instance like in Section 1.6)

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - \frac{\partial \Phi}{\partial x}\frac{\partial f}{\partial v} = 0$$
$$\frac{\partial^2 \Phi}{\partial x^2} = 1 - \int f dv,$$

which leads to the Landau damping if we set as initial condition a small perturbation on the equilibrium $f_0(x,v) = M(v)\left[1 + \epsilon\cos\left(\frac{x}{2}\right)\right]$, or to the two-stream instability, a typical vortex structure:

$$f_0(x,v) = Z\frac{2}{7\sqrt{2\pi}}\left(1+5v^2\right)\left[1+\alpha\left(\frac{\cos(2kx)+\cos(3kx)}{1.2}+\cos(kx)\right)\right]e^{-\frac{v^2}{2}}.$$

By these two tests we see how PWENO schemes are able to control spurious oscillations.

A semiconductor model is then studied in which the force field is self-consistently computed through the Poisson equation and the collisions are a correction of the linear BGK, the relaxation time being position-dependent

$$\tau(t,x) = \frac{m}{e}\frac{2\mu_0}{1+\sqrt{1+4((\mu_0/v_0)E(t,x))^2}}.$$

The main contribution of this part of the memory is the development of this non-linear interpolation procedure PWENO and its coupling with splitting techniques for solving kinetic equations. We also discuss the conservation of magnitudes like L^p-norms, total energy and entropy.

In Chapter 3, we apply the previously developed schemes to the simulation of an electronic device in which the single energy-band is given in the parabolic approximation, the electrostatic field is self-consistently computed via Poisson equation and collisions are taken into account with acoustic and optical phonons, in the linear approximation. The transport/collision equation is solved through recursive splitting schemes in rescaled cartesian coordinates, so that we reduce to the solution of 1D advection steps and collisions; as for the advection, the mass-preserving FBM scheme [42] has been chosen. The collision operator in the low-density approximation is

$$\mathcal{Q}[f](t,x,k) = \int_{\mathbb{R}^3} \left[S(k',k)f(t,x,k') - S(k,k')f(t,x,k)\right] dk'$$

with

$$\begin{aligned} S(k,k') = K\left[(n_q+1)\delta(\varepsilon(k')-\varepsilon(k)+\hbar\omega) + n_q\delta(\varepsilon(k')-\varepsilon(k)-\hbar\omega)\right] \\ + K_0\delta(\varepsilon(k')-\varepsilon(k)), \end{aligned}$$

where the reader can refer to Chapter 3 for the meaning and value of the physical constants. The solution of this operator requires performing integrations on an interval $[-\sqrt{\gamma},\sqrt{\gamma}]$ of the pdf $f(k_1,u)$ following a semicircle in the (k_1,u)-space, where $u = \|(k_2,k_3)\|$:

$$\int_{-\sqrt{\gamma}}^{\sqrt{\gamma}} f\left(k_1, \sqrt{\gamma-(k_1)^2}\right) dk_1.$$

Several interpolations are needed to compute the integral, which would not be the case in energy-adapted variables, the semicircle representing a level of the band-energy $\varepsilon(k)$ and the integration reducing to a single evaluation, so that just one interpolation may be needed. In our case several strategies have been tried, but without much improvement with respect to a plain along-the-line linear interpolation; refer to Section 1.4.1 for details.

In Chapter 5 we afford the numerical simulation of a 2D MOSFET (3D if we assume y-invariance) at a nanoscopic scale. The Silicon Debye length,

essentially the distance over which significant charge separation occurs, is given, at lattice temperature $T_L = 300K$, by

$$\lambda_D = \sqrt{\frac{\varepsilon_{Si}\varepsilon_0 k_B T_L}{e^2 N_D}} \approx 4088\sqrt{\frac{1}{N_D}}.$$

For the highly doped regions ($N^+ = 10^{20} cm^{-3}$), the Debye length is about $\lambda_D = 0.4nm$, while for the lowly doped regions ($N^+ = 10^{15} cm^{-3}$), the Debye length is about $\lambda_D = 130nm$. The device object of our study is $20nm$ long in the x-direction, in which we inject the carriers, and 8 nm in the z-direction, in which the confinement takes place, thanks to a built-in potential barrier of $3eV$ between the Si-layer and the SiO_2-layer. While along the transport dimension a classical description of the motion is satisfactory, the confined dimension is very short and energy levels become quantized: from now on, they are indexed by letter p, from lower energies, therefore more occupied, to higher energies, therefore less occupied. A quantum description in the confined dimension is more appropriate, therefore we assume the carriers behave like waves, their state being described by the 1D stationary-state Schrödinger equation (1D because the x-position only acts as a parameter):

$$-\frac{d}{dz}\left[\frac{1}{m_*}\frac{d\chi_p}{dz}\right] - q\left(V + V_c\right)\chi_p = \epsilon_p \chi_p.$$

Along the x-dimension carriers behave like particles, driven by the free motion, the self-consistent force field and having collisions with the phonons (but for the scope of this work we shall just use a linear BGK); the microscopic description of their motion is given by the Boltzmann Transport Equation:

$$\frac{\partial f_p}{\partial t} + \frac{1}{\hbar}\nabla_k \epsilon_p^{kin} \cdot \nabla_x f_p - \frac{1}{\hbar}\nabla_x \epsilon_p^{pot} \cdot \nabla_k f_p = \mathcal{Q}_p[f_p],$$

where the band kinetic energy is taken in its parabolic approximation:

$$\epsilon_p^{kin}(k) = \frac{\hbar^2 |k|^2}{2 m_* k_B T_L},$$

therefore it does not depend on the band. The force field is computed through the Poisson equation:

$$-\mathrm{div}_{x,z}\left[\varepsilon_R(x,z)\nabla_{x,z} V\right] = -\frac{q}{\varepsilon_0}\left[N[V] - N_D\right].$$

The Schrödinger and the Poisson equation cannot be decoupled, because we need the potential to compute the Schrödinger eigenproperties $\{\epsilon_p, \chi_p\}_p$, which are needed to compute the density

$$N[V] = \sum_p \rho_p |\chi_p[V]|^2.$$

The Schrödinger-Poisson problem is solved through a Newton iteration, in order to try a different strategy with respect to [101] where a Gummel iteration is used. Newton schemes have proven robustness and fast convergence, both in 1D and 2D. They require several times the solution of the 1D Schrödinger equation, which is discretized via standard finite differences and then diagonalized by means of a LAPACK routine called DSTEQR. They also need the solution of a "generalized" Poisson problem, i.e. a Poisson equation with non-local effects taken into account by an integral term. This fact makes the matrix full and not just tridiagonal.

As for the transport problem, several strategies have been tried: integrating either the original variable f_p or a slotboom variable g_p defined by

$$f_p(t,x,k) = g_p(t,x,k)e^{-\epsilon_p(t,x)-|k|^2}$$

by means of time splitting schemes [35] or Runge-Kutta schemes [28] based on Finite Differences. We have obtained the best results by integrating the original pdf f_p by Runge-Kutta-3. Any of these schemes is time consuming, because of two main reasons: Poisson has to be solved very often ($10^{-4}ps$) not to introduce oscillations due to its own overcorrections, and the drain-source potential should be applied gently not to initialize Newton schemes too far from the equilibrium.

Chapter 4 is slightly different from the other ones, it does not follow the line of direct application to some specific model. We are interested in intermediate approximations between the kinetic equation, which is microscopic therefore deeply detailed

$$\varepsilon\frac{\partial f_\varepsilon}{\partial t} + v\frac{\partial f_\varepsilon}{\partial x} = \frac{1}{\varepsilon}\mathcal{Q}[f_\varepsilon], \tag{1}$$

where we choose as collision operator a linear BGK

$$\mathcal{Q}[f_\varepsilon] = \int_V f_\varepsilon d\mu(v) - f_\varepsilon,$$

and the heat equation

$$\frac{\partial \rho}{\partial t} - \frac{\partial^2 \rho}{\partial x^2} = 0,$$

(the easiest macroscopic equation) this one being the formal limit of (1) as $\varepsilon \to 0$. We develop numerical schemes, based on splitting techniques, which are asymptotic-preserving as parameter ε tends to zero without need of resolving its smallness at meshes level. For the moment equations several closures have been proposed: we numerically show that the first order closure improves the zero-th order closure in the $\varepsilon \to 0$ limit.

One of the main contribution of this chapter is the scheme solving the kinetic equation (1). As speeds of propagation are order $\frac{1}{\varepsilon}$, a direct Semi-Lagrangian scheme or a Finite Differences scheme for Runge-Kutta discretization would become unbearably time consuming, either in order to

give a satisfactory precision or due the CFL constraint. In this work, we develop an asymptotic-preserving scheme based on the idea of decomposing f_ε into the sum of its mean value

$$\rho_\varepsilon = \int_V f_\varepsilon d\mu(v)$$

and fluctuations around this mean value, following the idea of a Hilbert expansion:

$$f_\varepsilon = \rho_\varepsilon + \varepsilon g_\varepsilon.$$

After plugging the decomposition into the kinetic equation and suppressing the non-leading terms in ε, we proceed by splitting schemes separating relaxations and advection:

Step 1.1 Relax g_ε

$$g_\varepsilon^{n+1/2} = e^{-\frac{\Delta t}{\varepsilon^2}} g_\varepsilon^n - (1 - e^{-\frac{\Delta t}{\varepsilon^2}}) v \frac{\partial \rho_\varepsilon^n}{\partial x}.$$

Step 1.2 Relax f_ε

$$f_\varepsilon^{n+1/2} = e^{-\frac{\Delta t}{\varepsilon^2}} f_\varepsilon^n + (1 - e^{-\frac{\Delta t}{\varepsilon^2}}) \rho_\varepsilon^n.$$

Step 1.3 In this first step the mean value does not change:

$$\rho_\varepsilon^{n+1/2} = \rho_\varepsilon^n.$$

Step 2.1 Solve for a Δt-time step

$$\frac{\partial f_\varepsilon}{\partial t} + v \frac{\partial g}{\partial x} = 0.$$

Step 2.2 Update the mean value

$$\rho_\varepsilon^{n+1} = \int_V f_\varepsilon d\mu(v).$$

Step 2.3 In this last step g_ε remains unchanged:

$$g_\varepsilon^{n+1} = g_\varepsilon^{n+1/2}.$$

Similar ideas are used to propose asymptotic-preserving schemes for the moment closure equations; refer to Chapter 4 for further details.

The organization of this thesis is as follows: in Chapter 1 the main proposed numerical techniques are described, Chapter 2 is devoted to the application of PWENO to the splitting schemes, Chapter 3 adapts these strategies to the semiconductor Boltzmann-Poisson problem, Chapter 4 introduces asymptotic-preserving schemes for diffusion approximations by splitting algorithms, and finally Chapter 5 deals with the quantum-kinetic coupled problem for solving charged particle transport in nanoMOSFETs.

Acknowledgments

The authors acknowledge

- the IGSOC/FI-IQUC Ph.D. fellowship of the Generalitat de Catalunya, from 2002 to 2006;
- the support from the project MTM2005-08024-C02 from the Spanish Ministry DGI-MEC (with reference to [26, 25]);
- the Marie Curie Early Stage Network DEASE: MEST-CT-2005-021122 funded by the European Union (which made possible my work in France during year 2006-2007);
- the computational resources of the Grid'5000 project which made the simulations possible (with reference to [23]);
- INRIA for the invitation in Lille in November 2006 (with reference to [23]);
- partial support of the bilateral project France-Spain HF2006-0198;
- the ACI Nouvelles Interfaces des Mathématiques Nb. ACINIM 176-2004 entitled "MOQUA" and funded by the French Linistry od Research (which made possible my participation to the International Workshop on Computational Microelectronics in Amherst, MA, Oct. 2007);
- Grup de recerca consolidat inside the IV Pla de Recerca de Catalunya (2005-2008), 2005SGR00611.

Chapter 1

Numerical instruments

In this chapter we develop most of the numerical techniques introduced and applied in this memory.

1.1 Pointwise Weighted Essentially Non Oscillatory interpolations

We describe the Pointwise Weighted Essentially Non Oscillatory (PWENO) interpolation: it is based on a direct reconstruction on the grid points and its goal is to avoid the introduction of spurious oscillations due to the Lagrange interpolation in presence of high gradients.

1.1.1 Introduction

We have an interval $[x_L, x_R]$, a division into a grid of N points

$$x_L = x_0 < x_1 < ... < x_{N-1} = x_R$$

and the values of a function f in these points:

$$\{f_i = f(x_i)\}_{i=0,...,N-1}.$$

Our goal is to reconstruct the function in the whole interval in a "non-oscillatory" way near the points where the function has high gradients or discontinuities.

In the numerical solution of conservation laws, hyperbolic and transport equations, the sharp shape of the solutions (shocks in conservation laws) and the total variation of the function have to be controlled.

1.1.2 Description of the method

We have an interval $[x_L, x_R]$ divided into a grid of N points

$$x_L = x_0 < x_1 < x_2 < ... < x_{N-1} = x_R$$

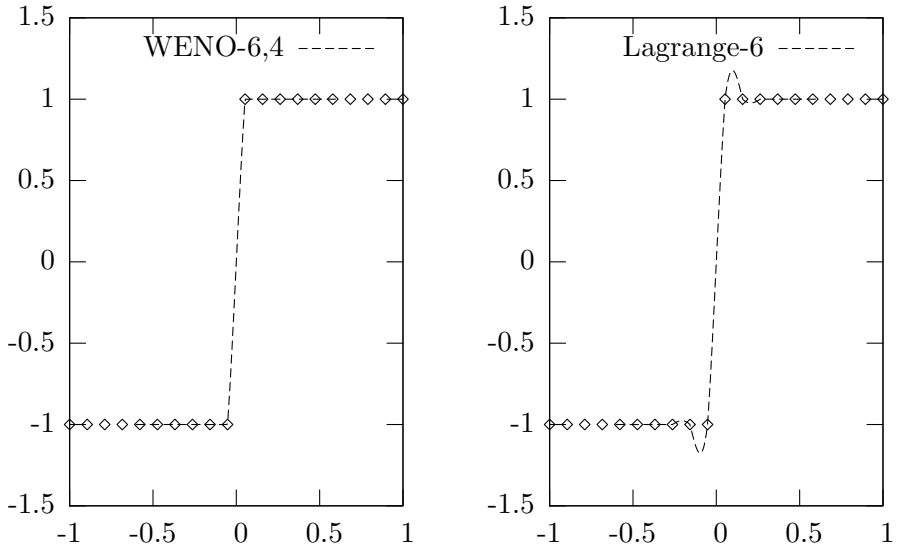


Figure 1: Comparison between an oscillating and a non-oscillating interpolation method.

$$\{f_i = f(x_i)\}_{i=0,...,N-1}.$$

We choose a stencil of *ntot* points

$$\mathcal{S} = \{x_{first}, ..., x_{last}\} = \{x_{first}, ..., x_{first+ntot-1}\}$$

and *nlp* substencils

$$\mathcal{S}_r = \{x_{last-lpo+1-r}, ..., x_{last-r}\} = \{x_{i-r}, ..., x_{last-r}\}$$

for $r = 0, ..., nlp-1$, where $lpo := ntot - nlp + 1$ and $i := last - lpo + 1$ (see Figure 2).

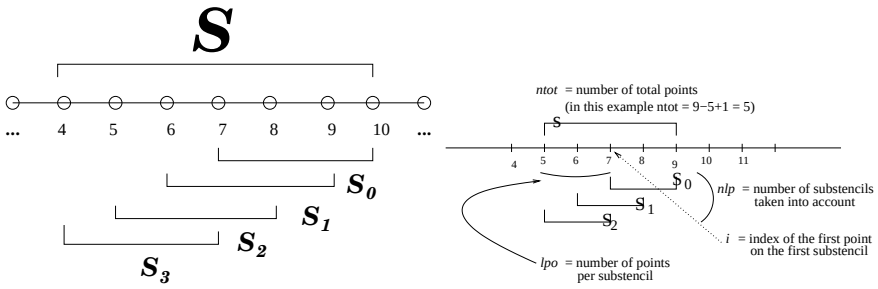


Figure 2: Left: the main stencil $\mathcal{S}$ with the substencils $\mathcal{S}_r$. Right: the meaning of parameters *nlp*, *lpo*, *ntot* and i.

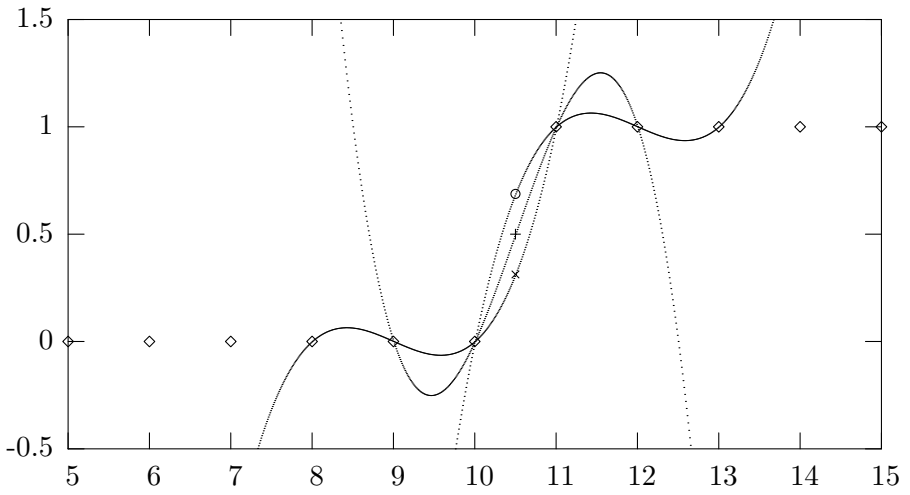


Figure 3: This figure shows how PWENO works with $ntot = 6$ and $lpo = 4$: it takes the (three, in this case) reconstructions given by the (three) Lagrange polynomials and then it will make an average between them.

The reconstructed value at points x will be a convex combination of the values given by the Lagrange polynomials in the stencils $\mathcal{S}_r$:

$$\sum_{r=0}^{nlp-1} \omega_r(x) p_r(x)$$

with p_r the Lagrange polynomials of degree $lpo - 1$ interpolating the stencil $\mathcal{S}_r$

$$p_r(x) = p_r^{lpo}(x) = \sum_{j=0}^{lpo-1} f_{i-r+j} c_{r,j}(x) \tag{1.1}$$

where

$$c_{r,j}(x) = \prod_{l=0, l\neq j}^{lpo-1} \frac{x - x_{i-r+l}}{x_{i-r+j} - x_{i-r+l}}.$$

$\omega_r(x)$ are weights which give more or less relevance to the stencils where the $p_r(x)$ are more or less "regular". Both the words "non-oscillatory" and "regular" will be rigorously defined in the next section.

The meaning of the parameters

1. cardinality of the main stencil $\mathcal{S}$ (**ntot**).
2. number of Lagrange polynomials (**nlp**).

3. Lagrange polynomial order (**lpo**), i.e. the number of points each polynomial interpolates.

1.1.3 The weights $\omega_r(x)$

WENO interpolation is given by

$$\sum_{r=0}^{nlp-1} \omega_r(x) p_r(x)$$

where $p_r(x)$ is defined in (1.1).

We need now to define the coefficients ω_r. First of all we need a measurement of the regularity of the Lagrange polynomials near x.

The smoothness indicators β_r

We shall call smoothness of the polynomial $p_r(x)$ a measure of its derivatives in the interval:

$$\mathcal{E} = [\mathcal{E}_L, \mathcal{E}_R] = \begin{cases} \left[x_{first+\left[\frac{ntot}{2}\right]-\frac{1}{2}}, x_{first+\left[\frac{ntot}{2}\right]+\frac{1}{2}}\right] & \text{if } ntot \text{ is odd} \\ \left[x_{first+\left[\frac{ntot}{2}\right]-1}, x_{first+\left[\frac{ntot}{2}\right]}\right] & \text{if } ntot \text{ is even} \end{cases} \tag{1.2}$$

where the dependencies of $\mathcal{E}_L$ and $\mathcal{E}_R$ will be omitted and $[x]$ means the integer part of x. Using the notation

$$i^* = first + \left[\frac{ntot}{2}\right]$$

we can also write

$$[\mathcal{E}_L, \mathcal{E}_R] = \begin{cases} \left[x_{i^*-\frac{1}{2}}, x_{i^*+\frac{1}{2}}\right] & \text{if } ntot \text{ is odd} \\ [x_{i^*-1}, x_{i^*}] & \text{if } ntot \text{ is even} \end{cases}$$

If the derivatives are large, the smoothness indicator is wanted to be large, and viceversa. The following measurement is proposed by Jiang and Shu in [64, page 207]:

$$\beta_r = \sum_{l=1}^{lpo-1} \int_{\mathcal{E}_L}^{\mathcal{E}_R} \Delta x^{2l-1} \left(D^l p_r\right)^2 dx$$

This is a weighted sum of L^2-norms of the derivatives, which we can see also as a weighted Sobolev norm of Dp_r in the interval $[\mathcal{E}_L, \mathcal{E}_R]$

$$\beta_r = \sum_{l=1}^{lpo-1} \Delta x^{2l-1} \left\| D^l p_r \right\|^2_{L^2(\mathcal{E}_L, \mathcal{E}_R)}$$

The weights Δx^{2l-1} are needed to make the terms of the sum independent of Δx, i.e., to make them all be of the same order. This will be clarified below. Other measurements would be possible, but we shall omit discussing this point.

Protoweights $\tilde{\omega}_r(x)$

Once we have computed the smoothness indicators, we define the $\tilde{\omega}_r(x)$ as

$$\tilde{\omega}_r(x) = \frac{d_r(x)}{(\epsilon + \beta_r)^p} \tag{1.3}$$

where $d_r(x)$ are some weights we need to optimize the order of the method (we shall discuss it later), and ϵ is a constant to avoid the denominator to be zero (in the code $\epsilon = 10^{-6}$ is used). The choice of $1 \leq p < \infty$ has no influence on the order of the method. In all the tests we have set $p = 2$. If we chose a greater p we would decide to give less weight to the stencils where the reconstruction is more irregular, and viceversa.

Weights $\omega_r(x)$

To get the weights $\omega_r(x)$ we just have to normalize the $\tilde{\omega}_r(x)$ given by (1.3).

$$\omega_r(x) = \frac{\tilde{\omega}_r(x)}{\sum_{j=0}^{nlp-1} \tilde{\omega}_j(x)}. \tag{1.4}$$

Still we have to find weights $d_r(x)$ to get the highest order method.

The weights $d_r(x)$

In order to get a high order method (if the function is smooth enough), we need coefficients $d_r(x)$ such that:

$$p(x) = \sum_{r=0}^{nlp-1} d_r(x) p_r(x) \tag{1.5}$$

where $p(x)$ is the $ntot - 1$-degree Lagrange polynomial interpolating the whole stencil $\mathcal{S}$

$$p(x) = \sum_{j=0}^{ntot-1} f_{first+j} \prod_{l=0, j\neq j}^{ntot-1} \frac{x - x_{first+j}}{x_{first+j} - x_{first+l}}. \tag{1.6}$$

Lagrange interpolation gives a *ntot*-order method; by mean of these coefficients we want PWENO-*ntot*,*lpo* interpolation to approach a *ntot*-order method non oscillatory for homogeneously regular functions, i.e., whenever all the weights β_r have the same order of magnitude.

Proposition 1.1.1 (Existence and uniqueness of the weights $d_r(x)$)**.** *Let $\mathcal{I} = [x_L, x_R] \subset \mathbb{R}$ be an interval, and let*

$$x_L = x_0 < x_1 < ... < x_{N-1} = x_R$$

be the grid. If x is not a point of the grid, then the weights $d_r(x)$ defined by (1.5) *are unique.*

Proof. The *ntot*-order Lagrange interpolation is exact on $\mathbb{P}^{ntot-1}$, i.e.,

$$f \in \mathbb{P}^{ntot-1} \rightarrow p[f](x) = f(x)$$

where $p[f]$ is the Lagrange polynomial which interpolates f in *ntot* points. If $\mathcal{B} = \{b_i\}_{i=0}^{ntot-1}$ is a basis of $\mathbb{P}^{ntot-1}$, $f(x) = \sum_{i=0}^{ntot-1} f_i b_i(x)$ and $c_{r,j}(x)$ are given by (1.1.2),

$$p\left[f\right](x) = p\left[\sum f_i b_i\right](x) = \sum_j \sum_i f_i b_i(x_j) c_{r,j}(x) \tag{1.7}$$

$$= \sum_i f_i \sum_j b_i(x_j) c_{r,j}(x) = \sum_i f_i p\left[b_i\right](x), \tag{1.8}$$

so we only need to impose condition (1.5) on the elements of a basis. Take as a basis of $\mathbb{P}^{ntot-1}$

$$\mathcal{B} = \{b_0, b_1, ..., b_{ntot-1}\} = \left\{1, \prod_{l=0}^{m}(x - x_{first+l}), m \in \{0, ..., ntot-2\}\right\}.$$

For $deg \leq lpo-1$, condition (1.5) gives $b_{deg}(x) = \sum_{r=0}^{nlp-1} d_r(x) b_{deg}(x)$ which means

$$\sum_{r=0}^{nlp-1} d_r(x) = 1.$$

For $lpo \leq deg \leq ntot-1$, condition (1.5) gives

$$b_{deg}(x) = \sum_{r=0}^{nlp-1} d_r(x) p_{r,deg}(x)$$

where $p_{r,deg}(x)$ is the Lagrange polynomial interpolating polynomial $b_{deg}(x)$ at points $\{x_{i-r}, ..., x_{i-r+lpo-1}\}$, i.e.,

$$p_{r,deg}(x) = \sum_{j=0}^{lpo-1} b_{deg}(x_{i-r+j}) c_{r,j}(x).$$

We get the following linear system

$$L = \left(\begin{array}{cccc|c} 1 & 1 & ... & 1 & 1 \\ p_{0,lpo}(x) & p_{1,lpo}(x) & ... & p_{nlp-1,lpo}(x) & b_{lpo}(x) \\ p_{0,lpo+1}(x) & p_{1,lpo+1}(x) & ... & p_{nlp-1,lpo+1}(x) & b_{lpo+1}(x) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{0,ntot-1}(x) & p_{1,ntot-1}(x) & ... & p_{nlp-1,ntot-1}(x) & b_{ntot-1} \end{array}\right).$$

Note that $b_{deg}(x_{i-r+j}) = 0$ for $i - r + j \leq deg - 1$, which implies that the matrix has the following appearance

$$L = \left(\begin{array}{ccccc|c} 1 & 1 & 1 & ... & 1 & 1 \\ p_{0,lpo}(x) & p_{1,lpo}(x) & ... & p_{nlp-2,lpo}(x) & 0 & b_{lpo}(x) \\ p_{0,lpo+1}(x) & p_{1,lpo+1}(x) & ... & 0 & 0 & b_{lpo+1}(x) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ p_{0,ntot-1}(x) & 0 & 0 & ... & 0 & b_{ntot-1}(x) \end{array} \right).$$

Call $L = (L_1|L_2)$. The computation of the determinant of L_1 is straightforward

$$\det(L_1) = \prod_{s=0}^{nlp-2} p_{s,(ntot-1)-s}(x) \neq 0$$

which is non-zero because none of the terms can be zero. Since

$$p_{s,(ntot-1)-s}(x) = \sum_{j=0}^{lpo-1} b_{(ntot-1)-s}(x_{i-s+j}) c_{s,j}(x)$$

and $b_{(ntot-1)-s}(x_{i-s+j}) = 0$, for $j = 0, ..., lpo - 2$, then

$$p_{s,(ntot-1)-s}(x) = b_{(ntot-1)-s}(x_{i-s+lpo-1}) c_{s,lpo-1}(x) \neq 0$$

because $b_{(ntot-1)-s}(x_{i-s+lpo-1}) \neq 0$; if not, we should get a contradiction: a non-zero polynomial of degree $(ntot - 1) - s$ would have $ntot - 1$ zeros. $c_{s,lpo-1}(x) \neq 0$ because x does not belong to the points of the grid. In this way, the existence and uniqueness of the $d_r(x)$ has been proven. □

Remark. If x is a grid point, the uniqueness would not be needed, because in fact any linear combination of the Lagrange polynomials interpolating that point would be suitable.

Code implementation of the weights $d_r(x)$

Imposing the definition of Lagrange polynomials (1.1) in (1.5), we get

$$\sum_{j=0}^{ntot-1} f_{i-r^*+j} c^{ntot}_{r^*,j}(x) = \sum_{r=0}^{nlp-1} d_r(x) \sum_{j=0}^{lpo-1} f_{i-r+j} c^{nlp}_{r,j}(x)$$

where r^* is defined in Figure 4. As the $d_r(x)$ do not depend on the values of f, we have to impose for every $0 \leq s \leq ntot - 1$ the coefficients of f_{i-r^*+s} to be equal:

$$c^{ntot}_{r^*,s}(x) = \sum_{r,j \ \ s.t. \ \ -r+j=s} d_r(x) c^k_{r,j}(x).$$

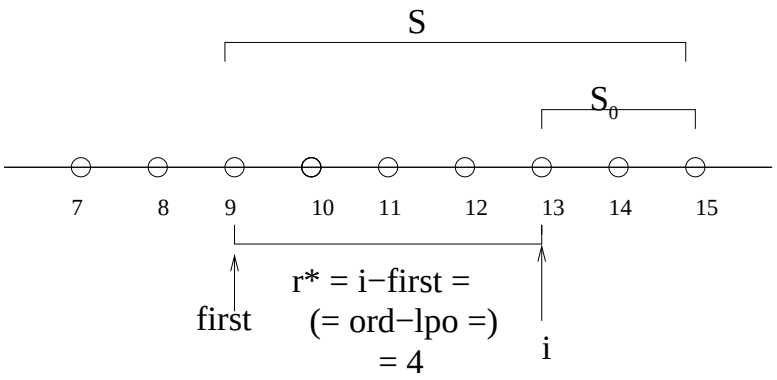


Figure 4: The parameter r^*

So, the linear system to be solved is represented by the following $ntot \times (nlp+1)ntot$ matrix:

$$(ls)_{i,j} = \begin{cases} c^{lpo}_{j,i+j-(nlp-1)}(x) & \text{if } nlp-1 \leq i+j \leq lpo-1+nlp-1 \\ 0 & \text{else} \end{cases},$$

with the known terms

$$(ls)_{i,nlp} = c^{ntot}_{r^*,i}.$$

Take now its submatrix $LS \in \mathbb{M}_{nlp \times nlp+1}$:

$$(LS)_{i,j} = \begin{cases} c^{lpo}_{j,i+j-(nlp-1)}(x) & if\ nlp-1 \leq i+j \leq lpo-1+nlp-1 \\ 0 & else \end{cases}.$$

Here we have nlp conditions for nlp unknowns, and the linear system is represented by an upper triangular matrix, which can be solved directly by a recursive procedure starting from the first line ($i = 0$).

A way for calculating explicitly the weights $d_r(x)$ as polynomials

We are able to get an explicit formula for the polynomials $d_r(x)$ by an iterative method. Assume the interpolation points are $\{x_0, x_1, ..., x_n\}$. Suppose we have two polynomials: $p(x)$ interpolates a function f at the points $x_0, x_1, ..., x_{n-1}$, and $q(x)$ interpolates at the points $x_1, x_2, ..., x_n$. A simple exactness argument (Aitken-Neville method, see [1, page 56]) allows to check that the polynomial interpolating f at the points $x_0, x_1, ..., x_n$ is given by

$$r(x) = \frac{p(x)(x-x_n) - q(x)(x-x_0)}{x_0 - x_n}.$$

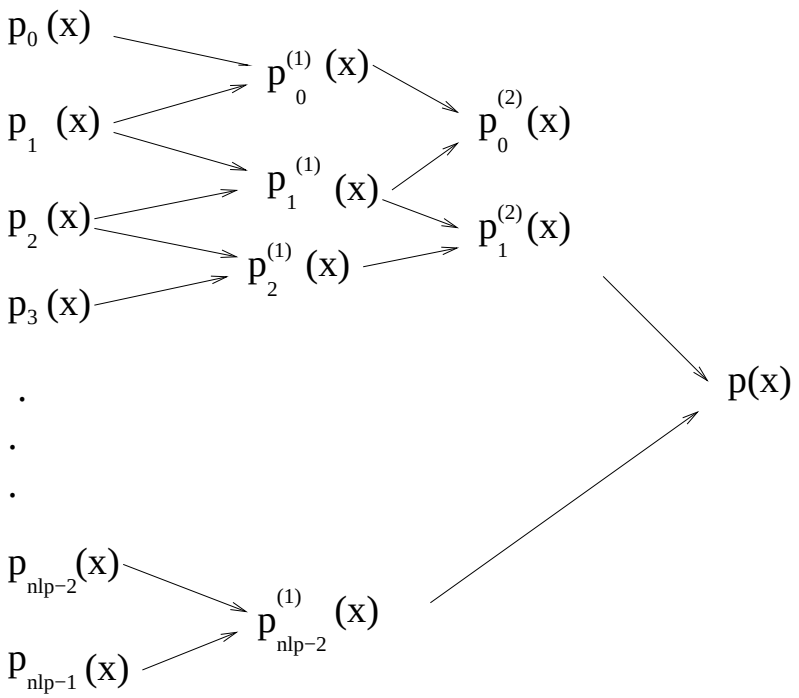


Figure 5: Explicit construction of the polynomials $\tilde{d}_r(x)$ by recursion.

An explicit recursion procedure based on the previous formula gives us the explicit value of $\tilde{d}_r(x)$. In the case of three points ($n = 2$) we get:

$$\begin{cases} \tilde{d}_0(x) = \frac{(x-x_{i-1})(x-x_{i-2})}{(x_{i-1}-x_{i+2})(x_{i-2}-x_{i+2})} \\ \tilde{d}_1(x) = -\left[\frac{(x-x_{i+2})(x-x_{i-2})}{(x_{i-1}-x_{i+2})(x_{i-2}-x_{i+2})} + \frac{(x-x_{i-2})(x-x_{i+2})}{(x_{i-2}-x_{i+1})(x_{i-2}-x_{i+2})}\right] \\ \tilde{d}_2(x) = \frac{(x-x_{i+1})(x-x_{i+2})}{(x_{i-2}-x_{i+1})(x_{i-2}-x_{i+2})} \end{cases}$$

Proposition 1.1.2 (Uniqueness of the weights $d_r(x)$)**.** *Let $\mathcal{I} = [x_L, x_R] \subset \mathbb{R}$ be an interval, and let*

$$x_L = x_0 < x_1 < ... < x_{N-1} = x_R$$

be the grid. The weights $d_r(x)$, recursively constructed as polynomials like in Figure 5, are unique in $\mathbb{P}^{ntot-lpo}$.

Proof. Polynomials $\tilde{d}_r(x)$ are explicitly constructed by the recursive method shown in Figure 5, so they exist (no denominator can be zero because $i \neq j \Rightarrow x_i \neq x_j$), are unique and their degree is $(ntot - 1) - (lpo - 1) = ntot - lpo$. Moreover, $\forall x \in \mathbb{R} \setminus \{x_i\}_{i=0,...,N-1}$,

$$\tilde{d}_r(x) = d_r(x)$$

for construction (they must verify the same property), then the weights $d_r(x)$, constructed as polynomials in $\mathbb{P}^{ntot-lpo}$, are unique $\forall x \in \mathbb{R}$. □

The choice of $d_r(x)$

Proposition 1.1.3. *If the weights $d_r(x)$ satisfty*

$$\sum_{r=0}^{nlp-1} d_r(x) = 1 \tag{1.9}$$

and

$$\omega_r(x) - d_r(x) = O(\Delta x^n),$$

then PWENO-ntot,lpo gives a $(lpo+n)$-order reconstruction.

The proof is developed in the following section.

The choice we have made for the weights $d_r(x)$ satisfy property 1.9, but it was also meant to approach the best order and the best accuracy. On one hand, if f is a homogeneous regular function (which means that all the β_r are of the same order), by this choice the $\omega_r(x)$ approach the $d_r(x)$, i.e. $p^{PWENO}(x)$ approaches $p(x)$ (defined in (1.6)). Even if PWENO is not *ntot*-order, we force it to behave like Lagrange, which is of highest order, in case of regular functions.

On the other hand, if the function is not regular, then the weights $\omega_r(x)$ are very different from the weights $d_r(x)$, like it had to be, because we want PWENO to behave differently from Lagrange near high-gradients.

1.1.4 The order of the method

First of all we recall a standard result about the error committed by Lagrange interpolation (see [68, page 291]):

Proposition 1.1.4. *Let $\mathcal{I} = [x_L, x_R] \subset \mathbb{R}$ be an interval. If $f \in \mathcal{C}^{n+1}[x_L, x_R]$ and $p \in \mathbb{P}^n$ is the polynomial interpolating f in $n+1$ different points $\{x_0, x_1, ..., x_n\}$ in $[x_L, x_R]$. Then, $\forall x \in [x_L, x_R]$ there is a $\xi_x \in]x_L, x_R[$ such that*

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i).$$

In the case of a regular grid this means that

$$p(x) = f(x) + O(\Delta x^{n+1}).$$

We shall now introduce a simple lemma.

Lemma 1.1.1. *If*

$$\beta_r = A(1 + O(\Delta x^n)) \tag{1.10}$$

where A is a non-zero quantity independent of r, then

$$\omega_r(x) - d_r(x) = O(\Delta x^n).$$

Proof of lemma 1.1.1. By straighforward calculations,

$$\frac{1}{{\beta_r}^p} = \frac{1}{D^p} + O(\Delta x^n)$$

and

$$\tilde{\omega_r} = \frac{d_r}{{\beta_r}^p} = d_r \left[\frac{1}{D^p} + O(\Delta x^n)\right] = \frac{1}{D^p} d_r + O(\Delta x^n)$$

with

$$\sum_j \tilde{\omega_j} = \frac{1}{D^p} d_r + O(\Delta x^n).$$

Finally,

$$\omega_r = \frac{\tilde{\omega_r}}{\sum_j \tilde{\omega_j}} = \frac{\frac{1}{D^p} d_r + O(\Delta x^n)}{\frac{1}{D^p} + O(\Delta x^n)} = d_r + O(\Delta x^n).$$

□

Proof of proposition 1.1.3

Since

$$\begin{aligned} p^W(x) - f(x) &= p^W(x) - p^L(x) + p^L(x) - f(x) = \\ &= \sum \omega_r(x) p_r(x) - \sum d_r(x) p_r(x) + \sum d_r(x) p_r(x) - f(x) \\ &= \sum \omega_r(x) p_r(x) - \sum d_r(x) p_r(x) + O(\Delta x^{ntot}), \end{aligned}$$

we already know that the method cannot be more than *ntot*-order, and we need to calculate

$$\mathcal{I} = \sum_{r=0}^{nlp-1} \omega_r(x) p_r(x) - \sum_{r=0}^{nlp-1} d_r(x) p_r(x)$$

to check which is the order. By simple manipulations

$$\begin{aligned} \mathcal{I} &= \sum_{r=0}^{nlp-1} \omega_r(x) p_r(x) - \sum_{r=0}^{nlp-1} d_r(x) p_r(x) \\ &= \sum_{r=0}^{nlp-1} \left[\omega_r(x) - d_r(x)\right] p_r(x) \\ &= \sum_{r=0}^{nlp-1} \left[\omega_r(x) - d_r(x)\right] \left[p_r(x) - f(x)\right]. \end{aligned}$$

Due to (1.9), we get

$$\begin{aligned}\mathcal{I} &= \sum_{r=0}^{nlp-1} \left[\omega_r(x) - d_r(x)\right] \left[p_r(x) - f(x)\right] \\ &= \sum_{r=0}^{nlp-1} \left[\omega_r(x) - d_r(x)\right] O(\Delta x^{lpo})\end{aligned}$$

and because of proposition 1.1.4 and lemma 1.10, we finally deduce

$$\begin{aligned}\mathcal{I} &= \sum_{r=0}^{nlp-1} \left[\omega_r(x) - d_r(x)\right] O(\Delta x^{lpo}) \\ &= \sum_{r=0}^{nlp-1} O(\Delta x^{n}) O(\Delta x^{lpo}).\end{aligned}$$

□

Manipulation on the β_r

Our problem is now translated to the Taylor expansion of the smoothness indicators β_r. Once we have been able to compute n in (1.10), then the method will be order $lpo + n$. The weights can be expressed as

$$\begin{aligned}\beta_r &= \sum_{l=1}^{lpo-1} \int_{\mathcal{E}} \Delta x^{2l-1} \left[D^l p_r(x)\right]^2 dx \\ &= \sum_{l} \Delta x^{2l-1} \int_{\mathcal{E}} \left[D^l \sum_{j=0}^{nlp-1} c_{r,j}^{lpo}(x) f_{i-r+j}\right]^2 dx \\ &= \sum_{l} \Delta x^{2l-1} \int_{\mathcal{E}} \left[\sum_{j} f_{i-r+j} D^l c_{r,j}(x)\right]^2 dx \\ &= \sum_{l} \Delta x^{2l-1} \int_{\mathcal{E}} \sum_{j,k=0}^{lpo-1} f_{i-r+j} f_{i-r+k} D^l c_{r,j}(x) D^l c_{r,k}(x) dx \\ &= \sum_{l} \Delta x^{2l-1} \sum_{j,k} f_{i-r+j} f_{i-r+k} \int_{\mathcal{E}} D^l c_{r,j}(x) D^l c_{r,k}(x) dx \\ &= \sum_{j,k} f_{i-r+j} f_{i-r+k} \sum_{l=1}^{lpo-1} \int_{\mathcal{E}} \Delta x^{2l-1} D^l c_{r,j}(x) D^l c_{r,k}(x) dx \\ &= \sum_{j,k} f_{i-r+j} f_{i-r+k} K_{j,k}^r\end{aligned}$$

where K^r is a symmetric matrix defined as

$$K^r_{j,k} = \sum_{l=1}^{lpo-1} \int_{\mathcal{E}} \Delta x^{2l-1} D^l c_{r,j} D^l c_{r,k} dx.$$

We can now expand $f(x)$ around the point x_i:

$$f(x) = \sum_n \frac{1}{n!} f^{(n)}(x_i)(x - x_i)^n$$

so that

$$\begin{aligned} f_{i-r+j} &= \sum_n \frac{1}{n!} f^{(n)}(x_i)(\Delta x(i - r + j - i))^n && (1.11) \\ &= \sum_n \frac{1}{n!} f^{(n)}(x_i)(j - r)^n \Delta x^n. && (1.12) \end{aligned}$$

Therefore,

$$\begin{aligned} \beta_r &= \sum_{j,k} K^r_{j,k} \sum_{n,m} \frac{1}{n!m!} f^{(n)}(x_i) f^{(m)}(x_i)(j - r)^n (k - r)^m \Delta x^{n+m} K \\ &= \sum_{n,m} \Delta x^{n+m} f^{(n)}(x_i) f^{(m)}(x_i) \frac{1}{n!m!} \sum_{j,k} K^r_{j,k} (j - r)^n (k - r)^m \\ &= \sum_{n,m} \Delta x^{n+m} f^{(n)} f^{(m)} \mathcal{D}^r_{n,m} \end{aligned}$$

where $\mathcal{D}^r$ is a symmetric matrix defined as

$$\mathcal{D}^r_{n,m} = \frac{1}{n!m!} \sum_{j,k=0}^{lpo-1} K^r_{j,k} (j - r)^n (k - r)^m.$$

Numerical computation of $\mathcal{D}^r$ for method WENO-6,4 are given in (1.13)-(1.15) in the appendix of this chapter. These computations show that

$$\forall 0 \le r \le 2, \qquad \mathcal{D}^r_{1,1} = 1, \mathcal{D}^r_{2,2} = \frac{4}{3}$$

and therefore, if $f' \neq 0$ then

$$\beta_r = (f'(x_i)\Delta x)^2 (1 + O(\Delta x))$$

(where $A = (f'(x_i)\Delta x)^2$, $n = 1$ in (1.10)), and if $f'(x_i) = 0$ then

$$\beta_r = \frac{4}{3} (f''(x_i)\Delta x^2)^2 (1 + O(\Delta x))$$

($A = \frac{4}{3}(f''(x_i)\Delta x^2)^2$, $n = 1$ in (1.10)), i.e. that the method is $lpo + 1 = 4 + 1 = 5$ order.
As for WENO-5,3, we see from (1.16)-(1.18) that if $f'(x_i) \neq 0$ then

$$\beta_r = (f'(x_i)\Delta x)^2(1 + O(\Delta x^2))$$

(where $A = (f'(x_i)\Delta x)^2$, $n = 2$ in (1.10)). This implies that WENO-5,3 is $lpo + 2 = 3 + 2 = 5$ order. □

Numerical example

Take $u_0(x) = \exp(x)$ in $[-1; 1]$, reconstruct the value at $x = 0$ (if the number of points is even it will never belong to the grid) by PWENO-6,4 method and compute the difference $|\exp(0) - num.val.|$:

$$\left(\begin{array}{l|l|l} points & L^\infty - error & L^\infty - order \\ 20 & 6.88 \times 10^{-9} & \\ 40 & 8.95 \times 10^{-11} & 6.263432 \\ 80 & 1.28 \times 10^{-12} & 6.119669 \\ 160 & 1.93 \times 10^{-14} & 6.059640 \\ 320 & 2.22 \times 10^{-16} & 6.442943 \end{array}\right)$$

While at least 5 was expected, we see that we get 6, due to the homogeneous regularity of the function.

Results for matrix $\mathcal{D}$

For WENO-6,4 (the values are given in absolute values):

$$\mathcal{D}^0 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.500 & 0.166 & 0.041 & 0.008 & 0.001 & 0.000 \\ 0.000 & 0.500 & 1.333 & 0.625 & 0.159 & 0.402 & 0.312 & 0.167 \\ 0.000 & 0.166 & 0.625 & 1.383 & 1.543 & 1.154 & 0.657 & 0.305 \\ 0.000 & 0.041 & 0.159 & 1.543 & 2.472 & 2.101 & 1.272 & 0.610 \\ 0.000 & 0.008 & 0.402 & 1.154 & 2.101 & 1.847 & 1.134 & 0.548 \\ 0.000 & 0.001 & 0.312 & 0.657 & 1.272 & 1.134 & 0.700 & 0.339 \\ 0.000 & 0.000 & 0.166 & 0.305 & 0.610 & 0.548 & 0.339 & 0.165 \end{bmatrix}, \quad (1.13)$$

$$\mathcal{D}^1 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.500 & 0.166 & 0.041 & 0.008 & 0.001 & 0.000 \\ 0.000 & 0.500 & 1.333 & 0.625 & 0.381 & 0.139 & 0.048 & 0.013 \\ 0.000 & 0.166 & 0.625 & 1.383 & 0.729 & 0.340 & 0.114 & 0.033 \\ 0.000 & 0.041 & 0.381 & 0.729 & 0.393 & 0.181 & 0.061 & 0.018 \\ 0.000 & 0.008 & 0.139 & 0.340 & 0.181 & 0.084 & 0.028 & 0.008 \\ 0.000 & 0.001 & 0.048 & 0.114 & 0.061 & 0.028 & 0.009 & 0.002 \\ 0.000 & 0.000 & 0.013 & 0.033 & 0.018 & 0.008 & 0.002 & 0.000 \end{bmatrix}, \quad (1.14)$$

$$\mathcal{D}^2 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.500 & 0.166 & 0.041 & 0.008 & 0.001 & 0.000 \\ 0.000 & 0.500 & 1.333 & 0.625 & 0.159 & 0.139 & 0.041 & 0.013 \\ 0.000 & 0.166 & 0.625 & 1.383 & 0.625 & 0.340 & 0.111 & 0.033 \\ 0.000 & 0.041 & 0.159 & 0.625 & 0.300 & 0.157 & 0.052 & 0.015 \\ 0.000 & 0.008 & 0.139 & 0.340 & 0.157 & 0.084 & 0.027 & 0.008 \\ 0.000 & 0.001 & 0.041 & 0.111 & 0.052 & 0.027 & 0.009 & 0.002 \\ 0.000 & 0.000 & 0.013 & 0.033 & 0.015 & 0.008 & 0.002 & 0.000 \end{bmatrix}. \quad (1.15)$$

For WENO-5,3:

$$\mathcal{D}^0 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.333 & 0.250 & 0.116 & 0.041 & 0.012 \\ 0.000 & 0.000 & 1.083 & 1.083 & 0.631 & 0.270 & 0.093 & 0.027 \\ 0.000 & 0.333 & 1.083 & 1.194 & 0.715 & 0.309 & 0.107 & 0.031 \\ 0.000 & 0.250 & 0.631 & 0.715 & 0.431 & 0.187 & 0.064 & 0.018 \\ 0.000 & 0.116 & 0.270 & 0.309 & 0.187 & 0.081 & 0.028 & 0.008 \\ 0.000 & 0.041 & 0.093 & 0.107 & 0.064 & 0.028 & 0.009 & 0.002 \\ 0.000 & 0.012 & 0.027 & 0.031 & 0.018 & 0.008 & 0.002 & 0.000 \end{bmatrix}, \quad (1.16)$$

$$\mathcal{D}^1 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.166 & 0.000 & 0.008 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.083 & 0.000 & 0.090 & 0.000 & 0.003 & 0.000 \\ 0.000 & 0.166 & 0.000 & 0.027 & 0.000 & 0.001 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.090 & 0.000 & 0.007 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.008 & 0.000 & 0.001 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.003 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}, \quad (1.17)$$

$$\mathcal{D}^2 = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.333 & 0.250 & 0.116 & 0.041 & 0.012 \\ 0.000 & 0.000 & 1.083 & 1.083 & 0.631 & 0.270 & 0.093 & 0.027 \\ 0.000 & 0.333 & 1.083 & 1.194 & 0.715 & 0.309 & 0.107 & 0.031 \\ 0.000 & 0.250 & 0.631 & 0.715 & 0.431 & 0.187 & 0.064 & 0.018 \\ 0.000 & 0.116 & 0.270 & 0.309 & 0.187 & 0.081 & 0.028 & 0.008 \\ 0.000 & 0.041 & 0.093 & 0.107 & 0.064 & 0.028 & 0.009 & 0.002 \\ 0.000 & 0.012 & 0.027 & 0.031 & 0.018 & 0.008 & 0.002 & 0.000 \end{bmatrix}. \quad (1.18)$$

1.1.5 Explicit coefficients for some PWENO interpolations

Here, we give the explicit polynomials $p_r(x)$ and weights β_r (in the case of regular grids) for the most used PWENO interpolations, in order to optimize the codes.

PWENO-4,3

$$\begin{cases} d_0(x) &= \frac{x-x_{i-1}}{3\Delta x} \\ d_1(x) &= -\frac{x-x_{i-2}}{3\Delta x} \\ \beta_0 &= \frac{13}{12}f_{i+2}f_{i+2} - \frac{13}{3}f_{i+1}f_{i+2} + \frac{13}{12}f_if_{i+2} \\ & \quad +\frac{16}{3}f_{i+1}f_{i+1} - \frac{19}{3}f_if_{i+1} + \frac{25}{12}f_if_i \\ \beta_1 &= \frac{13}{12}f_{i-1}f_{i-1} + \frac{16}{3}f_{i-1}f_{i+1} - \frac{13}{3}f_{i-1}f_i \\ & \quad +\frac{25}{12}f_{i+1}f_{i+1} - \frac{19}{3}f_if_{i+1} + \frac{16}{3}f_if_i \end{cases}$$

PWENO-5,3

$$\begin{cases} d_0(x) &= \frac{(x-x_{i-1})(x-x_{i-2})}{12\Delta x^2} \\ d_1(x) &= -\frac{(x-x_{i+2})(x-x_{i-2})}{6\Delta x^2} \\ d_2(x) &= \frac{(x-x_{i+1})(x-x_{i+2})}{12\Delta x^2} \\ \beta_0 &= \frac{10}{3}f_i^2 + \frac{25}{3}f_{i+1}^2 + \frac{4}{3}f_{i+2}^2 \\ & \quad -\frac{31}{3}f_if_{i+1} + \frac{11}{3}f_if_{i+2} - \frac{19}{3}f_{i+1}f_{i+2} \\ \beta_1 &= \frac{4}{3}f_{i-1}^2 + \frac{13}{3}f_i^2 + \frac{4}{3}f_{i+1}^2 \\ & \quad -\frac{13}{3}f_{i-1}f_i + \frac{5}{3}f_{i-1}f_{i+1} - \frac{13}{3}f_if_{i+1} \\ \beta_2 &= \frac{4}{3}f_{i-2}^2 + \frac{25}{3}f_{i-1}^2 + \frac{10}{3}f_i^2 \\ & \quad -\frac{19}{3}f_{i-2}f_{i-1} + \frac{11}{3}f_{i-2}f_i - \frac{31}{3}f_{i-1}f_i \end{cases}$$

PWENO-6,4

$$\begin{cases} d_0(x) &= \frac{(x-x_{i-1})(x-x_{i-2})}{20\Delta x^2} \\ d_1(x) &= -\frac{(x-x_{i+3})(x-x_{i-2})}{10\Delta x^2} \\ d_2(x) &= \frac{(x-x_{i+2})(x-x_{i+3})}{20\Delta x^2} \\ \beta_0 &= \frac{248}{15}f_{i+2}^2 - \frac{2309}{60}f_{i+1}f_{i+2} + \frac{439}{30}f_if_{i+2} - \frac{553}{60}f_{i+2}f_{i+3} + \frac{721}{30}f_{i+1}^2 \\ & \quad -\frac{1193}{60}f_if_{i+1} + \frac{103}{10}f_{i+1}f_{i+3} + \frac{407}{90}f_i^2 - \frac{683}{180}f_if_{i+3} + \frac{61}{45}f_{i+3}^2 \\ \beta_1 &= \frac{61}{45}f_{i+2}^2 + \frac{61}{45}f_{i-1}^2 + \frac{179}{30}f_{i-1}f_{i+1} - \frac{141}{20}f_{i-1}f_i - \frac{293}{180}f_{i-1}f_{i+2} \\ & \quad -\frac{141}{20}f_{i+1}f_{i+2} + \frac{179}{30}f_if_{i+2} + \frac{331}{30}f_{i+1}^2 - \frac{1259}{60}f_if_{i+1} + \frac{331}{30}f_i^2 \\ \beta_2 &= \frac{248}{15}f_{i-1}^2 + \frac{439}{30}f_{i-1}f_{i+2} - \frac{2309}{60}f_{i-1}f_i + \frac{407}{90}f_{i+1}^2 - \frac{1193}{60}f_if_{i+1} \\ & \quad +\frac{721}{30}f_i^2 + \frac{103}{10}f_{i-2}f_i - \frac{553}{60}f_{i-2}f_{i-1} + \frac{61}{45}f_{i-2}^2 - \frac{683}{180}f_{i-2}f_{i+1} \end{cases}$$

FDWENO-5,3

Here, we also summarize for completeness and comparison the weights and formulas for the case of Finite-Differences WENO methods developed by [64].

$$\hat{f}^{\pm}_{j+\frac{1}{2}} = \omega_0^{\pm} p_0^{\pm} + \omega_1^{\pm} p_1^{\pm} + \omega_2^{\pm} p_2^{\pm}.$$

$$\begin{aligned}
\beta_0^+ &= \frac{13}{12}\left[f_{j-2} - 2f_{j-1} + f_j\right]^2 + \frac{1}{4}\left[f_{j-2} - 4f_{j-1} + 3f_j\right]^2 \\
\beta_1^+ &= \frac{13}{12}\left[f_{j-1} - 2f_j + f_{j+1}\right]^2 + \frac{1}{4}\left[f_{j-1} - f_{j+1}\right]^2 \\
\beta_2^+ &= \frac{13}{12}\left[f_j - 2f_{j+1} + f_{j+2}\right]^2 + \frac{1}{4}\left[3f_j - 4f_{j+1} + f_{j+2}\right]^2 \\
d_0^+ &= \frac{1}{10}, \qquad d_1^+ = \frac{6}{10}, \qquad d_2^+ = \frac{3}{10} \\
p_0^+ &= \frac{1}{3}f_{j-2} - \frac{7}{6}f_{j-1} + \frac{11}{6}f_j \\
p_1^+ &= -\frac{1}{6}f_{j-1} + \frac{5}{6}f_j + \frac{1}{3}f_{j+1} \\
p_2^+ &= \frac{1}{3}f_j + \frac{5}{6}f_{j+1} - \frac{1}{6}f_{j+2}.
\end{aligned}$$

$$\begin{aligned}
\beta_0^- &= \frac{13}{12}\left[f_{j+3} - 2f_{j+2} + f_{j+1}\right]^2 + \frac{1}{4}\left[f_{j+3} - 4f_{j+2} + 3f_{j+1}\right]^2 \\
\beta_1^- &= \frac{13}{12}\left[f_{j+2} - 2f_{j+1} + f_j\right]^2 + \frac{1}{4}\left[f_{j+2} - f_j\right]^2 \\
\beta_2^- &= \frac{13}{12}\left[f_{j+1} - 2f_j + f_{j-1}\right]^2 + \frac{1}{4}\left[3f_{j+1} - 4f_j + f_{j-1}\right]^2 \\
d_0^- &= \frac{3}{10}, \qquad d_1^- = \frac{6}{10}, \qquad d_2^- = \frac{1}{10} \\
q_0^- &= \frac{1}{3}f_{j+3} - \frac{7}{6}f_{j+2} + \frac{11}{6}f_{j+1} \\
q_1^- &= -\frac{1}{6}f_{j+2} + \frac{5}{6}f_{j+1} + \frac{1}{3}f_j \\
q_2^- &= \frac{1}{3}f_{j+1} + \frac{5}{6}f_j - \frac{1}{6}f_{j-1}.
\end{aligned}$$

1.2 Strang's time splitting

Take the equation

$$\frac{\partial \Psi}{\partial t} = \mathcal{L}\Psi$$

where $\mathcal{L}$ is a linear operator generator of a $\mathcal{C}_0$-semigroup. Its formal solution is

$$\Psi(t) = \exp(t\mathcal{L})\Psi(0).$$

We split the operator $\mathcal{L}$ into two parts:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$$

and we take them as the infinitesimal generators of the $\mathcal{C}_0$-semigroups

$$\mathcal{F}_i(t) = \exp\left(t\mathcal{L}_i\right).$$

Note that $\mathcal{L}_1 + \mathcal{L}_2$ is also the infinitesimal generator of a $\mathcal{C}_0$-semigroup because of Trotter's product theorem. *Strang's splitting* consists in taking

$$\mathcal{F}(\Delta t) = \mathcal{F}_1\left(\frac{\Delta t}{2}\right)\mathcal{F}_2\left(\Delta t\right)\mathcal{F}_1\left(\frac{\Delta t}{2}\right).$$

This can be proven to be a second order scheme in the sense that

$$\mathcal{F}\left(\Delta t\right) = \exp\left(\Delta t\mathcal{L}\right) + O\left(\Delta t^3\right).$$

Consider now two cases of Strang's splittings, the two we are going to use.

1.2.1 Strang's time splitting between Vlasov and Boltzmann

Given equation

$$\frac{\partial f}{\partial t} + v\cdot\nabla_x f + F\cdot\nabla_v f = \mathcal{Q}[f]$$

where the force field $F(t,x)$ and the Boltzmann operator $\mathcal{Q}[f]$ are known, we advance a step in time by advancing separately in Vlasov and in Boltzmann parts, i.e., given

$$f(t^n, x_i, v_j)$$

we proceed in this way:

1. Perform a $\frac{\Delta t}{2}$ time step in Vlasov part, i.e., solving

   $$\frac{\partial f}{\partial t} + v\cdot\nabla_x f + F\cdot\nabla_v f = 0$$

   to get

   $$f^*(t^n, x_i, v_j).$$

2. Perform a Δt time step in Boltzmann part, i.e., solving

   $$\frac{\partial f}{\partial t} = \mathcal{Q}[f]$$

   to get

   $$f^{**}(t^n, x_i, v_j).$$

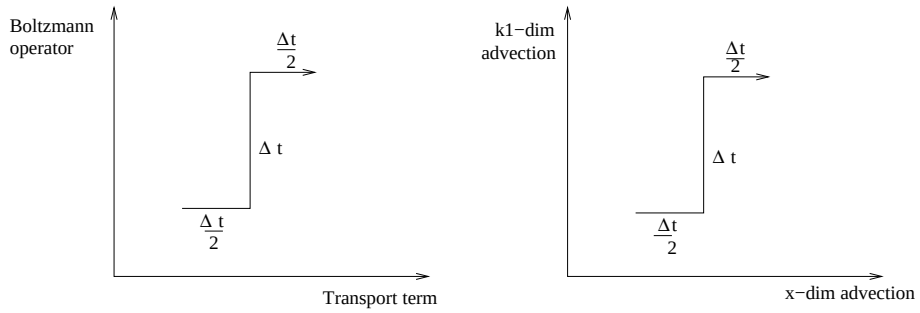


Figure 6: Time splitting schemes.

3. Perform a $\frac{\Delta t}{2}$ time step in Vlasov part, i.e., solving

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + F \cdot \nabla_v f = 0$$

to get

$$f(t^{n+1}, x_i, v_j).$$

like in Figure 6.

Remark. In our previous notation, we split the operator

$$\mathcal{L} = -v \cdot \nabla_x - F \cdot \nabla_v + \mathcal{Q}$$

into

$$\begin{cases} \mathcal{L}_1 = -v \cdot \nabla_x - F \cdot \nabla_v \\ \mathcal{L}_2 = \mathcal{Q} \end{cases}.$$

1.2.2 Strang's splitting between x and v

Our purpose is now to solve

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + F \cdot \nabla_v f = 0$$

given the force field $F(t, x)$. The procedure (which was originally introduced by Cheng and Knorr [35]) is to split the Vlasov equation into either phases x and v, in this way: given $f\left(t^n, x, v\right)$,

1. Consider v fixed, and take the free transport equation

$$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = 0$$

and perform a $\frac{\Delta t}{2}$ time step in the x-direction to get

$$f^*\left(t^n, x, v\right) = f\left(t^n, x - v\frac{\Delta t}{2}, v\right)$$

by one of the advection algorithms previously described.

2. Compute the force field $F\left(f^*\left(t^n, x, v\right)\right)$.

3. Consider x fixed, and take equation

$$\frac{\partial f^*}{\partial t} + F\frac{\partial f^*}{\partial v} = 0$$

and perform a Δt time step in the v-direction to obtain

$$f^{**}\left(t^n, x, v\right) = f^*\left(t^n, x, v - F\Delta t\right).$$

4. Consider v fixed, and take the free transport equation

$$\frac{\partial f^{**}}{\partial t} + v\frac{\partial f^{**}}{\partial x} = 0$$

and perform a $\frac{\Delta t}{2}$ time step in the x-direction to obtain

$$f\left(t^{n+1}, x, v\right) = f^{**}\left(t^n, x - v\frac{\Delta t}{2}, v\right).$$

This scheme is second order in time.

Remark. In our former terms, we split the operator

$$\mathcal{L} = -v \cdot \nabla_x - F \cdot \nabla_v$$

into

$$\begin{cases} \mathcal{L}_1 = -v \cdot \nabla_x \\ \mathcal{L}_2 = -F \cdot \nabla_v \end{cases}.$$

1.3 Semi-lagrangian solvers for the linear advection

When solving a transport step, the fundamental block which we need to solve is the linear advection

$$\begin{cases} \dfrac{\partial f}{\partial t} + a\dfrac{\partial f}{\partial x} = 0 \\ f(t_0, x) = f_0(x) \end{cases},$$

where $a \in \mathbb{R}$. Given $f(t^n, x)$ we want to compute $f(t^{n+1}, x) = f(t^n + \Delta t, x)$.

Three instruments will be explained, each of them presenting advantages and disadvantages with respect to the other ones.

1.3.1 Introduction

Before developing the solvers for the linear advection, a mathematical introduction is needed about the transport equation.

The transport (or advection) equation is:

$$\begin{cases} \dfrac{\partial f}{\partial t} + a(t,x)\cdot\nabla_x f = 0, \quad (t,x)\in[0,T]\times\mathbb{R}^N \\ f(0,x) = f_0(x) \end{cases}$$

where $a : [0,T]\times\mathbb{R}^N \longrightarrow \mathbb{R}^N$.

We want to give results about existence and uniqueness of the solutions for such an equation. In order to do this, first of all we need to introduce the definition of characteristics.

Proposition 1.3.1 (Uniqueness of characteristic $\mathcal{X}$)**.** *If*

$$a \in \mathcal{C}^1([0,T]\times\mathbb{R}^N), \tag{1.19}$$

for all $T>0$, and there exists $k>0$ such that

$$|a(t,x)| \le k(1+|x|),\ \forall (t,x)\in\mathbb{R}_{\ge 0}\times\mathbb{R}^N, \tag{1.20}$$

then there exists a unique solution,

$$\mathcal{X}(s;t,x)\in\mathcal{C}^1([0,T]\times[0,T]\times\mathbb{R}^N),$$

for all $T>0$, of the Cauchy problem

$$\begin{cases} \dfrac{d\mathcal{X}}{dt} = a(t,\mathcal{X}(t;s,x)) \\ \mathcal{X}(s;s,x) = x \end{cases}$$

The proof can be found in any standard analysis book for ODE's systems and in this particular case in [14].

Going back to problem (1.3.1) the following theorem can be stated:

Theorem 1.3.1 (Existence and uniqueness of strong solutions)**.** *Given the advection field $a(t,x)$ satisfying* (1.19) *and* (1.20)*, and the initial data $f_0\in\mathcal{C}^1(\mathbb{R}^N)$, then there exists a unique solution of the Cauchy problem* (1.3.1)*, given by*

$$f(t,x) = f(s,\mathcal{X}(s;t,x)).$$

so, in particular,

$$f(t,x) = f_0(\mathcal{X}(0;t,x)).$$

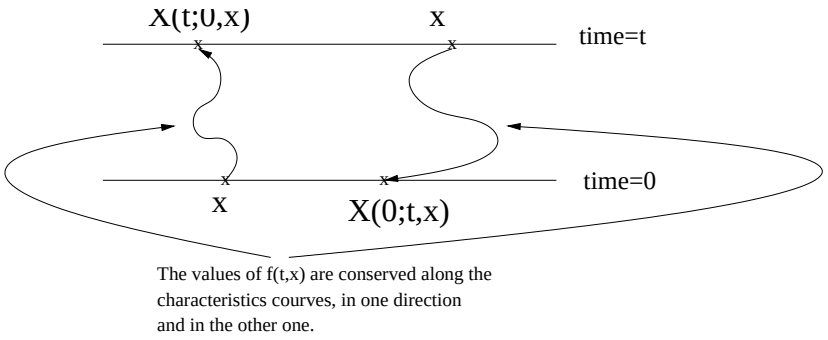


Figure 7: This is why this equation is called transport equation: the values of $f(t,x)$ are transported along the characteristics.

Proof. By hypotheses (1.19) and (1.20) Proposition 1.3.1 applies, so characteristics $\mathcal{X}(s;t,x)$ exist globally and are unique. We define

$$f(t,x) = f_0(\mathcal{X}(0;t,x)).$$

This $f(t,x)$ is a solution because of the semigroup property of $\mathcal{X}(s)$, and uniqueness comes from proving that if $f(t,x)$ is a solution, then $f(t,x)$ is constant over characteristics, that is,

$$\frac{d}{ds} f(s, \mathcal{X}(s;t,x)) = 0$$

and thus, it must be of the form we have written (see [14] for details).

Linear advection

The computation of characteristics in case $a(t,x)$ is a real constant is straighforward:

$$\begin{cases} \dfrac{d\mathcal{X}}{ds} = a \\ \mathcal{X}(t) = x \end{cases}$$

gives

$$\mathcal{X}(s;t,x) = x + a(s-t)$$

so that the solution of the initial value problem

$$\begin{cases} \dfrac{\partial f}{\partial t} + a\dfrac{\partial f}{\partial x} = 0 \\ f(t_0,x) = f_0(x) \end{cases}$$

is

$$f(t,x) = f_0(x - a(t-t_0))$$

This is the only result that we need to implement in all the routines concerning advection: the different reconstructions of f_0 will give different properties, like mass conservation or total variation control.

Mass conservation

It is trivial to remark that linear advection is mass-conservative:

$$\int_{\mathbb{R}} f(t,x)dx = \int_{\mathbb{R}} f_0(x-at)dx = \int_{\mathbb{R}} f_0(x)dx = M$$

That is why we would like numerical methods to preserve this property.

1.3.2 Direct semi-lagrangian method

This method is based on straightforwardly following backwards the characteristics.

Knowing $f(t^n, x_i)$, we want to compute $f(t^{n+1}, x_i)$. Following the characteristics, we know that

$$f(t^{n+1}, x_i) = f(t^n, x_i - a\Delta t)$$

which means that we have to reconstruct the values of $f(t^n, \cdot)$ (of which we do not dispose), by some interpolation method. Of course, the choice of which reconstruction influences the properties of the solution; Lagrange interpolation, for instance, may induce spurious oscillations but have very low computational cost.

This method is very easy to implement, but it has an important disadvantage: it is not conservative.

The α parameter

The parameter

$$\alpha = a\frac{\Delta t}{\Delta x}$$

describes how close to the grid points we are interpolating: from $f(t^{n+1}, x_i) = f(t^n, x_i - a\Delta t)$,

$$x_i - a\Delta t = x_i - \alpha\Delta x$$

which means that the nearer is α to an integer number, the better is hoped the interpolation to be.

1.3.3 The Flux Balance Method

FBM (Flux Balance Method) is used in [42] to construct a conservative method. We already know that

$$f(t+\Delta t, x) = f(t, x - a\Delta t).$$

Now, let us integrate over an interval $[b_1, b_2]$, to get

$$\begin{aligned}\int_{b_1}^{b_2} f(t+\Delta t,\xi)d\xi &= \int_{b_1}^{b_2} f(t,\xi-a\Delta t)d\xi \\ &= \int_{b_1-a\Delta t}^{b_2-a\Delta t} f(t,\xi)d\xi \\ &= \int_{b_1-a\Delta t}^{b_1} f(t,\xi)d\xi + \int_{b_1}^{b_2} f(t,\xi)d\xi - \int_{b_2-a\Delta t}^{b_2} f(t,\xi)d\xi.\end{aligned}$$

If we use as notation

$$\Phi(t,x) = \int_{x-a\Delta t}^{x} f(t,\xi)d\xi,$$

we get

$$\int_{b_1}^{b_2} f(t+\Delta t,\xi)d\xi = \int_{b_1}^{b_2} f(t,\xi)d\xi + \Phi(t,b_1) - \Phi(t,b_2),$$

and dividing by $\Delta = b_2 - b_1$

$$\frac{\int_{b_1}^{b_2} f(t+\Delta t,\xi)d\xi}{\Delta} = \frac{\int_{b_1}^{b_2} f(t,\xi)d\xi}{\Delta} + \frac{\Phi(t,b_1)-\Phi(t,b_2)}{\Delta},$$

which means

$$\bar{f}_{(b_1,b_2)}(t+\Delta t) = \bar{f}_{(b_1,b_2)}(t) + \frac{\Phi(t,b_1)-\Phi(t,b_2)}{\Delta}.$$

This is the local description of mass conservation.
Call $F(t,\cdot)$ the primitive of $f(t,\cdot)$, the numerical method we get is the following:

$$f_i^{n+1} = f_i^n + \frac{\Phi^n(x_{i-\frac{1}{2}}) - \Phi^n(x_{i+\frac{1}{2}})}{\Delta x},$$

where

$$\Phi^n(x_{i-\frac{1}{2}}) = \int_{x_{i-\frac{1}{2}}-a\Delta t}^{x_{i-\frac{1}{2}}} f(t^n,\xi)d\xi = F(x_{i-\frac{1}{2}}) - F(x_{i-\frac{1}{2}} - a\Delta t).$$

Now we need some method to reconstruct what we do not have: either directly $\Phi^n(x_{i-\frac{1}{2}})$ or $F(_{i-\frac{1}{2}}-a\Delta t)$ (if we can compute $F(x_{i-\frac{1}{2}})$).

Reconstruction of $F(t,\cdot)$

We can compute $F(t,x_{i+\frac{1}{2}})$ by putting $F(t,x_{i+\frac{1}{2}}) = \sum_{j=0}^{i} f(t,x_i)\Delta x$ and reconstruct the values $F(t,x_{i-\frac{1}{2}} - a\Delta t)$ and $F(t,x_{i+\frac{1}{2}} - a\Delta t)$ by using some interpolation method.

The problems we could find by using this method is that the positivity is not guaranteed and the oscillations could be uncontrolled, especially by using Lagrange interpolation.

1.3.4 PFC-3 method

In order to assure the positivity and the control of the oscillations in the reconstruction, this method has been introduced in [42, page 70-72]. It is a third order method, and the flux $\Phi_{i+\frac{1}{2}}$ is directly computed: if the wind propagation velocity is positive, let j be the index of the cell which contains $x_{i+\frac{1}{2}} - a\Delta t$, let $\alpha_i = x_{j+\frac{1}{2}} - (x_{i+\frac{1}{2}} - a\Delta t)$, compute slope correctors ϵ_i^+ and ϵ_i^-, then

$$\begin{aligned}\Phi_{i+\frac{1}{2}} &= \Delta x \sum_{k=j+1}^{i} f_k + \alpha_i \left[f_j + \frac{\epsilon_i^+}{6}\left(1 - \frac{\alpha_i}{\Delta x}\right)\left(2 - \frac{\alpha_i}{\Delta x}\right)(f_{j+1} - f_j) \right. \\ &+ \left. \frac{\epsilon_i^-}{6}\left(1 - \frac{\alpha_i}{\Delta x}\left(1 + \frac{\alpha_i}{\Delta x}\right)(f_j - f_{j-1})\right)\right].\end{aligned}$$

Otherwise, if the wind propagation velocity is negative, let $\alpha_i = x_{j-\frac{1}{2}} - (x_{i+\frac{1}{2}} - a\Delta t)$, then

$$\begin{aligned}\Phi_{i+\frac{1}{2}} &= \Delta x \sum_{k=i+1}^{j-1} f_k + \alpha_i \left[f_j - \frac{\epsilon_i^+}{6}\left(1 - \frac{\alpha_i}{\Delta x}\right)\left(1 + \frac{\alpha_i}{\Delta x}\right)(f_{j+1} - f_j) \right. \\ &- \left. \frac{\epsilon_i^-}{6}(2 + \frac{\alpha_i}{\Delta x}(1 + \frac{\alpha_i}{\Delta x})(f_j - f_{j-1})\right].\end{aligned}$$

Correctors ϵ_i^+ and ϵ_i^- are defined

$$\epsilon_i^+ = \begin{cases} \min\left(1; 2\frac{f_i}{f_{i+1}-f_i}\right) & f_{i+1} > f_i \\ \min\left(1; -2\frac{f_\infty - f_i}{f_{i+1}-f_i}\right) & f_{i+1} < f_i \end{cases}$$

$$\epsilon_i^- = \begin{cases} \min\left(1; 2\frac{f_\infty - f_i}{f_i - f_{i-1}}\right) & f_i > f_{i-1} \\ \min\left(1; -2\frac{f_i}{f_i - f_{i-1}}\right) & f_i < f_{i-1} \end{cases}$$

where $f_\infty = \sup_{i=0}^{N-1} f_i$.

1.4 Collisions

In this section the two fundamental steps for solving the collision steps are developed; basically we mean to integrate a two variable function

$$\begin{aligned} f : \mathbb{R}^2 &\rightarrow \mathbb{R} \\ (x, y) &\mapsto f(x, y) \end{aligned}$$

either on a semicircle, following the undergoing segment

$$\int_{-R}^{R} f(x, \sqrt{R^2 - x^2})dx,$$

or on a circle, in the Riemann way:

$$\int_0^{2\pi} f\left[R\cos(x), R\sin(x)\right] dx.$$

1.4.1 Integration on a segment following a semicircle

Numerical Scheme: Collision Step.- In order to solve the collision step, we need to compute some integrals along semicircles of radius $\gamma_0(k)$, $\gamma_+(k)$ and $\gamma_-(k)$ in the (k_1, k_{23})-space. Figure 8 explains two different ways in which we can perform it. We may first use a direct linear interpolation between the closest points in the cartesian grid as specified in Figure 8 left. The integration rule to compute the final approximation of (3.12) is coherently chosen in terms of accuracy as the trapezoidal rule.

On the other hand, we can choose a WENO-4,3 interpolation along the $k_1 = (k_1)_j$ line where the point lies as in Figure 8 right. This interpolation is performed on the stencil $((k_1)_j, (k_{23})_{l-1}), ..., ((k_1)_j, (k_{23})_{l+2})$. If not enough points are available (i.e. $l = 0$ or $l \geq N_{k_{23}} - 2$), we use Lagrange-3 on the proper stencil. Coherently with a degree-2 polynomial interpolation, as integration rule for (3.12), we choose Simpson's rule. A similar interpolation procedure was used in [93] to cope with analogous problems in a computational fluid dynamics problem.

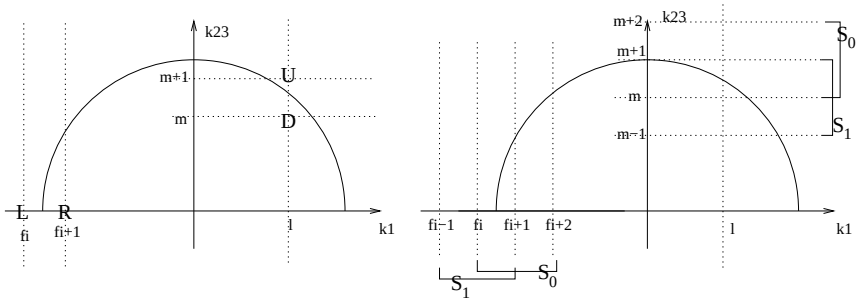


Figure 8: The integration on the interval $[-\sqrt{\gamma}, \sqrt{\gamma}]$. Left: the needed values are obtained through a linear interpolation on the two closest points lying either on the $k_{23} = 0$ line (for the first and last point) or on the $k_1 = l$ line (for the other points). Right: the needed values are obtained through PWENO-4,3 interpolation on the closest points lying either on the $k_{23} = 0$ line (for the first and last point) or on the $k_1 = l$ line (for the other points).

1.4.2 Riemann integration along a circle

We want to compute the integrals of $f(k_1, k_2)$ along the circles of radius $\gamma_{p,p'}$ in the (k_1, k_2)-space.

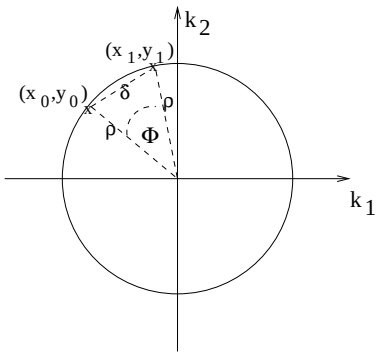


Figure 9: The computation of the angle Φ starting from two points on the circle.

The idea is the exploitation of a along-the-line interpolation each time we cross either a $k_1 = (k_1)_l$-line or a $k_2 = (k_2)_m$-line.

The angle between two points, like in Figure 1.4.2, is expressed by

$$\Phi \quad = \quad 2\arcsin\left[\frac{\sqrt{(x_1-x_0)^2+(y_1-y_0)^2}}{2\rho}\right].$$

The algorithm

Step 1. We check that the algorithm is going to perform for the integration of $f_p(x, k_1, k_2)$ along a circle of radius ρ in the (k_1, k_2)-plane.

Step 0(i). if

$$\rho \geq (k_1)_{N_{k_1}-1}$$

return 0.

Step 0(ii). if (look at Figure 1.4.2)

$$\rho < \frac{\Delta k_1}{2}$$

we may face four different cases:

Step 0(ii)a. N_{k_1} is odd, N_{k_2} is odd: in this case we have

$$f_p(0,0) = f_{p,i,\left[\frac{N_{k_1}}{2}\right],\left[\frac{N_{k_2}}{2}\right]},$$

then

$$\int_0^{2\pi} f\left(\rho\cos(\Phi), \rho\sin(\Phi)\right) d\Phi \approx 2\pi\rho f_p(0,0).$$

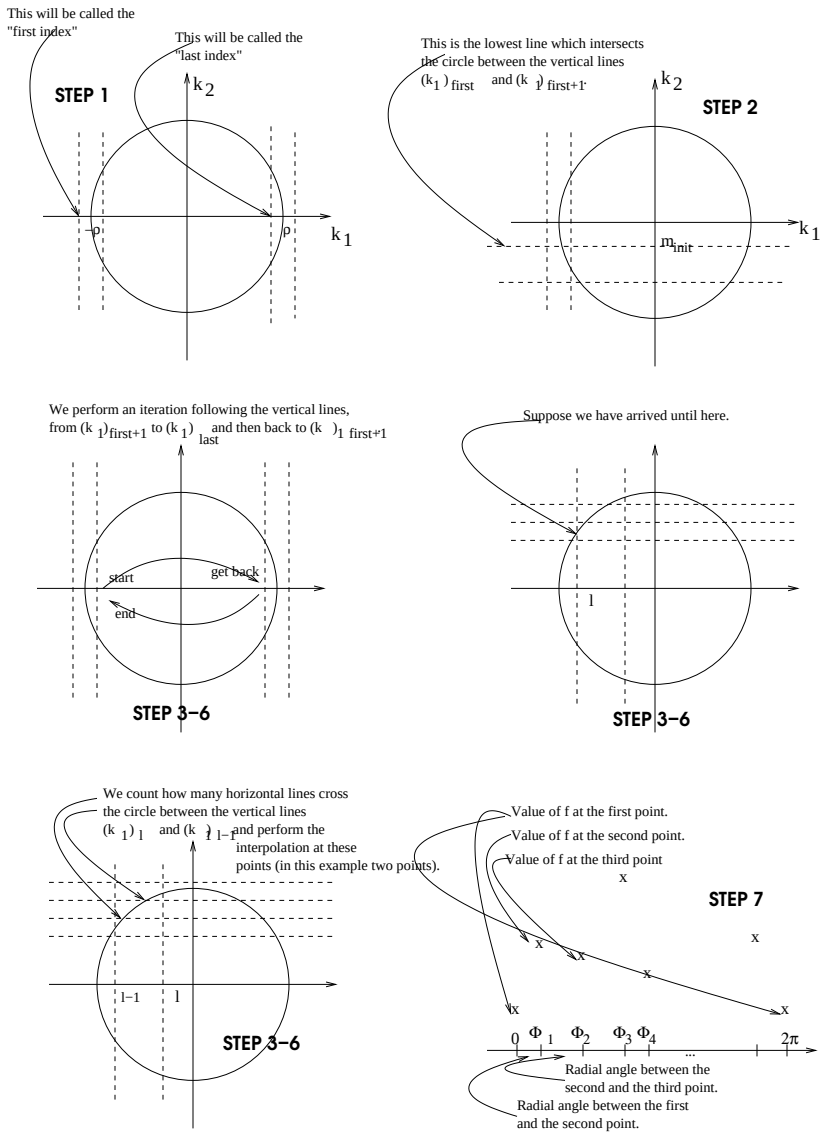


Figure 10: Resumé of the algorithm for the interpolation of f along a circle of radius ρ.

Step 0(ii)b. N_{k_1} is even, N_{k_2} is odd: in this case we have

$$f_p(0,0) = \text{linear interp. between } f_{p,i,\frac{N_{k_1}}{2}-1,\left[\frac{N_{k_2}}{2}\right]} \text{ and } f_{p,i,\frac{N_{k_1}}{2},\left[\frac{N_{k_2}}{2}\right]}$$

then

$$\int_0^{2\pi} f\left(\rho\cos(\Phi),\rho\sin(\Phi)\right) d\Phi \approx 2\pi\rho f_p(0,0).$$

Step 0(ii)c. N_{k_1} is odd, N_{k_2} is even: in this case we have

$$f_p(0,0) = \text{linear interp. between } f_{p,i,\left[\frac{N_{k_1}}{2}\right],\frac{N_{k_2}}{2}-1} \text{ and } f_{p,i,\left[\frac{N_{k_1}}{2}\right],\frac{N_{k_2}}{2}}$$

then

$$\int_0^{2\pi} f\left(\rho\cos(\Phi),\rho\sin(\Phi)\right) d\Phi \approx 2\pi\rho f_p(0,0).$$

Step 0(ii)d. N_{k_1} is even, N_{k_2} is even: in this case we evaluate

$$\begin{array}{l} f_p(0,0) \\ = \quad \dfrac{f_{p,i,\frac{N_{k_1}}{2}-1,\frac{N_{k_2}}{2}-1} + f_{p,i,\frac{N_{k_1}}{2}-1,\frac{N_{k_2}}{2}} + f_{p,i,\frac{N_{k_1}}{2},\frac{N_{k_2}}{2}-1} + f_{p,i,\frac{N_{k_1}}{2},\frac{N_{k_2}}{2}}}{4} \end{array}$$

then

$$\int_0^{2\pi} f\left(\rho\cos(\Phi),\rho\sin(\Phi)\right) d\Phi \approx 2\pi\rho f_p(0,0).$$

Step 1. compute the first and the last index, called l_{first} and l_{last}

Step 2. we have to look for the first index m such that the line $k_2 = (k_2)_m$ intersects the circle between the lines $k_1 = (k_1)_{l_{first}}$ and $k_1 = (k_1)_{l_{first}+1}$, i.e.

$$m_{init} = \min\left\{m \text{ such that } (k_1)_{l_{first}} < \sqrt{\rho^2 - (k_2)_m^2} < (k_1)_{l_{first}+1}\right\}.$$

Now, initialize

$$l = l_{first} + 1, \qquad m = m_{init}.$$

Step 3. perform the iteration:

$$\text{for}(; l \le l_{last}; l = l + 1)$$

Step 3(i). compute $M(l) \in \{0, ..., N_{k_2} - 1\}$, the index such that the line $k_1 = (k_1)_l$ intersects the circle between the lines $k_2 = (k_2)_{M(l)}$ and $k_2 = (k_2)_{M(l)+1}$; **remark:** there are two solutions for $M(l)$, because, for

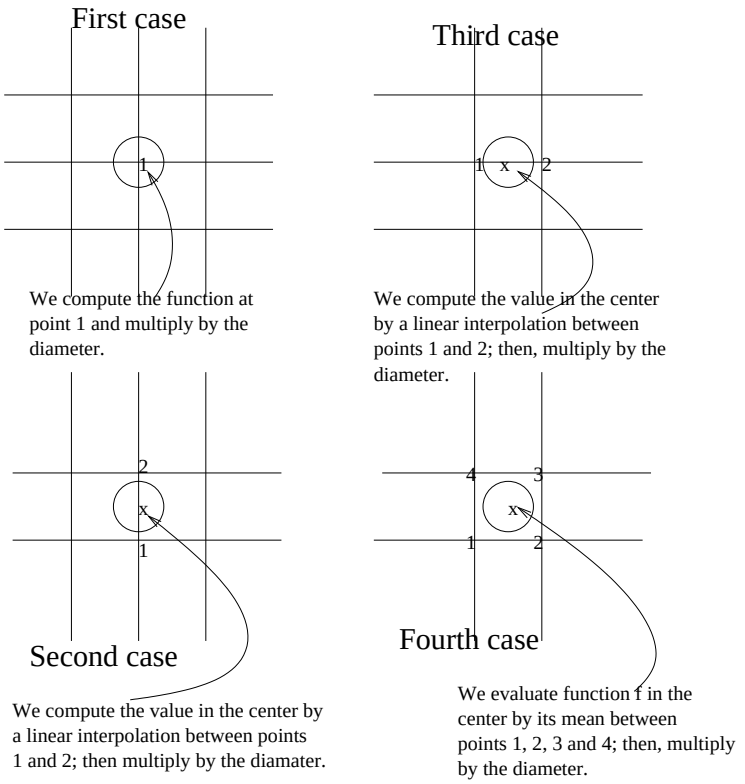


Figure 11: The particular case when the radius is very small.

symmetry reasons, there will be two pairs of horizontal lines containing the point

$$\left((k_1)_l, \sqrt{\rho^2 - (k_1)_l^2}\right),$$

one pair in the positive half-plane, and one in the negative half-plane. We shall denote by $M(l)$ the upper index, and by $\bar{M}(l)$ the lower index.

Step 3(ii). compute the difference

$$\Delta m = M(l) - m + 1$$

Step 3(iii). if $\Delta m > 0$

3(iii)a

$$\text{for}(; m \leq M(l); m = m + 1)$$

perform interpolation to get the value at point

$$\begin{cases} \left(\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l > 0 \\ \left(-\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l \leq 0 \end{cases}$$

Step 3(iv). else if $\Delta m < 0$

Step 3(iv)a.

$$\text{for}(m = m - 1; m > M(l); m = m - 1)$$

perform interpolation to get the value at point

$$\begin{cases} \left(\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l > 0 \\ \left(-\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l \leq 0 \end{cases}$$

Step 3(iv)b. set

$$m = m + 1$$

Step 3(v). perform interpolation to reconstruct the value at point

$$\left((k_1)_l, \sqrt{\rho^2 - (k_1)_l^2}\right)$$

Step 4. set

$$l = l_{last}$$

Step 5. set

$$m = M(l)$$

Step 6. perform the iteration (like before)

$$\text{for}(; l \geq l_{first} + 1; l = l - 1)$$

Step 6(i). compute the index $\bar{M}(l)$

Step 6(ii). compute

$$\Delta m = \bar{M}(l) - m + 1.$$

Step 6(iii). If $\Delta m > 0$ then

Step 6(iii)a.

$$\text{for}(; m \leq M(l); m = m + 1)$$

perform interpolation to get the value at point

$$\begin{cases} \left(\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l \geq 0 \\ \left(-\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l < 0 \end{cases}$$

Step 6(iv). else if $\Delta m < 0$

Step 6(iv)a.

$$\text{for}(m = m - 1; m > M(l); m = m - 1)$$

perform interpolation to get the value at point

$$\left(\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right)$$

Step 6(iv)b.

$$m = m + 1$$

Step 6(v). interpolate to reconstruct the value at point

$$\begin{cases} \left(\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l \geq 0 \\ \left(-\sqrt{\rho^2 - (k_2)_m^2}, (k_2)_m\right) & \text{if } (k_1)_l < 0 \end{cases}$$

Step 7. perform Riemann integration and return the computed value.

1.5 1D stationary-state Schrödinger solver

In this section, we solve the eigenvalue problem to the Schrödinger equation: given potential $V(z)$, find $\{\epsilon^p, \chi^p\}$ such that the following problems have solutions

$$-\frac{\hbar^2}{2m_e}\frac{d}{dz}\left[\frac{1}{m_*}\frac{d\chi^p}{dz}\right] - q\left(V + V_c\right)\chi^p = \epsilon^p\chi^p$$
$$\text{with } \left\{\chi^p\right\}_p \subseteq H_o^1(0, l_z) \text{ orthonormal basis}$$

where $\hbar$ is the reduced Planck constant, m_e is the electron mass, $m_*(z)$ is the effective electron mass (in pure number units, i.e. its ratio with respect to the electron mass), V_c is a fixed external potential (where "c" stands for "confining"): in the concrete case of the MOSFET described in Chapter 5, V_c is a built-in potential drop between the Si-layer and the SiO_2-layer which confines the carriers inside the Si-layer; $\epsilon^p[V](x)$ are noted $\epsilon_p^{pot}(t, x)$ and are referred to as band-potential energies; $\chi^p[V]$ are noted $\chi_p(t, x, z)$ and give information on how the free electrons are distributed along the z-dimension: band densities are a mixed quantum-classical state, where the classical part is given by the occupation numbers $\rho_p(x)$ and the quantum part by Schrödinger eigenfunctions $\chi_p(x, z)$

$$N_p(x, z) \quad = \quad \rho_p(x)|\chi_p(x, z)|^2.$$

1.5.1 1D Schrödinger equation discretization via Finite Differences

We propose a solver based on discretization via Finite Differences and the use of a LAPACK routine called DSTEQR. Finite differences for the two z-derivatives have to be taken in alternate direction in order to rescue the classical three-point discretization for the Laplacian.

We solve the Schrödinger equation in the interval $z \in [0, l_z]$, given a potential $V : [0, l_z] \longrightarrow \mathbb{R}$:

$$-\frac{C^{S,1}}{2}\frac{d}{dz}\left[\frac{1}{m(z)}\frac{d\chi^p}{dz}\right] + C^{S,2}V\chi^p = \epsilon^p\chi^p$$
$$\chi^p \in H_o^1(0, l_z), \qquad \langle\chi^p, \chi^q\rangle = \int_0^{l_z} \chi^p\chi^q dz = \delta_{p,q}, \qquad (1.21)$$

where we have included the confining potential into potential V, and we have grouped all the physical constants ($\hbar$, m_e, q) into one constant per term: when solving the concrete case of the MOSFET in Chapter 5, equations are reduced to adimensionalized units, so dimensionless parameters appear, and we have followed the same notations.
Meshing the z-dimension with a uniform grid

$$\{z_j = j\Delta z\}_{j=0,\ldots,N_z-1}, \qquad \Delta z = \frac{l_z}{N_z - 1},$$

and discretizing the z-derivatives by alternate finite differences, (1.21) becomes

$$-\frac{C^{S,1}}{2}\frac{\frac{1}{m_j}\chi^p_{j+1}-\left(\frac{1}{m_j}+\frac{1}{m_{j-1}}\right)\chi^p_j+\frac{1}{m_{j-1}}\chi^p_{j-1}}{\Delta z^2}+C^{S,2}V\chi^p \;=\; \epsilon^p\chi^p.$$

Thanks to the boundary conditions, which fix the bottom point and the top point, our system is $(N_z-2)\times(N_z-2)$.

If we let (for $(r,s)\in\{0,...,N_Z-1\}\times\{0,...,N_Z-1\}$))

$$\begin{aligned}(M_1)_{r,s} &= -\frac{C^{S,1}}{2\Delta z^2}\left[\frac{1}{m_r}\delta_{r,s+1}-\left(\frac{1}{m_r}+\frac{1}{m_{r+1}}\right)\delta_{r,s}+\frac{1}{m_s}\delta_{r+1,s}\right]\\ M_2 &= C_2^S\begin{bmatrix}V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_{N_z-2}\end{bmatrix}\end{aligned}$$

the matrix we want to diagonalise is

$$M=M_1+M_2,$$

i.e. we have to compute a certain number of eigenvalues ϵ^p of M, the lowest ones (physically representing the most occupied energy bands) and their relative eigenvectors $(\chi^p_j)_j$.

The discretized system reads

$$M\begin{bmatrix}\chi^P_1\\ \chi^P_2\\ \vdots\\ \chi^P_{N_z-3}\\ \chi^P_{N_z-2}\end{bmatrix}=\epsilon^p\begin{bmatrix}\chi^p_1\\ \chi^p_2\\ \vdots\\ \chi^p_{N_z-3}\\ \chi^p_{N_z-2}\end{bmatrix}. \tag{1.22}$$

1.6 "Generalized" 1D Poisson equations

We call "generalized" Poisson problem the following system

$$-\frac{d}{dz}\left[\varepsilon(z)\frac{dV}{dz}(z)\right]+\mathcal{C}\int\mathcal{A}(z,\zeta)V(\zeta)d\zeta=\mathcal{B}(z), \tag{1.23}$$

plus Dirichlet, Neumann or Robin boundary conditions: for $z\in\{0,l_z\}$

$$\begin{cases}V(z)=\bar{V}(z) & \text{Dirichlet}\\ \frac{dV}{dn}=g & \text{Neumann}\\ \frac{dV}{dn}+\alpha(z)(V(z)-\bar{V}(z))=g & \text{Robin}\end{cases}.$$

Such kind of system appears in the modelling of the MOSFET in Chapter 5 when we have to compute the border potential respecting the electrical neutrality. Here we have already grouped all the physical constant, because in the concrete case in which we shall use this scheme, equations are reduced to adimensionalized units. Here, $\varepsilon(z)$ is the relative dielectric permittivity along the z-slice, $\mathcal{C}$ is a constant which has contributions from many physical constants, $\mathcal{A}(z,\zeta)$ is a kernel which includes the Gâteaux-derivative of the density $N[V]$ which does not explicitly appear in these equations being included in term $\mathcal{B}$.

1.6.1 Discretization

We propose a solver based on discretization via alternate Finite Differences for the two z-derivatives (in order to rescue the classical centered three-points scheme for the Laplacian), and trapezoid rule for the integration. The resulting matrix is solved by means of a LAPACK routine called DGESV.

1.6.2 Discretized system

z-dimension is supposed to be uniformly meshes. System (1.23) is therefore discretized:

$$\begin{aligned} & -\frac{\varepsilon_j}{\Delta z^2}V_{j+1} + \frac{\varepsilon_j+\varepsilon_{j-1}}{\Delta z^2}V_j - \frac{\varepsilon_{j-1}}{\Delta z^2}V_{j-1} \\ & +\mathcal{C}\Delta z\left[\sum_{l=1}^{N-2}\mathcal{A}_{j,l}V_l + \frac{1}{2}\mathcal{A}_{j,0}V_0 + \frac{1}{2}\mathcal{A}_{j,N-1}V_{N-1}\right] \\ = \;& \mathcal{B}_j, \qquad \text{for } j = 1,...,N-2 \end{aligned} \tag{1.24}$$

completed by boundary conditions for $j = 0$

$$\begin{cases} V_0 = \bar{V}_0 & \text{Dirichlet} \\ V_0 - V_1 = \Delta z g_0 & \text{Neumann} \\ (\Delta z\alpha_0 + 1)V_0 - V_1 = \alpha_0\Delta z\bar{V}_0 + \Delta z g_0 & \text{Robin} \end{cases} \tag{1.25}$$

and $j = N-1$

$$\begin{cases} V_{N-1} = \bar{V}_{N-1} & \text{Dirichlet} \\ V_{N-1} - V_{N-2} = \Delta z g_{N-1} & \text{Neumann} \\ (1+\alpha_{N-1}\Delta z)V_{N-1} - V_{N-2} = \alpha_{N-1}\Delta z\bar{V}_{N-1} + \Delta z g_{l_z} & \text{Robin} \end{cases} \tag{1.26}$$

The integration is performed through standard trapezoids rule

$$\begin{aligned} & \left[\int \mathcal{A}(z,\zeta)V(\zeta)d\zeta\right]_j \\ \approx \;& \Delta z\left[\frac{1}{2}\mathcal{A}_{j,0}V_{i,0} + \sum_{l=1}^{N_z-2}\mathcal{A}_{j,l}V_l + \frac{1}{2}\mathcal{A}_{j,N_z-1}V_{N_z-1}\right]. \end{aligned} \tag{1.27}$$

Putting (1.24)-(1.27) together, we obtain a $N_z \times N_z$ system with extra-diagonal terms due to the presence of (1.27) which introduces non-local effects. The system can be solved by any suitable linear system solver: here we have chosen a LAPACK routine called DGESV, which has proven to be fast and robust. Other possible choices would be for instance Gauss-Siegel iteration or Successive OverRelaxation method.

1.7 "Generalized" 2D Poisson equations

We give now the method for solving the 2D "generalized" Poisson system

$$-\mathrm{div}_{x,z}\left[\varepsilon(x,z)\nabla_{x,z}V(x,z)\right] + \mathcal{C}\int \mathcal{A}(x,z,\zeta)V(x,\zeta)d\zeta = \mathcal{B}(x,z).$$

Dirichlet, Neumann or Robin boundary conditions

Such systems are solved many times per time step during the numerical simulation of the MOSFET described in Chapter 5: potential is computed through a Newton method, and the evaluation of the self-consistent electrostatic field via the solution of a "generalized" Poisson equation is required at each iteration. $\varepsilon(x,z)$ is the relative dielectric permittivity of the material (in the concrete case of the MOSFET, the Si and SiO_2 permittivity), $\mathcal{C}$ is a parameter which has the contribution of many physical constants (and rescalings), the system being reduced to adimensionalized units. $\mathcal{B}$ usually has the classical contribution of the total electron density plus an extra term coming from the non-local effects due to the presence of the Gâteaux-derivative of the density $N[V]$.

1.7.1 Discretization

We propose a solver based on discretization via Finite Differences of the derivatives and trapezoid rule for the integration. The resulting matrix is then solved via a LAPACK routine called DGESV. Meshes are assumed regular.

The "Laplacian"

First of all we remark that the derivatives of the divergence and the derivatives of the gradient must be in alternate directions, because in the case of $\epsilon = 1$ we want to recover the standard centered three-points discretization of the Laplacian.

We choose to go forward as for the gradient and backwards as for the

divergence:

$$(\nabla V)_{i,j} = \left[\left(\frac{\partial V}{\partial x}\right)_{i,j}, \left(\frac{\partial V}{\partial z}\right)_{i,j}\right] = \left(\frac{V_{i+1,j}-V_{i,j}}{\Delta x}, \frac{V_{i,j+1}-V_{i,j}}{\Delta z}\right)$$

$$\mathrm{div}[(F^1,F^2)]_{i,j} = \frac{F^1_{i,j}-F^1_{i-1,j}}{\Delta x} + \frac{F^2_{i,j}-F^2_{i,j-1}}{\Delta z}.$$

So, once we put everything together

$$\begin{aligned}
&\left[\mathrm{div}_{x,z}\left(\varepsilon(x,z)\nabla_{x,z}V(x,z)\right)\right]_{i,j} \\
\approx\ & \frac{\varepsilon_{i-1,j}}{\Delta x^2}V_{i-1,j} + \frac{\varepsilon_{i,j-1}}{\Delta z^2}V_{i,j-1} - \left(\frac{\varepsilon_{i,j}+\varepsilon_{i-1,j}}{\Delta x^2} + \frac{\varepsilon_{i,j}+\varepsilon_{i,j-1}}{\Delta z^2}\right)V_{i,j} \\
&+\frac{\varepsilon_{i,j}}{\Delta z^2}V_{i,j+1} + \frac{\varepsilon_{i,j}}{\Delta x^2}V_{i+1,j} \qquad (1.28)
\end{aligned}$$

The integral

The integration is performed through standard trapezoids rule

$$\begin{aligned}
&\left[\int \mathcal{A}(x,z,\zeta)V(x,\zeta)d\zeta\right]_{i,j} \\
\approx\ & \Delta z\left[\frac{1}{2}\mathcal{A}_{i,j,0}V_{i,0} + \sum_{l=1}^{N_z-2}\mathcal{A}_{i,j,l}V_{i,l} + \frac{1}{2}\mathcal{A}_{i,j,N_z-1}V_{i,N_z-1}\right]. \qquad (1.29)
\end{aligned}$$

Boundary conditions

If Dirichlet condition is set, we impose

$$V_{i,j} = \bar{V}_{i,j}.$$

If Neumann condition is set

$$\frac{\partial V}{\partial n} = g(x,z),$$

we discretize the normal derivative by finite differences, either forward or backward:

$(x,0)$	(L_x,z)	(x,l_z)	$(0,z)$
$\frac{\partial V}{\partial n} = -\frac{\partial V}{\partial z}$ $\frac{\partial V}{\partial n} \approx \frac{V_{i,0}-V_{i,1}}{\Delta z}$	$\frac{\partial V}{\partial n} = \frac{\partial V}{\partial x}$ $\frac{\partial V}{\partial n} \approx \frac{V_{N_x-1,j}-V_{N_x-2,j}}{\Delta x}$	$\frac{\partial V}{\partial n} = \frac{\partial V}{\partial z}$ $\frac{\partial V}{\partial n} \approx \frac{V_{i,N_z-1}-V_{i,N_z-2}}{\Delta z}$	$\frac{\partial V}{\partial n} = -\frac{\partial V}{\partial x}$ $\frac{\partial V}{\partial n} \approx \frac{V_{0,j}-V_{1,j}}{\Delta x}$

therefore

$$\frac{\partial V}{\partial n} = g(x,z)$$

$\approx$

$(x,0)$	(L_x,z)
$\frac{V_{i,0}-V_{i,1}}{\Delta z} = g_{i,0}$ $\frac{1}{\Delta z}V_{i,0} - \frac{1}{\Delta z}V_{i,1} = g_{i,0}$	$\frac{V_{N_x-1,j}-V_{N_x-2,j}}{\Delta x} = g_{N_x-1,j}$ $\frac{1}{\Delta x}V_{N_x-1,j} - \frac{1}{\Delta x}V_{N_x-2,j} = g_{N_x-1,j}$
(x,l_z)	$(0,z)$
$\frac{V_{i,N_z-1}-V_{i,N_z-2}}{\Delta z} = g_{i,N_z-1}$ $\frac{1}{\Delta z}V_{i,N_z-1} - \frac{1}{\Delta z}V_{i,N_z-2} = g_{i,N_z-1}$	$\frac{V_{0,j}-V_{1,j}}{\Delta x} = g_{0,j}$ $\frac{1}{\Delta x}V_{0,j} - \frac{1}{\Delta x}V_{1,j} = g_{0,j}$

If Robin condition

$$\frac{\partial V}{\partial n} + \alpha(x,z)(V - \bar{V}) = 0$$

is set, we discretize it the following way:

$(x,0)$	(L_x,z)	(x,l_z)	$(0,z)$
$\frac{\partial V}{\partial n} = -\frac{\partial V}{\partial z}$ $\frac{\partial V}{\partial n} \approx \frac{V_{i,0}-V_{i,1}}{\Delta z}$	$\frac{\partial V}{\partial n} = \frac{\partial V}{\partial x}$ $\frac{\partial V}{\Delta n} \approx \frac{V_{N_x-1,j}-V_{N_x-2,j}}{\Delta x}$	$\frac{\partial V}{\partial n} = \frac{\partial V}{\partial z}$ $\frac{\partial V}{\Delta n} \approx \frac{V_{i,N_z-1}-V_{i,N_z-2}}{\Delta z}$	$\frac{\partial V}{\partial n} = -\frac{\partial V}{\partial x}$ $\frac{\partial V}{\Delta n} \approx \frac{V_{0,j}-V_{1,j}}{\Delta x}$

therefore

$$\frac{\partial V}{\partial n} + \alpha(x,z)(V - \bar{V}) = g(x,z)$$

$\approx$

$(x,0)$
$\frac{V_{i,0}-V_{i,1}}{\Delta z} + \alpha_{i,0}(V_{i,0} - \bar{V}_{i,0}) = g_{i,0}$ $\frac{1}{\Delta z}V_{i,0} - \frac{1}{\Delta z}V_{i,1} + \alpha_{i,0}V_{i,0} - \alpha_{i,0}\bar{V}_{i,0} = g_{i,0}$ $\left(\frac{1}{\Delta z} + \alpha_{i,0}\right)V_{i,0} - \frac{1}{\Delta z}V_{i,1} = \alpha_{i,0}\bar{V}_{i,0} + g_{i,0}$

(L_x,z)
$\frac{V_{N_x-1,j}-V_{N_x-2,j}}{\Delta x} + \alpha_{N_x-1,j}(V_{N_x-1,j} - \bar{V}_{N_x-1,j}) = g_{N_x-1,j}$ $\frac{1}{\Delta x}V_{N_x-1,j} - \frac{1}{\Delta x}V_{N_x-2,j} + \alpha_{N_x-1,j}V_{N_x-1,j} - \alpha_{N_x-1,j}\bar{V}_{N_x-1,j} = g_{N_x-1,j}$ $\left(\frac{1}{\Delta x} + \alpha_{N_x-1,j}\right)V_{N_x-1,j} - \frac{1}{\Delta x}V_{N_x-2,j} = \alpha_{N_x-1,j}\bar{V}_{N_x-1,j} + g_{N_x-1,j}$

(x,l_z)
$\frac{V_{i,N_z-1}-V_{i,N_z-2}}{\Delta z} + \alpha_{i,N_z-1}(V_{i,N_z-1} - \bar{V}_{i,N_z-1}) = g_{i,N_z-1}$ $\frac{1}{\Delta z}V_{i,N_z-1} - \frac{1}{\Delta z}V_{i,N_z-2} + \alpha_{i,N_z-1}V_{i,N_z-1} - \alpha_{i,N_z-1}\bar{V}_{i,N_z-1} = g_{i,N_z-1}$ $\left(\frac{1}{\Delta z} + \alpha_{i,N_z-1}\right)V_{i,N_z-1} - \frac{1}{\Delta z}V_{i,N_z-2} = \alpha_{i,N_z-1}\bar{V}_{i,N_z-1} + g_{i,N_z-1}$

$(0,z)$
$\frac{V_{0,j}-V_{1,j}}{\Delta x} + \alpha_{0,j}(V_{0,j} - \bar{V}_{0,j}) = g_{0,j}$ $\frac{1}{\Delta x}V_{0,j} - \frac{1}{\Delta x}V_{1,j} + \alpha_{0,j}V_{0,j} - \alpha_{0,j}\bar{V}_{0,j} = g_{0,j}$ $\left(\frac{1}{\Delta x} + \alpha_{0,j}\right)V_{0,j} - \frac{1}{\Delta x}V_{1,j} = \alpha_{0,j}\bar{V}_{0,j} + g_{0,j}$

Still we have to decide what to do with the four corner points

$$(0,0), (L_x,0), (L_x,l_z), (0,l_z).$$

For instance we could make them be the mean between the two nearest boundary points:

$$\begin{aligned}
V_{(0,0)} &= \frac{V_{(1,0)} + V_{(0,1)}}{2} \\
V_{(N_x-1,0)} &= \frac{V_{(N_x-2,0)} + V_{(N_x-1,1)}}{2} \\
V_{(N_x-1,N_z-1)} &= \frac{V_{(N_x-2,N_z-1)} + V_{(N_x-1,N_z-2)}}{2} \\
V_{(0,N_z-1)} &= \frac{V_{(1,N_z-1)} + V_{(0,N_z-2)}}{2}.
\end{aligned}$$

Remark

In order to improve the numerical results, in the code we have multiplied the discretization of the boundary conditions by constants $\beta_{i,j}$: analytically it is completely uninfluent, but numerically it has improved the results, because it makes all the terms of the same magnitude.

1.7.2 Solution

Putting together (1.28), (1.29), the boundary conditions and the corner points condition, the resulting system is $(N_x \times N_z)^2$ and has non-local terms due to the presence of (1.29). The solution could be performed through any linear system solver; we have chosen to use a LAPACK routine called DGESV which has proven to be fast and robust. Other possible choices would be, for instance, the Gauss-Siegel method and the SOR method.

1.8 The parameterized eigenvalue problem

When simulating the MOSFET described in Chapter 5, we need to solve some Schrödinger-Poisson problems for the computation of the electrostatic field and the eigenproperties. In order to do that, we have chosen to use a Newton iteration, so that we need to compute the Gâteaux (directional) derivative of the density $N[V]$. Its formulation requires the derivatives, with respect to the potential V of the Schrödinger eigenproperties $\varepsilon_p[V]$ and $\chi_p[V]$. That is why we compute the derivatives of the eigenvalues and eigenvectors in the case of an eigenvalue matrix problem and then apply these results to the case of the Schrödinger eigenvalue problem.

1.8.1 The matrix eigenvalue problem

Consider the parameterized eigenvalue problem

$$M(t)e_k(t) = \lambda_k(t)e_k(t), \qquad \langle e_k, e_{k'} \rangle = \delta_{k,k'}. \tag{1.30}$$

There exist results of regularity on the eigenvalues and the eigenfunctions starting from the regularity of the (Hermitian = real symmetric) matrix $M(t)$.

Some manipulations to obtain what we need. Differentiate with respect to parameter t,

$$M'(t)e_k(t) + M(t)e'_k(t) = \lambda'_k(t)e_k(t) + \lambda_k(t)e'_k(t). \tag{1.31}$$

As the diagonalization form an orthonormal basis,

$$\langle e_k(t), e_k(t) \rangle = 1, \qquad \langle e_k(t), e'_k(t) \rangle = 0, \tag{1.32}$$

which means that $e'_k(t)$ has the following form (it has any component but the k-th):

$$e'_k(t) = \sum_{k' \neq k} \alpha_{k,k'} e_{k'}(t). \tag{1.33}$$

Take equation (1.31) and consider its scalar product by $e_k(t)$ (we omit the time dependence):

$$\begin{aligned}
\langle M' e_k, e_k \rangle + \langle M e'_k, e_k \rangle &= \lambda'_k \langle e_k, e_k \rangle + \lambda_k \langle e'_k, e_k \rangle \\
\langle M' e_k, e_k \rangle + \langle e'_k, M e_k \rangle &= \lambda'_k + 0 \\
\langle M' e_k, e_k \rangle + \lambda_k \langle e'_k, e_k \rangle &= \lambda'_k \\
\langle M' e_k, e_k \rangle &= \lambda'_k.
\end{aligned}$$

Take now equation (1.31) and consider its scalar product by $e_{k'}(t)$ (with $k' \neq k$):

$$\begin{aligned}
\langle M' e_k, e_{k'} \rangle + \langle M e'_k, e_{k'} \rangle &= \lambda'_k \langle e_k, e_{k'} \rangle + \lambda_k \langle e'_k, e_{k'} \rangle \\
\langle M' e_k, e_{k'} \rangle + \langle \sum_{p \neq k} \alpha_{k,p} e_p, M e_{k'} \rangle &= 0 + \lambda_k \langle \sum_{p \neq k} \alpha_{k,p} e_p, e_{k'} \rangle \\
\langle M' e_k, e_{k'} \rangle + \lambda_{k'} \langle \sum_{p \neq k} \alpha_{k,p} e_p, e_{k'} \rangle &= \lambda_k \alpha_{k,k'} \\
\langle M' e_k, e_{k'} \rangle + \lambda_{k'} \alpha_{k,k'} &= \lambda_k \alpha_{k,k'} \\
\langle M' e_k, e_{k'} \rangle &= (\lambda_k - \lambda_{k'}) \alpha_{k,k'} \\
\frac{\langle M' e_k, e_{k'} \rangle}{\lambda_k - \lambda_{k'}} &= \alpha_{k,k'}.
\end{aligned}$$

The most important relations we have obtained are

$$\lambda'_k(t) = \langle M'(t) e_k(t), e_k(t) \rangle \tag{1.34}$$

$$e'_k(t) = \sum_{k' \neq k} \frac{\langle M' e_k, e_{k'} \rangle}{\lambda_k - \lambda_{k'}} e_{k'}(t). \tag{1.35}$$

1.8.2 The Schrödinger eigenvalue problem

Our goal is now to adapt the same results as (1.34) and (1.35) for the case of the Schrödinger eigenvalue problem

$$-\frac{C^{S,1}}{2} \frac{d}{dz} \left[\frac{1}{m} \frac{d\chi_p[V]}{dz} \right] + C^{S,2}(V + V_c)\chi_p[V] = \epsilon_p[V]\chi_p[V]$$

$\{\chi_p[V]\}_p \subseteq H^1_o(0, l_z)$ orthonormal basis.

In order to achieve this, we have to perform the following replacements

inside problem (1.30):

$$\begin{aligned}
t &\rightarrow V \\
M(t) &\rightarrow \text{Schrödinger operator} \\
&\quad S[V](\cdot) = -\frac{C^{S,1}}{2}\frac{d}{dz}\left[\frac{1}{m}\frac{d\cdot}{dz}\right] + C^{S,2}(V+V_c)\cdot \\
\lambda_k &\rightarrow \epsilon_k \\
e_k &\rightarrow \chi_k \\
\langle f,g\rangle &\rightarrow \int fg dz \\
\frac{d\cdot}{dt} &\rightarrow d\cdot(V,U) = \text{Gâteaux-der. with respect to } V \text{ in direction } U.
\end{aligned}$$

The t-derivative of the matrix $M(t)$ in (1.30) transforms into

$$dS(V,U) = \lim_{t\rightarrow 0}\frac{S[V+tU]-S[V]}{t} = C^{S,2}U,$$

the derivative of the eigenvalues (1.34) becomes

$$d\epsilon_p(V,U) = \int dS(V,U)\chi_p[V]\chi_p[V]dz = C^{S,2}\int U|\chi_p[V]|^2 dz,$$

the $\alpha_{p,p'}$ parameters in (1.35) become

$$\alpha_{p,p'} = \frac{\int dS(V,U)\chi_p[V]\chi_{p'}[V]dz}{\epsilon_p[V]-\epsilon_{p'}[V]} = C^{S,2}\frac{\int U\chi_p[V]\chi_{p'}[V]dz}{\epsilon_p[V]-\epsilon_{p'}[V]}$$

so that the derivative of the eigenvectors (1.33) reads

$$\begin{aligned}
& d\chi_p(U,V) \\
= & \sum_{p'\neq p}\alpha_{p,p'}\chi_{p'}[V] \\
= & C^{S,2}\sum_{p'\neq p}\frac{1}{\epsilon_p[V]-\epsilon_{p'}[V]}\int U(\zeta)\chi_p[V](\zeta)\chi_{p'}[V](\zeta)d\zeta\chi_{p'}[V](z).
\end{aligned}$$

Resumé

We have obtained the following relations:

$$\begin{aligned}
d\epsilon_p(V,U) &= C^{S,2}\int U(x,\zeta)|\chi_p[V](x,\zeta)|^2 d\zeta \\
d\chi_p(V,U) &= C^{S,2}\sum_{p'\neq p}\frac{\int U(x,\zeta)\chi_p[V](x,\zeta)\chi_{p'}[V](x,\zeta)d\zeta}{\epsilon_p[V](x)-\epsilon_{p'}[V](x)}\chi_{p'}[V](x,z).
\end{aligned}$$

1.9 Newton schemes for the Schrödinger-Poisson problem

In this section the instruments for solving the Schrödinger-Poisson system are developed. We refer again to Chapter 5 for the notation, meaning and discussion of this model and adimensional quantities in our application to nanodevices. We develop the scheme with physical units and then adimensionalize in order to have the correct rescaling (we cannot reverse these two steps: dimensionless parameters would be incorrect).

Our goal is now the solution of the 1D or 2D Poisson problem under the 1D Schrödinger eigenvalue problem (which is always 1D because x just acts as a parameter)

$$-\frac{\hbar^2}{2m_e}\frac{d}{dz}\left[\frac{1}{m}\frac{d\chi_p[V]}{dz}\right] - q(V+V_c)\chi_p[V] = \epsilon_p[V]\chi_p[V] \tag{1.36}$$

$$-\mathrm{div}\,(\varepsilon_R\nabla V) = -\frac{q}{\varepsilon_0}\left(N[V]-N_D\right) \tag{1.37}$$

plus boundary conditions.

From (1.37) we define the Poisson functional

$$P[V] = -\mathrm{div}\,(\varepsilon_R\nabla V) + \frac{q}{\varepsilon_0}\left(N[V]-N_D\right), \tag{1.38}$$

which has to be minimized under the constraint of the Schrödinger equation (1.36) for the computation of the eigenproperties.

The Newton scheme for the minimization of (1.38) is

$$dP(V^{old}, V^{new}-V^{old}) = -P[V^{old}], \tag{1.39}$$

where $dP(V,U)$ means Gâteaux-differentiation at point V in direction U, which in our case is

$$\begin{aligned} dP(V,U) &= -\mathrm{div}\,(\varepsilon_R\nabla U) + \frac{q}{\varepsilon_0}dN(V,U) \\ &= -\mathrm{div}\,(\varepsilon_R\nabla U) + \frac{q}{\varepsilon_0}\int \mathcal{A}[V](z,\zeta)U(\zeta)d\zeta \end{aligned} \tag{1.40}$$

because the Gâteaux-derivative of the density can always be written in this form thanks to the formulae given in Section 1.8.

Plugging (1.40) into (1.39), after simplifying and separating the *old* and *new* terms, we obtain

$$\begin{aligned} &-\mathrm{div}\,(\varepsilon_R\nabla V^{new}) + \frac{q}{\varepsilon_0}\int \mathcal{A}[V^{old}](z,\zeta)V^{new}(\zeta)d\zeta \\ = \;& -\frac{q}{\varepsilon_0}\left(N[V^{old}]-N_D\right) + \frac{q}{\varepsilon_0}\int \mathcal{A}[V^{old}](z,\zeta)V^{old}(\zeta)d\zeta. \end{aligned} \tag{1.41}$$

Adimensionalization of the scheme

Rescaling the Newton scheme (1.41) by the parameters proper of the MOSFET device, which can be found in Table 5.6, we obtain the adimensionalized Newton scheme

$$\begin{aligned} & - \quad \mathrm{div}\,(\varepsilon_R \nabla V^{new}) \\ & + \quad C^{New,1} \int \mathcal{A}[V^{old}](x,z,\zeta) V^{new}(x,\zeta) d\zeta = \mathcal{B}[V^{old}](x,z) \\ & \quad \text{plus boundary conditions for } V, \end{aligned}$$

where we have set

$$\begin{aligned} C^{New,1} &= \frac{q\rho^* \chi^* (l^*)^3}{\varepsilon_0} \\ C^P &= \frac{qN^*(l^*)^2}{\varepsilon_0 V^*} \\ \mathcal{B}[V^{old}](x,z) &= -C^P \left(N[V^{old}] - N_D \right) \\ &\quad + C^{New,1} \int \mathcal{A}[V^{old}](x,z,\zeta) V^{old}(x,\zeta) d\zeta. \end{aligned}$$

This is a "generalized" Poisson problem which can be solved by the schemes developed either in Section 1.6 or in Section 1.7.

The iteration

We summarize here the iteration for solving the Schrödinger-Poisson problem:

Step 0 Choose an initialization, i.e. a potential V^{old}.

Step 1 Perform the following loop:

Step 1.1 Diagonalize Schrödinger by the scheme exposed in Section 1.5 to obtain its eigenvalues and eigenfunctions $\left\{ \epsilon_p[V^{old}], \chi_p[V^{old}] \right\}_p$

Step 1.2 Compute $N[V^{old}]$ (from its definition, depending on the problem which is being solved) and $\mathcal{A}[V^{old}](z,\zeta)$ by means of the formulae given in Section 1.8.

Step 1.3 Compute

$$\begin{aligned} \mathcal{B}[V^{old}] &= -C^P \left[N[V^{old}] - N_D \right] \\ &+ C^{New,1} \int \mathcal{A}[V^{old}](z,\zeta) V^{old}(\zeta) d\zeta. \end{aligned}$$

Step 1.4 Solve the "generalized" Poisson problem

$$\begin{aligned} - \quad & \operatorname{div}\left[\varepsilon_R \nabla V^{new}\right] \\ + \quad & C^{New,1} \int \mathcal{A}[V^{old}](z,\zeta) V^{new}(\zeta) d\zeta = \mathcal{B}[V^{old}] \end{aligned}$$

by the schemes developed either in Section 1.6 or in Section 1.7.

Step 1.5 Check convergence: if

$$\|V^{old} - V^{new}\|_{\text{some norm}} < \lambda_{tolerance}$$

(for instance, choose the L^∞ norm and $\lambda = 10^{-6}$) then exit the loop and set $V = V^{new}$, else set $V^{old} = V^{new}$ and go back to **Step 1.1**.

Chapter 2

Non oscillatory interpolation methods applied to Vlasov-based models

This chapter corresponds to a work [26] in collaboration with J.A. Carrillo whose reference is: "Non oscillatory interpolation methods applied to Vlasov-based models", SIAM Journal of Scientific Computing 29, 1179-1206, 2007.

2.1 Introduction

The Vlasov's equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \frac{F}{m} \cdot \nabla_v f = 0$$

is the basic kinetic model for the description of the motion of charged particles, in electronic devices and plasmas, under the effect of a force field $F(t,x)$. Here, $f(t,x,v)$ represents the particle number density in phase space (x,v) at time $t > 0$. In order to compute the force field

$$F = -\nabla_x \Phi$$

it is coupled with the Poisson's equation

$$\epsilon_0 \Delta_x \Phi = e\left[\rho[f] - C\right] = e\left[\int f dv - C\right]$$

for the computation of the electric potential $\Phi(t,x)$ due to both the self-consistent electric potential and an external density $C(x)$, like the doping profile in semiconductors or a background ion density in plasmas. Moreover, when collisional effects have to be taken into account, a Boltzmann's

operator appears as the right-hand side of the Vlasov's equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \frac{F}{m} \cdot \nabla_v f = \mathcal{Q}[f].$$

One of the main problems while simulating Vlasov-like systems is the presence of violent gradients in the distribution function, which, if not properly controlled by the numerical method, may drive to a false physical description of the phenomena due to a massive formation of spurious oscillations. Several approaches in order to solve Vlasov-based equations have been proposed in the literature, we refer to [45] for a good review and comparison of different Eulerian solvers. These oscillations appear naturally due to the energy conservation of the system and its Hamiltonian character, being the phenomena of the Landau damping an example of this behaviour [103, 45, 42]. When collisions are considered, oscillations are damped, but the control of spurious oscillations continues to be important since high gradients on the moments of the solution (density, mean velocity and temperature) are typical in the simulations, see [30, 28, 29] for instance in the semiconductor case.

One of the approaches to attack the numerical simulation of these models is based on a classical time splitting method, whose theoretical bases were given by Strang [98] and whose application for scientific computing by Cheng and Knorr [35]. Time splitting algorithms are used to subdivide a problem into simpler tasks being in our case the transport and the collision steps. In device and plasma simulations, the transport (Vlasov) step in (x, v) or (x, p)-phase space can be reduced by dimensional splitting to performing advection steps in either dimension, see [63] for fully relativistic models in which this cannot be done. Semi-Lagrangian methods were firstly introduced for meteorology problems [12] and for turbulence study [6]. The application to the Vlasov's equation was started by Sonnendrücker, Roche, Bertrand and Ghizzo in [97]. This method is based on following backwards the characteristics, and makes use of some interpolation method: Lagrange polynomials, spline [97], Hermite and ENO [95]; see [45] for an overview. Flux balance methods were proposed in [42] being the Positive Flux Conservative method (PFC-3) a particular case. This method preserves mass and positivity of the solution apart from controlling the total variation of the solution although it is only third-order accurate. This method has been modified in [38] in order to conserve the total energy too. Several improvements concerning unstructured and adaptive grids and different interpolation techniques have recently been devoloped in [61, 13].

Another successful approach in the case of collisional semiconductor and plasma models is based on the method of lines with high-order non oscillatory finite differences reconstructions of the derivatives of the distribution function $f(t, x, v)$ in phase space [64, 94] while performing a TVD (total-variation-diminishing) third order explicit Runge-Kutta method for the time

discretization. In this case, spurious oscillations are controlled by ENO [95] (Essentially Non Oscillatory) or WENO [64, 94] (Weighted Essentially Non Oscillatory) reconstructions of the fluxes and thus of the derivatives.

We propose in this paper a Semi Lagrangian (SL) and a Flux Balance Method (FBM) for Vlasov's systems based on a Weighted Essentially Non Oscillatory interpolations for the computation of point values (PWENO), see [93]. We will review this numerical technique for point values reconstruction in Section 2.

In Section 3, the Semi Lagrangian (SL) and the Flux Balance Method (FBM) are described. Both methods are based on integrating backwards in the characteristics, the second one being mass-conservative and the first one being better as for the control of the total variation. The main advantage with respect to previous semi-Lagrangian methods is the oscillation control while keeping a high-order approximation of the solution. Compared to WENO finite-differences methods, our approach avoids the restrictive CFL condition on the time stepping [72]. The proposed numerical method allows the time step Δt to be larger and constant, and the method requires much less steps (of an equivalent computational cost) than WENO finite-differences. Finally, these numerical techniques could be extended for multi-domain or multi-grid computations as done in [93].

In Section 4, we show the numerical simulations of four problems:

1. A $1D$ Vlasov-Boltzmann equation with given confining potential and a linear relaxation operator. Its asymptotic behavior was described in [27, 62].

2. A $1D$ non linear Landau damping [42], where we can observe the filamentation of the phase space during the first phase (strong oscillations in the (x, v)-space) and check whether the method is able not to add noise.

3. A $1D$ symmetric two stream instability, where we want to observe an vortex appearing after initial time, rotate and form a typical hole structure.

4. A Silicon $n^+ - n - n^+$-structured semiconductor [33, 30], where collisions are modelled through a linear Boltzmann operator: we compare results given by time splitting methods and finite differences methods.

The simulation results demonstrate overall the main features of this numerical approach: no CFL condition, good control of numerical oscillations and high accuracy on the approximations.

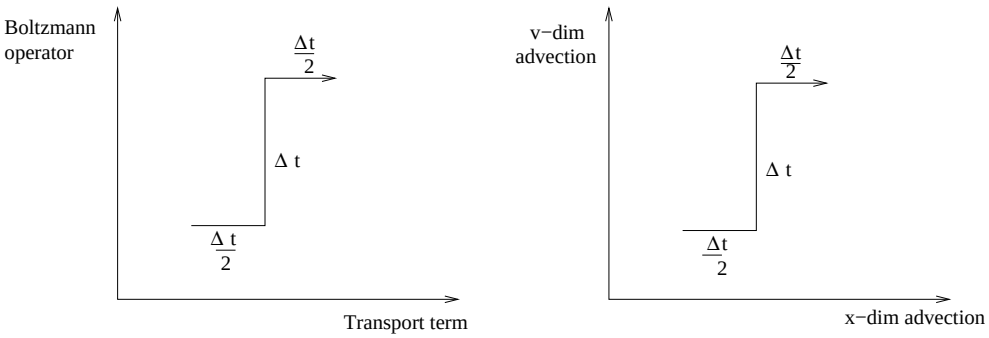


Figure 1: Time splitting schemes

2.2 Time splittings and linear advection

For integrating both the Vlasov-Boltzmann coupling

$$\frac{\partial f}{\partial t} + \underbrace{v \cdot \nabla_x f + F \cdot \nabla_v f}_{\text{Vlasov}} = \underbrace{\mathcal{Q}[f]}_{\text{Boltzmann}}$$

and the Vlasov's equation

$$\frac{\partial f}{\partial t} + \underbrace{v \cdot \nabla_x f}_{\text{adv. in } x} + \underbrace{F \cdot \nabla_v f}_{\text{adv. in } v} = 0$$

we use Strang's splitting schemes shown in Figure 1, in our simulations all reduces to solving the one-dimensional linear advection equation

$$\begin{cases} \frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = 0 \\ f(t=0,x) = f_0(x) \end{cases}$$

whose solution is the translation with velocity v of the initial function

$$f(t,x) = f_0(x - vt)$$

and is obviously mass-conservative.

Two methods are proposed: the Semi Lagrangian (SL) and the Flux Balance Method (FBM), the second one being mass-conservative and the first one being better as for the control of oscillations.

2.2.1 Semi-Lagrangian method

This method is based in directly following backwards the characteristics of the system, i.e. knowing

$$f_i^n = f(t^n, x_i)$$

we perform the step in time this way:

$$f_i^{n+1} = f(t^{n+1}, x_i) = f(t^n, x_i - v\Delta t) \simeq p^n(x_i - v\Delta t),$$

where p^n is any chosen interpolation from point values of $f(t^n, \cdot)$. Here, we will choose PWENO as interpolation method. This method is not mass-conservative.

2.2.2 Flux Balance Method

FBM (Flux Balance Method) is used in [42] to construct a conservative method. We already know that

$$f(t + \Delta t, x) = f(t, x - v\Delta t).$$

Now, let us integrate over an interval $[b_1, b_2]$, to get

$$\begin{aligned}\int_{b_1}^{b_2} f(t + \Delta t, \xi)d\xi &= \int_{b_1}^{b_2} f(t, \xi - v\Delta t)d\xi = \int_{b_1 - v\Delta t}^{b_2 - v\Delta t} f(t, \xi)d\xi \\ &= \int_{b_1 - v\Delta t}^{b_1} f(t, \xi)d\xi + \int_{b_1}^{b_2} f(t, \xi)d\xi - \int_{b_2 - v\Delta t}^{b_2} f(t, \xi)d\xi.\end{aligned}$$

If we use as notation

$$\Psi(t, x) = \int_{x - v\Delta t}^{x} f(t, \xi)d\xi,$$

we get

$$\int_{b_1}^{b_2} f(t + \Delta t, \xi)d\xi = \int_{b_1}^{b_2} f(t, \xi)d\xi + \Psi(t, b_1) - \Psi(t, b_2),$$

and dividing by $\Delta = b_2 - b_1$

$$\frac{\int_{b_1}^{b_2} f(t + \Delta t, \xi)d\xi}{\Delta} = \frac{\int_{b_1}^{b_2} f(t, \xi)d\xi}{\Delta} + \frac{\Psi(t, b_1) - \Psi(t, b_2)}{\Delta},$$

which means

$$\bar{f}_{(b_1, b_2)}(t + \Delta t) = \bar{f}_{(b_1, b_2)}(t) + \frac{\Psi(t, b_1) - \Psi(t, b_2)}{\Delta},$$

i.e., the local description of mass conservation. By denoting by $F(t, \cdot)$ the primitive of $f(t, \cdot)$, the numerical method we get applying the previous argument on the interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ is the following:

$$f_i^{n+1} = f_i^n + \frac{\Psi^n(x_{i-\frac{1}{2}}) - \Psi^n(x_{i+\frac{1}{2}})}{\Delta x},$$

where we have approximated the mean value of the function over the interval by its value at the center and

$$\Psi^n(x_{i-\frac{1}{2}}) = \int_{x_{i-\frac{1}{2}}-v\Delta t}^{x_{i-\frac{1}{2}}} f(t^n,\xi)d\xi = F(x_{i-\frac{1}{2}}) - F(x_{i-\frac{1}{2}} - v\Delta t).$$

Finally, $F(x_{i-\frac{1}{2}} - v\Delta t)$ will be approximated by a chosen interpolation from the known point values of F at $x_{i+\frac{1}{2}}$. Here, we will choose PWENO as interpolation method.

2.2.3 Total variation control

The Discrete Total Variation (DTV) is defined as

$$\sum_{i=0}^{N-2} \|f_i^n - f_{i+1}^n\|.$$

The exact solution of the linear advection equation obviously conserves the total variation. If the numerical method is able to respect this property, then it is not adding spurious oscillations due to the interpolation to the shape of the function. Take as initial datum

$$f^{step}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

for the linear advection equation with periodic boundary conditions on $x \in [-1, 1]$. As we see in Figure 2, Lagrange interpolation has the worst behaviour both in SL and in FBM, while PWENO interpolation has a very good conservation of the DTV in SL method. In FBM PWENO behaves a little bit worse, but in exchange this second method conserves the mass.

2.2.4 Disphasement errors

We want to focus now the attention on how important is the choice of the substencils, and thus, of the parameters *ntot* and *lpo* in the PWENO-*ntot*,*lpo* interpolation. In Figure 3, we see that WENO-6,4 is more accurate than WENO-5,3 both in semi-lagrangian method and in flux balance methods. This difference is due to the fact that in WENO-5,3 not all the substencils "feel" that there is a jump point. Suppose we are interpolating close to x_i, and suppose the jump point is situated between x_i and x_{i+1}. If there is a substencil which does not contain the irregularity, WENO method will give weights 0 to all of them but this one. This means that the jump runs with wrong speed $\frac{\Delta x}{\Delta t}$ with an error after time T with respect to the real speed of $\frac{\Delta x}{\Delta t}\,[\alpha - 1] \times T$ with $\alpha = v\frac{\Delta t}{\Delta x}$. Then, it is essential to use a WENO method

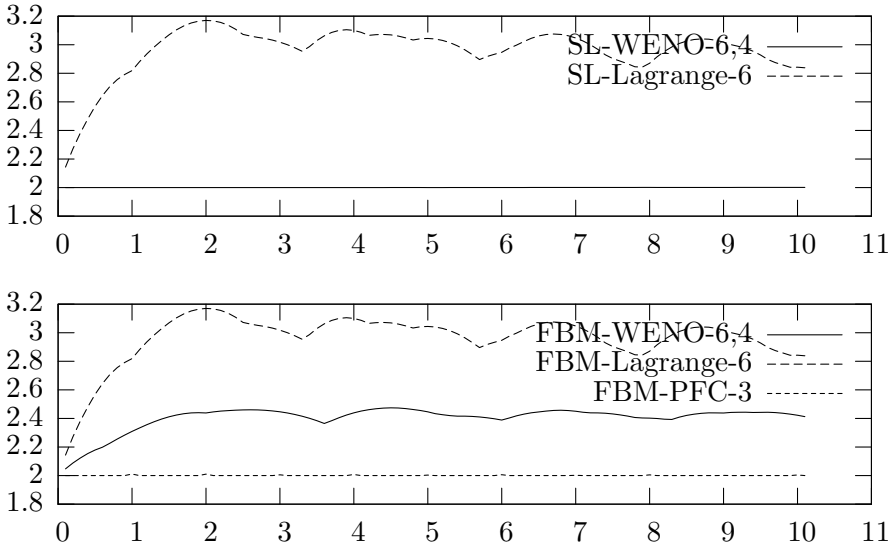


Figure 2: The evolution of Discrete Total Variation against time for the step as initial function.

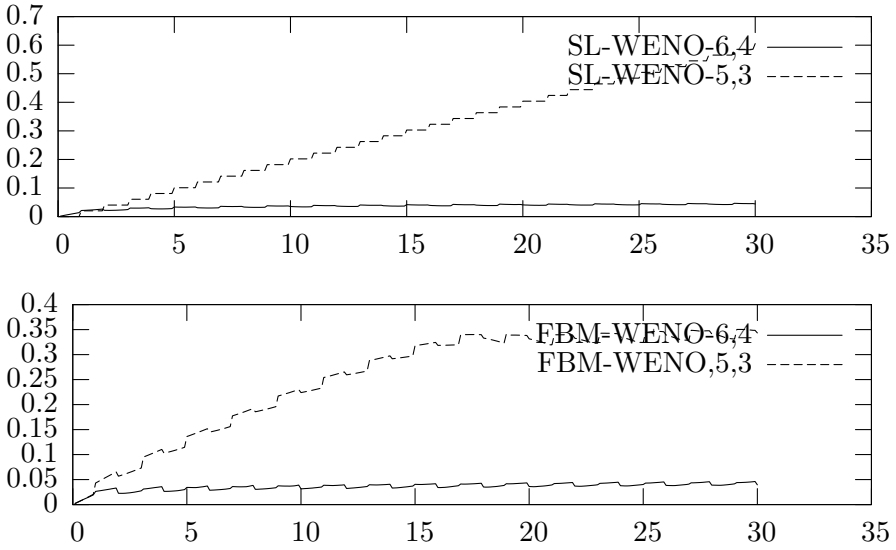


Figure 3: The evolution of L^1-norm error against the time for $N = 100$, $x \in [-\pi, \pi]$, $\Delta t = 0.1$, $t_{max} = 30$, $f_0(x) = f^{step}(x)$.

such that every substencil contains any discontinuity which may appear, like WENO-6,4. In general, it must be

$$lpo \geq \left[\frac{ntot+1}{2}\right] + 1.$$

2.3 Numerical simulations

We expose and compare results concerning the simulation of four models:

- A $1D$ Vlasov system with a linear Boltzmann operator and a given confining potential.
- A non collisional coupled Vlasov-Poisson system, by which we study the non linear Landau damping.
- A non collisional coupled Vlasov-Poisson system, by which we study the symmetric two stream instability problem.
- A collisional Vlasov-Poisson system, by which a silicon $n^+ - n - n^+$-structured semiconductor is simulated.

2.3.1 Vlasov-Boltzmann with confining potential

We solve the $1D$ Vlasov-Boltzmann equation with a confining potential $\Phi_0(x)$ and the simplest linear collision operator:

$$\begin{cases} \dfrac{\partial f}{\partial t} + v\dfrac{\partial f}{\partial x} - \dfrac{\partial \Phi_0}{\partial x}\dfrac{\partial f}{\partial v} = \dfrac{1}{\tau}\left[\rho M_1 - f\right] \\ f(0,x) = f_0(x). \end{cases}$$

From [27] we know that the solution tends to a global equilibrium given by

$$f_s = M\left(\int_{\mathbb{R}} \exp\left(-\Phi_0(x)\right) dx\right)^{-1} \exp\left(-\Phi_0(x)\right) M_1(v)$$

in L^1 norm, whenever the external potential verifies the confinement conditions:

$$\begin{cases} \bullet\, \Phi_0 \geq 0, \Phi_0 \in C^\infty(\mathbb{R}), \\ \bullet\, \exp(-\Phi_0(x)) \in L^1(\mathbb{R}), \\ \bullet\, \Phi_0 \text{ is a bounded perturbation of a uniformly convex potential on } \mathbb{R}: \\ \Phi_0 = \Phi_0^{uc} + \Phi_0^{bp} \text{ such that} \\ \qquad \text{there exists } \lambda_1 > 0 \text{ such that} \frac{\partial^2}{\partial^2 x}\Phi_0^{uc}(x) \geq \lambda_1,\ \forall x \in \mathbb{R}, \\ \text{and} \\ \qquad \text{there exists } a \text{ and } b \text{ such that } 0 < a \leq \Phi_0^{bp}(x) \leq b < \infty,\ \ \forall x \in \mathbb{R}. \end{cases}$$

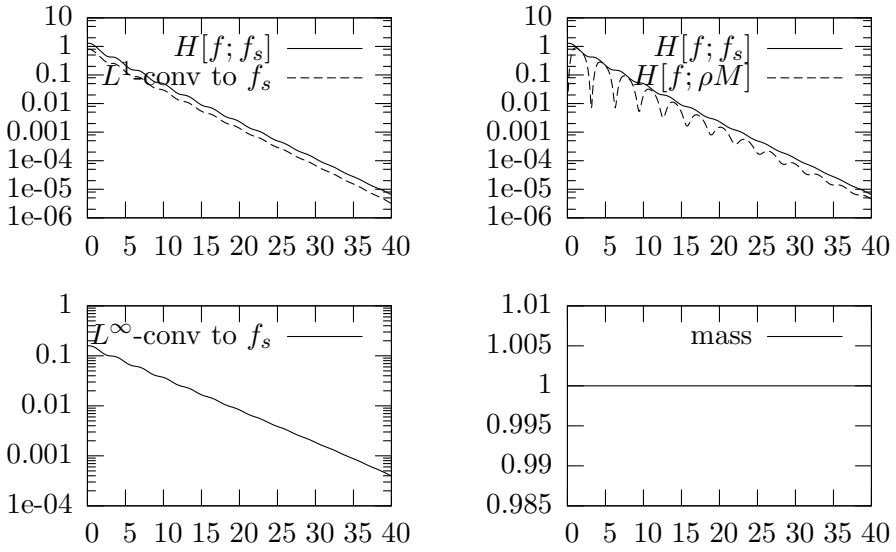


Figure 4: $1D$ Vlasov with confining potential. Test performed with 400×400 points, $\Delta t = 0.1$, SL method with PWENO-6,4, $f_0^{(1)}(x) = Z_1 \sin^2\left(\frac{x^2}{2}\right) e^{-\frac{x^2+v^2}{2}}$, $\tau = 3.5$.

The decay rate was proved to be "almost exponential" (see [27]), i.e.,

$$\|f - f_s\|_{L^1}^2 \leq H[f;f_s] \leq C(\epsilon, f_0) t^{-\frac{1}{\epsilon}}, \tag{2.1}$$

for all $\epsilon > 0$. Global and local relative entropies are measures of how far is f from the global equilibrium f_s and the local equilibrium $\rho(t,x)M_1$.

$$\begin{cases} H[f;f_s] = \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|f-f_s|^2}{f_s}\, dvdx \\ \tilde{H}[f;\rho M_1] = \displaystyle\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|f-\rho M_1|^2}{f_s}\, dvdx. \end{cases} \tag{2.2}$$

Global and local relative entropies satisfy the ODE inequalities system (see [27]):

$$\begin{cases} -\dfrac{d}{dt} H[f;f_s] \geq K \tilde{H}[f;\rho M_{\theta_0}] \\ \dfrac{d^2}{dt^2} \tilde{H}[f;\rho M_{\theta_0}] \geq K' H[f;f_s] - C(f,\epsilon)\tilde{H}[f;\rho M_{\theta_0}]. \end{cases} \tag{2.3}$$

We show numerical results in the particular case of $\Phi_0(x) = \frac{x^2}{2}$. In this case, the Vlasov part gives a rotation of the initial function $f_0(x,v)$ while the collision part thermalizes the velocity distribution towards $M_1(v)$.

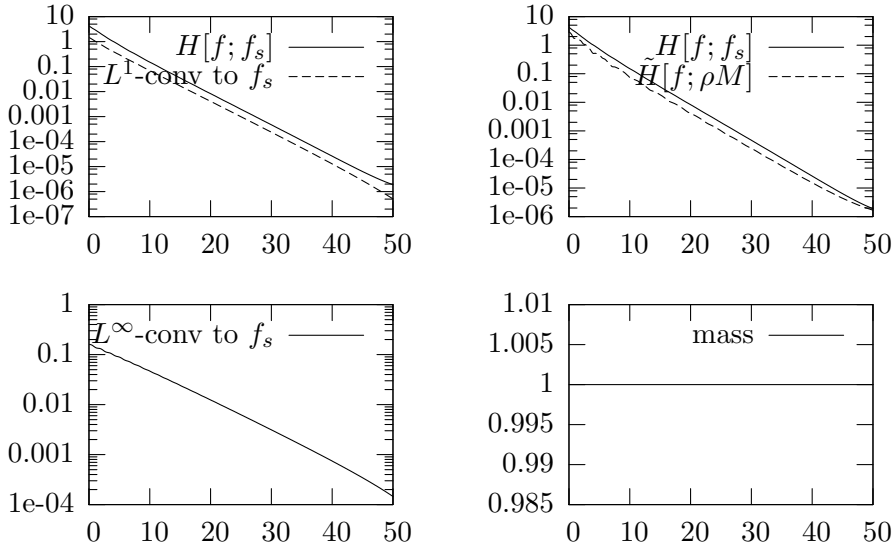


Figure 5: $1D$ Vlasov with confining potential. Test performed with 400×400 points, $\Delta t = 0.1$, SL method with PWENO-6,4, $f_0^{(2)}(x) = Z_2 \sin^2\left(\frac{x^2}{2}\right) sin^2\left(\frac{v^2}{2}\right) e^{-\frac{x^2+v^2}{2}}$, $\tau = 3.5$.

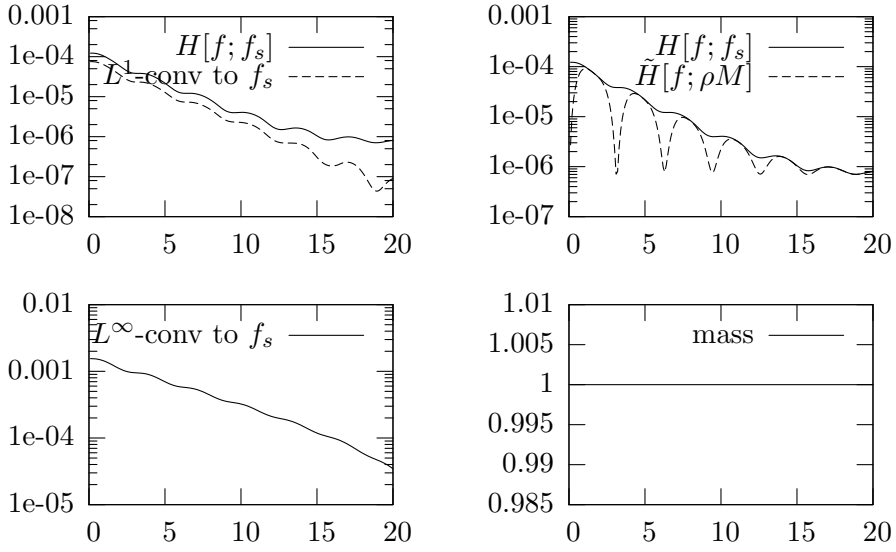


Figure 6: $1D$ Vlasov with confining potential. Test performed with 400×400 points, $\Delta t = 0.1$, SL method with PWENO-6,4, $f_0^{(3)}(x) = Z_3 \left[1 + 0.05 \sin^2\left(\frac{x^2}{2}\right)\right] e^{-\frac{x^2+v^2}{2}}$, $\tau = 3.5$.

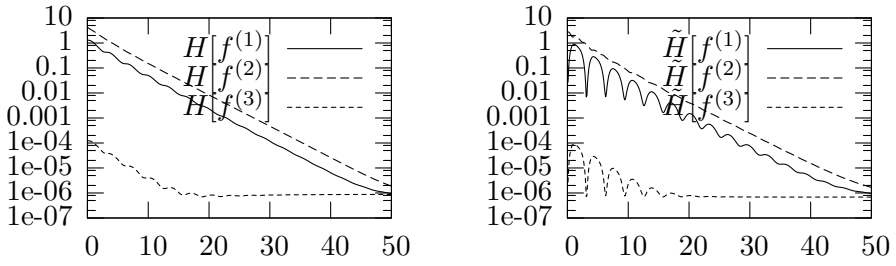


Figure 7: $1D$ Vlasov with confining potential. From this comparison, it is quite clear that the oscillation rate is the same for both initial functions, even if the behaviour is different.

Moreover, an analysis of the spectrum of operator

$$f \longmapsto -v\frac{\partial f}{\partial x} + x\frac{\partial f}{\partial v} + \frac{1}{\tau}\left[\rho M_1 - f\right]$$

has recently been done in [62] showing the existence of an spectral gap in a suitable L^2-weighted space and thus, of the exponential convergence in that space and as a consequence in L^1 towards f_s.

Both the previous ODEs inequalities (2.3) and the spectral analysis suggest the appearance of oscillations in the trend of solutions towards global equilibrium. The ODEs inequalities (2.3) shows that the trend of convergence towards local equilibria is compensated by the transport term that should push the solution out of the local equilibria manifold whenever the solution approaches a local equilibrium which is not the global one. This fact suggests an oscillation both in the local and global relative entropy. On the other hand, assuming the first non-zero eigenvalues in the spectrum are given by a pair of conjugate eigenvalues λ_1 and λ_2 then, we expect oscillations of the L^2-weighted norm with a slope decay given by $\Re(\lambda_i) < 0$ and oscillation frequency given by the absolute value of $\Im(\lambda_i)$. We refer to [62] for details.

In the upper left graphs of Figures 4 and 5 we see, for two different initial functions normalized to have unit mass, that the L^1-convergence of f towards f_s is led by the decay of $H[f; f_s]$, like in (2.1). The convergence is clearly exponential as the results of [62] prove for the L^1-norm.

In the lower left graphs we see that L^∞-convergence is also expected to be exponential, although this result has not been proven yet. In the upper right graphs we see that the oscillations of global and relative entropies correspond, due to the couplings (2.3). In the lower right graphs we observe that the mass seems pretty well conserved.

We shall now perform tests with different domain lengths and different initial functions: the oscillation rate and the decay slope should not change, because we are not using periodic conditions. In fact, we are solving the

Cauchy problem by neglecting the distribution function f outside a suitable domain chosen in such a way that the values of f near the border are almost negligible.

In Figure 7 we compare the decay of global and relative entropies for two different initial functions. Even if the amplitude of the oscillation is different, it seems evident that the oscillation rate and the decay slope correspond.

The system seems to "hesitate" between states where it is close to a local equilibrium ρM_1 and the convergence to the global equilibrium f_s. In [43] similar oscillations have been reported in the case of the full non-linear Boltzmann equation for rarefied gases in a box with periodic boundary conditions. A numerical approximation of the slope γ and the frequency ω of the decaying oscillations towards global equilibrium gives:

$$\left[\begin{array}{c|c|c|c} f_0(x) & L & \omega & \gamma \\ \hline f_0^{(1)}(x) & 4\pi & 3.15 & -0.298368 \\ f_0^{(1)}(x) & 6\pi & 3.15 & -0.298872 \\ f_0^{(2)}(x) & 4\pi & 3.125 & -0.304400 \\ f_0^{(2)}(x) & 6\pi & 3.125 & -0.304858 \end{array}\right].$$

A refinement study has been perfomed to check that the oscillation frequency and the decay slope do not depend on the dimensions of the domain, nor on the initial datum we choose: they are determined by the system itself.

2.3.2 $1D$ Vlasov-Fokker Planck with confining potential

We solve the $1D$ Vlasov equation with a confining potential $\Phi_0(x) = \frac{x^2}{2}$ and a Fokker Planck operator as collision operator:

$$\begin{cases} \dfrac{\partial f}{\partial t} + v\dfrac{\partial f}{\partial x} - x\dfrac{\partial f}{\partial v} = \dfrac{1}{\tau}\dfrac{\partial}{\partial v}\left[vf + \Theta\dfrac{\partial f}{\partial v}\right] \\ f(0,x) = f_0(x). \end{cases} \tag{2.4}$$

We show in Figures 8 and 9 that the solution tends, as well as for the case with a relaxation time operator, to a global equilibrium given by

$$f_s = M\left(\int_{\mathbb{R}} \exp\left(-\frac{x^2}{2}\right) dx\right)^{-1} \exp\left(-\frac{x^2}{2}\right) M_\Theta(v)$$

in L^1 norm. Global and local relative entropies are defined in (2.2) and are measures of how far is f from the global equilibrium f_s and the local equilibrium ρM_Θ.

Equation (2.4) becomes, after an obvious manipulation,

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - \frac{\partial f}{\partial v}\left[-\left(x + \frac{v}{\tau}\right) f\right] = \frac{\Theta}{\tau}\frac{\partial^2 f}{\partial v^2}.$$

In this case the v-advection velocity is not constant: the term $\left[-\left(x+\frac{v}{\tau}\right)f\right]$ depends on v. Therefore, in order to follow backwards the characteristics, we need to solve the ODE

$$\frac{dv}{dt} = -\frac{v}{\tau} - x,$$

which gives

$$v(t^{n+2/4}) = v(t^{n+1/4})\exp(-\Delta t/\tau) + \tau x\left[1-\exp(-\Delta t/\tau)\right].$$

(with $\Delta t = t^{n+2/4} - t^{n+1/4}$) i.e., solving with respect to $v(t^{n+1/4})$,

$$v(t^{n+1/4}) = v(t^{n+2/4})\exp(\Delta t/\tau) - \tau x\left[\exp(\Delta t/\tau)-1\right].$$

So, the solution to the v-advection is given by

$$f(t^{n+2/4},x,v) = f(t^{n+1/4},x,ve^{\Delta t/\tau} - \tau x\left(e^{\Delta t/\tau}-1\right)J(x,v)$$

where $J(t,x,v)=\exp(t/\tau)$ is the Jacobian of the change from to $v(t^{n+2/4})$ to $v(t^{n+1/4})$.

As for the "collision step"

$$\frac{\partial f}{\partial t} = \frac{\Theta}{\tau}\frac{\partial^2 f}{\partial v^2}.$$

it is solved through a plane Euler step, the second derivative being approximated by a standard finite difference.

2.3.3 $1D$ non linear Landau damping

The (normalized) model is:

$$\begin{cases} \dfrac{\partial f}{\partial t} + v\dfrac{\partial f}{\partial x} - \dfrac{\partial \Phi}{\partial x}\dfrac{\partial f}{\partial v} = 0 \\ \dfrac{\partial^2 \Phi}{\partial x^2} = 1 - \displaystyle\int_{\mathbb{R}} f dv \\ f_0(x,v) = f_0(x,v) = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}\left[1+0.5\cos\left(\dfrac{x}{2}\right)\right]. \end{cases}$$

Landau theory has been developed in order to study the propagation of small amplitude waves in a uniform plasma (with no magnetic field and no collisions). It conjectures an interchange of energy between the electric field (potential energy) and resonant particles (kinetic energy) driven by the wave, which produces an oscillating evolution of the electric energy

$$\int_0^L |E(t,x)|^2 dx.$$

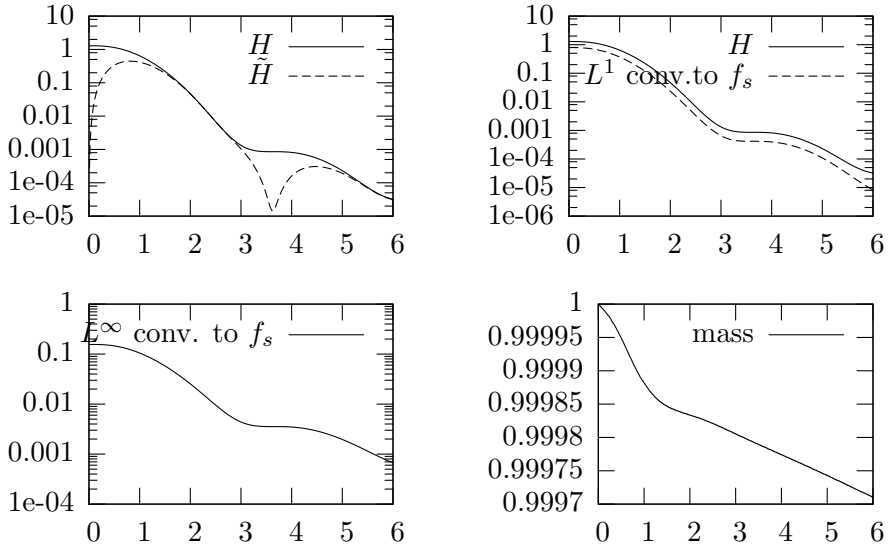


Figure 8: $1D$ Vlasov Fokker Planck. Test performed with 64×64 points, $\Delta t = 0.01$, SL method with PWENO-6,4, $f_0^{(2)}(x) = Z_1 \sin^2\left(\frac{x^2}{2}\right) e^{-\frac{x^2+v^2}{2}}$, $\tau = 1$, $\Theta = 1$.

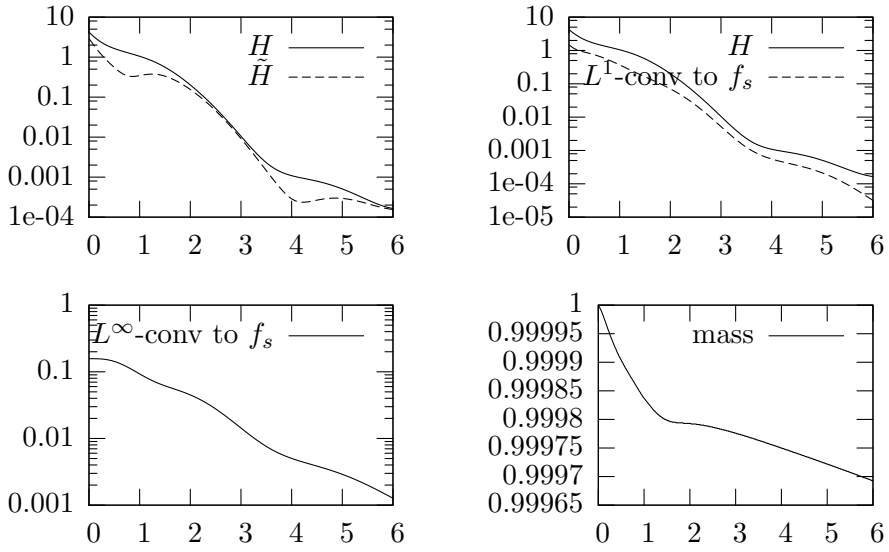


Figure 9: $1D$ Vlasov Fokker Planck. Test performed with 64×64 points, $\Delta t = 0.01$, SL method with PWENO-6,4, $f_0^{(2)}(x) = Z_2 \sin^2\left(\frac{x^2}{2}\right) \sin^2\left(\frac{v^2}{2}\right) e^{-\frac{x^2+v^2}{2}}$, $\tau = 1$, $\Theta = 1$.

Its decay is expected to be exponential in the first phase of the evolution, then it should start oscillating around some equilibrium value.

The numerical method must properly take into account two aspects: the filamentation of the phase space and the conservation properties.

The better results are given by FBM-PWENO-5,3: this method is able to describe the filamentation of the phase space, it controls the strong oscillations (part of the physical phenomenon), which, after some time, should start to disappear. It has a good conservation of the total energy (oscillations about 0.03%, 0.015% for Lagrange and 0.05% for PFC-3, like we see in Figure 10). As for FBM-Lagrange-5 and FBM-PFC-3, both give physically non reliable results, the first one because of a violent, spurious growth of the oscillations, and the second one because of a violent repression of them, as we can see in Figure 11. From Figure 12 we can see that the reconstruction given by PFC-3 is, in effect, less definite than the other ones, due to the low order. Lagrange reconstruction is more irregular, due to the parasite oscillations.

2.3.4 Two stream instability

The (normalized) model is:

$$\begin{cases} \dfrac{\partial f}{\partial t} + v\dfrac{\partial f}{\partial x} - \dfrac{\partial \Phi}{\partial x}\dfrac{\partial f}{\partial v} = 0 \\ \dfrac{\partial^2 \Phi}{\partial x^2} = 1 - \displaystyle\int_{\mathbb{R}} f dv \\ f_0(x,v) = Z\dfrac{2}{7\sqrt{2\pi}}\left(1+5v^2\right)\left[1+\alpha\left(\dfrac{\cos(2kx)+\cos(3kx)}{1.2}+\cos(kx)\right)\right]e^{-\frac{v^2}{2}} \end{cases}$$

with $\alpha = 0.01$, $k = 0.5$ and Z a normalizing factor to impose periodicity boundary conditions on the electric field. This test has been studied in [45]; there, Filbet and Sonnendrücker compare several methods for a coarse points grid and observe which ones have good conservation properties and which ones are able to follow thin details of the distribution function: among their cases, a semi-Lagrangian method with a cubic spline interpolation has given the most reliable results.

Our simulations show that even if both WENO-5,3 and WENO-6,4 are fifth order, there may be a substancial difference in their reconstructions, due to the problem the first method has in following shock-like situations, which is the case of this test. Dispersion errors become too important, in the phase space violent oscillations are produced (look at the level curves in Figure 13) which destabilize the simulation (the variation of the L^2-norm is about 400%, as we can remark from Figure 14). As for PFC-3 method, it has good conservation properties (Figure 14), but, compared with WENO-6,4, observing the level curves in Figure 13 it is evident that it is lower order and the reconstruction is less detailed.

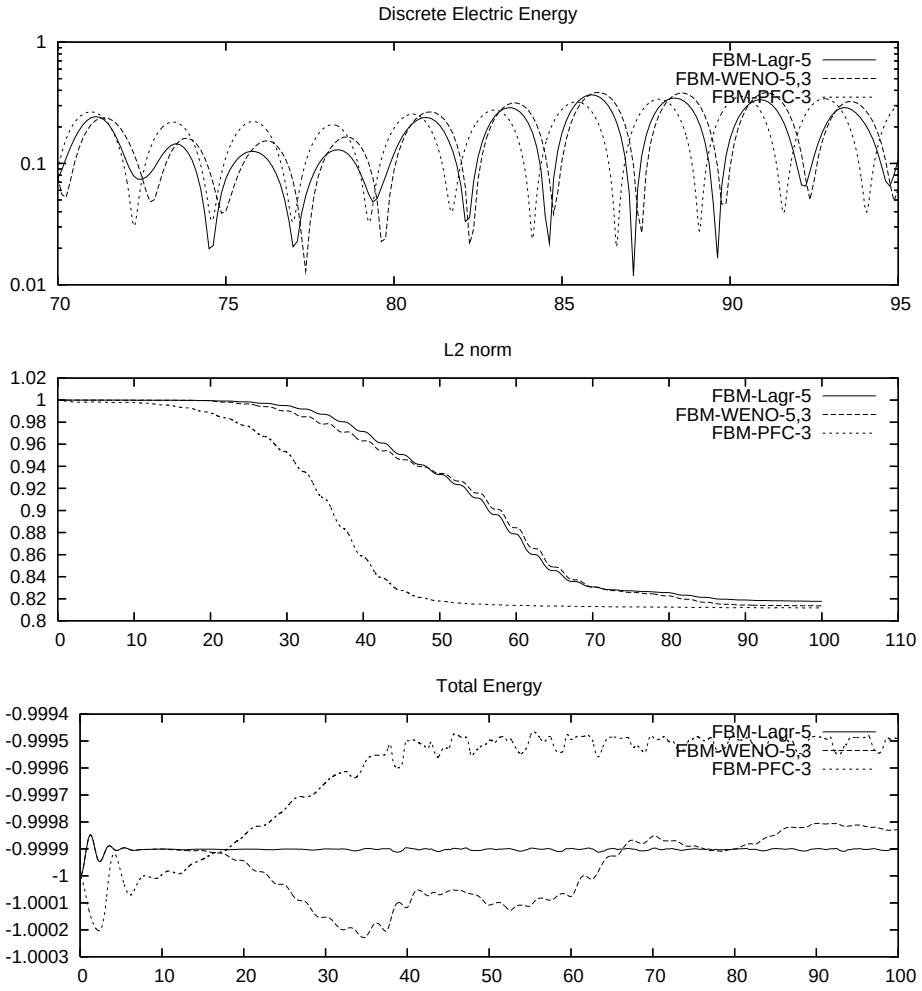


Figure 10: $1D$ non linear Landau damping. Evolution of the discrete electric energy, the L^2-norm and the total energy. Test performed with 256×256 points, $\Delta t = 0.125$ for WENO, $\Delta t = 0.01$ for PFC-3, $x \in [0, 4\pi]$, $v \in [-6, 6]$.

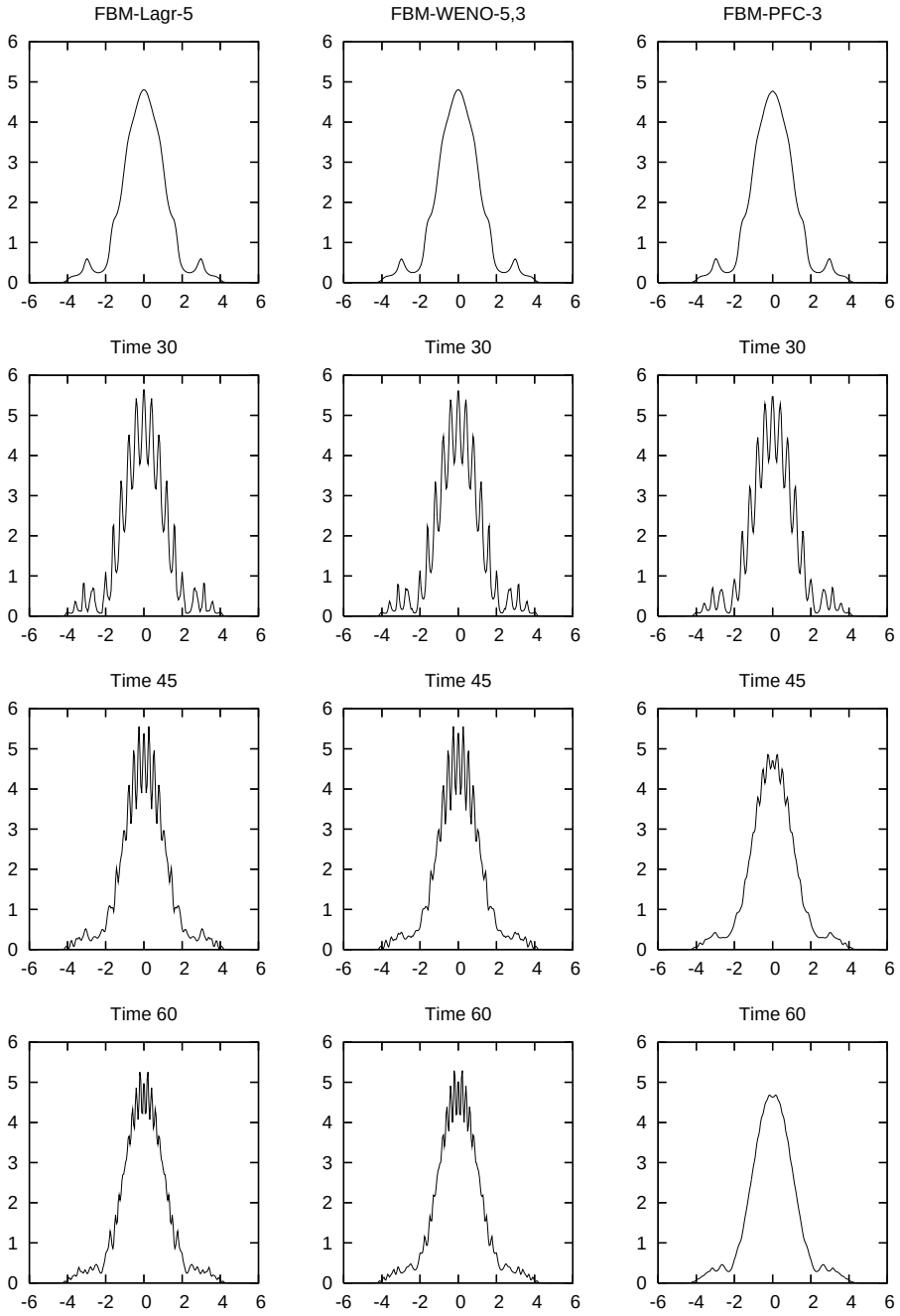


Figure 11: $1D$ non linear Landau damping. Evolution of the x-integrated distribution function, for several methods. Test performed with 256×256 points, $\Delta t = 0.125$ for WENO, $\Delta t = 0.01$ for PFC-3, $x \in [0, 4\pi]$, $v \in [-6, 6]$.

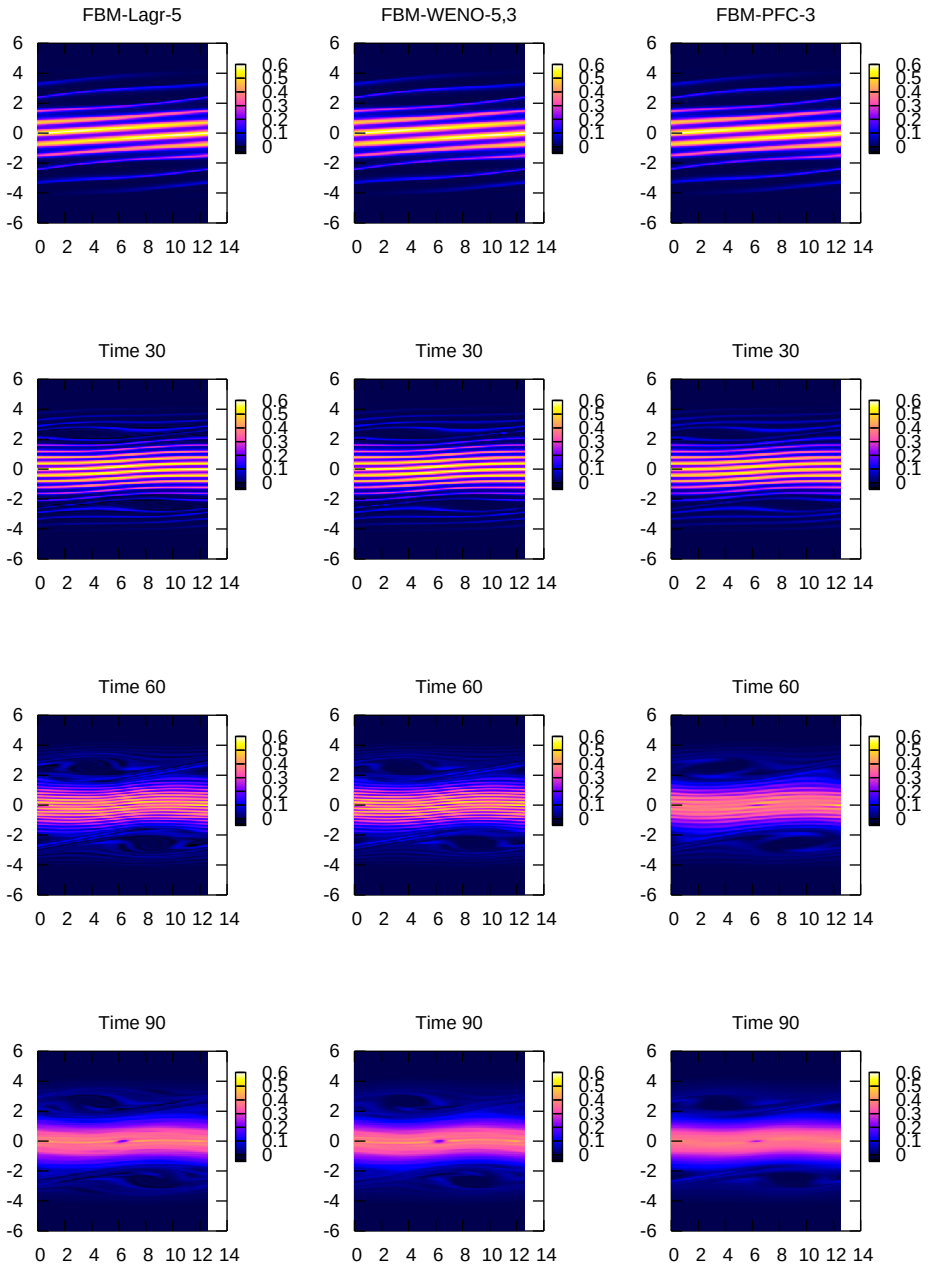


Figure 12: $1D$ non linear Landau damping. Evolution of the level curves of the distribution function, for several methods. Test performed with 256×256 points, $\Delta t = 0.125$ for WENO, $\Delta t = 0.01$ for PFC-3, $x \in [0, 4\pi]$, $v \in [-6, 6]$.

We have not shown level curves about semi-Lagrangian simulations, but a little remark is worth: SL-WENO-5,3 is unstable too, even if it behaves better than FBM-WENO-5,3. SL-WENO-6,4 as well as FBM-WENO-6,4 gives a fairly detailed reconstruction (in fact they are almost indistinguishable); the L^2-norm and the total energy are a bit worse conserved, but in any case acceptable.

2.3.5 Semiconductor

The model is

$$\begin{cases} \dfrac{\partial f}{\partial t} + v\dfrac{\partial f}{\partial x} - \dfrac{e}{m}E\dfrac{\partial f}{\partial v} = \dfrac{1}{\tau}\left[M_{\Theta_0}\rho[f] - f\right] \\ \epsilon_{Si}\dfrac{\partial^2\Phi}{\partial x^2} = -e\left[\rho[f] - C\right], \qquad E = -\dfrac{\partial\Phi}{\partial x} \\ f_0(x,v) = M_{\Theta_0}(v)C(x) \\ \tau(t,x) = \dfrac{m}{e}\dfrac{2\mu_0}{1+\sqrt{1+4((\mu_0/v_0)E(t,x))^2}}, \end{cases}$$

which describes the electron transport in a $n^+ - n - n^+$ diode where the interaction with the semiconductor crystal is taken into account by an effective relaxation-time operator. The lattice temperature is set $T_0 = 300K$, and is related to the thermal energy of the lattice by $\Theta_0 = \frac{k_B}{m}T_0$. The relaxation time depends nonlinearly on the electric field producing a saturation on the drift speed, see [30, 33] for more details. We consider a diode of channel length $L = 0.4\mu m$ and doping profile:

$$C(x) = \begin{cases} 5\times 10^5 \frac{1}{\mu m^3} & 0 \le x \le 0.1 \\ 2\times 10^3 \frac{1}{\mu m^3} & 0.1 < x < 0.5 \\ 5\times 10^5 \frac{1}{\mu m^3} & 0.5 \le x \le 0.6. \end{cases}$$

We apply a fixed potential drop at the drain, $\Phi(t,0) = 0$ V, and $\Phi(t,L) = V_{bias}$ V.

In Figure 15, density, mean velocity, electric field, potential, energy and the distribution function in v at a point near the end of the channel at the stationary state are plotted. We are comparing the results of our time splitting method based on FBM-WENO-6,4 with time step $\Delta t = 0.01$ and Finite Differences WENO5 coupled with third order Runge-Kutta for time discretization [30, 33]. Results of both simulations are almost indistinguishable.

In the following table, we compare the results given by Finite Differences WENO5 method and by FBM-WENO-6,4 for several time steps. One of the main advantages of SL or FBM-WENO schemes is that they do not have

an a priori restriction on the time step as the CFL condition for Finite Differences WENO5. Results given up to $\Delta t = 0.16$ are acceptable in mean L^∞ relative difference. Therefore, our scheme is much less costly from the computational point of view, since 62 time-steps corresponding to $\Delta t = 0.16$ of FBM-WENO-6,4 are enough to give equivalent results to 9000 time-steps corresponding to $\Delta t \simeq 0.0007$ of WENO5 method with comparable cost for each time-step between the methods.

time step	mean L^∞ relative difference
0.01	0.040116%
0.10	0.061354%
0.16	0.094815%
0.17	1.128900%
0.20	3.309995%

2.4 Appendix

We summarize the physical constants used for the simulation:

constant	physical meaning	magnitude
m	effective electron mass	$0.26 \times 0.9109(10^{-30}Kg)$
e	elementary electric charge	$0.1602(10^{-18}C)$
k_b	Boltzmann's constant	$0.138046 \times 10^{-4}(10^{-18}\frac{J}{K})$
ϵ_{Si}	Silicon dielectric permittivity	$11.7 \times 8.85418(10^{-18}\frac{F}{\mu m})$
μ_0	bulk mobility	0.1323
v_0	saturation velocity	0.13.

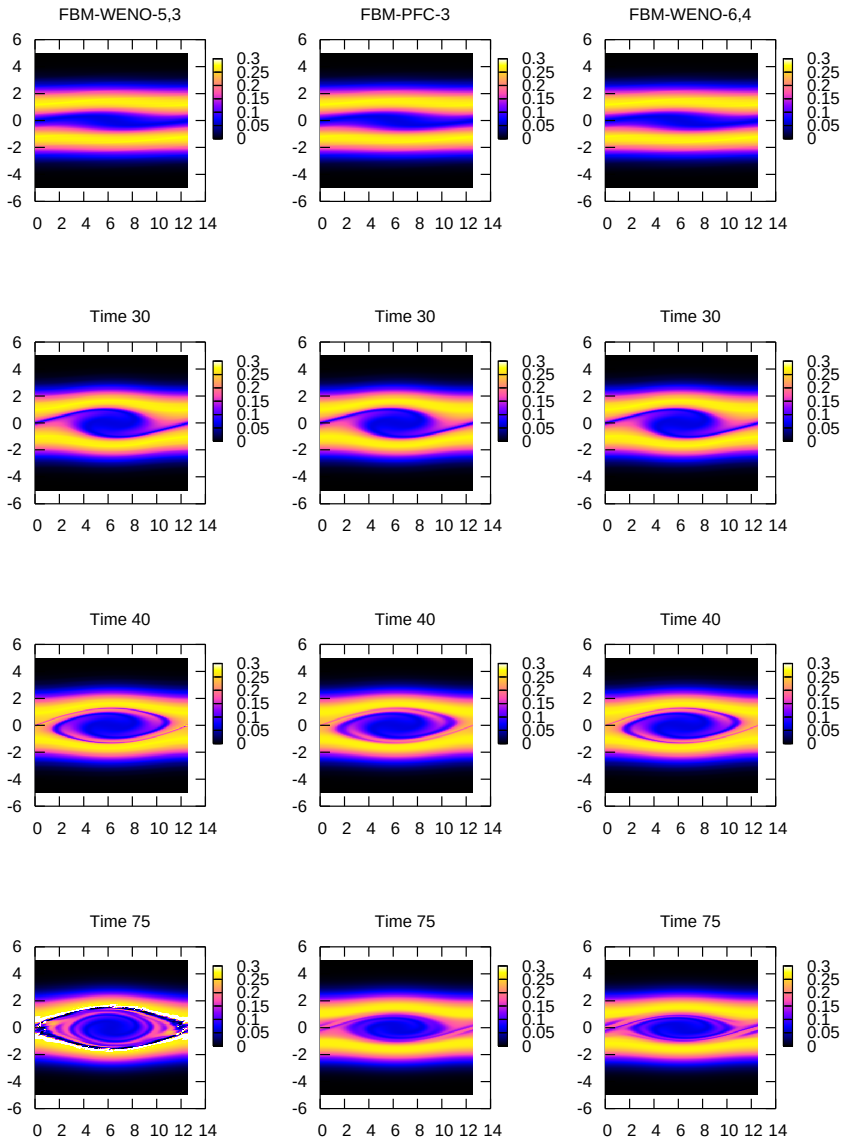


Figure 13: Two stream instability. Evolution of the level curves. Simulation performed with 256×256 points, $\Delta t = 0.125$ for WENO methods, $\Delta t = 0.01$ for PFC-3. $x \in [0, 4\pi]$, $v \in [-5, 5]$.

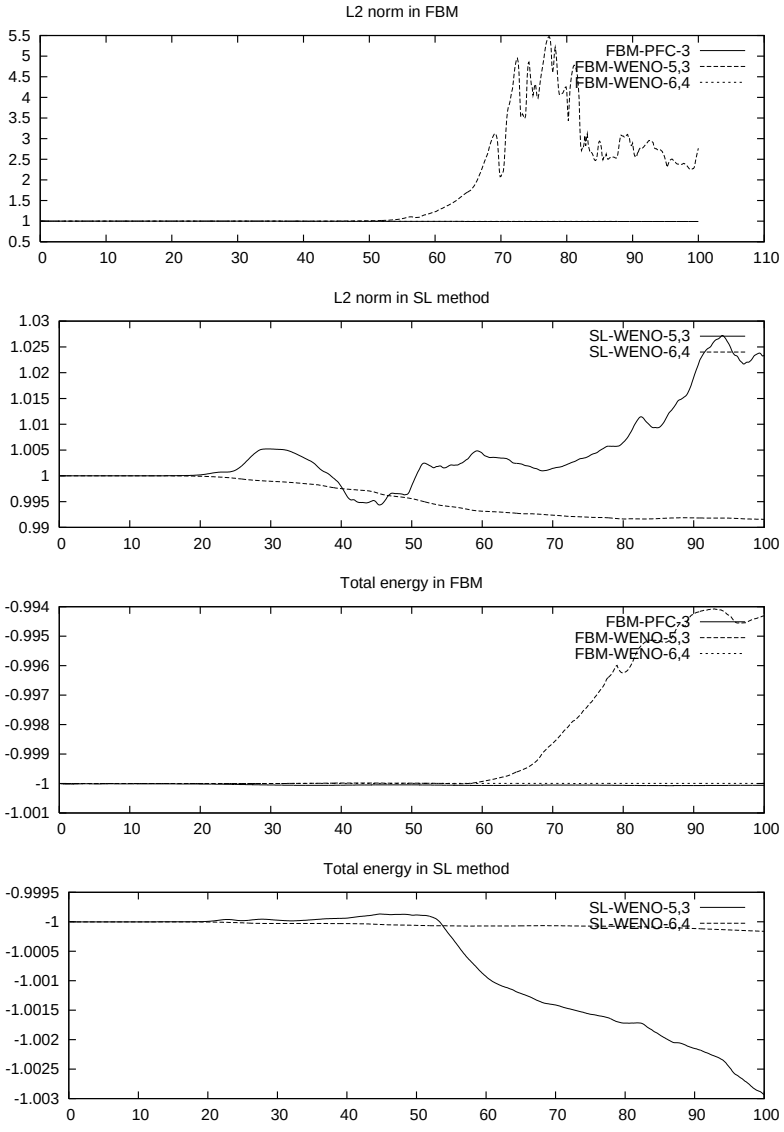


Figure 14: Two stream instability. Evolution of the L^2-norm and the total energy. Simulation performed with 256×256 points, $\Delta t = 0.125$ for WENO methods, $\Delta t = 0.01$ for PFC-3. $x \in [0, 4\pi]$, $v \in [-5, 5]$.

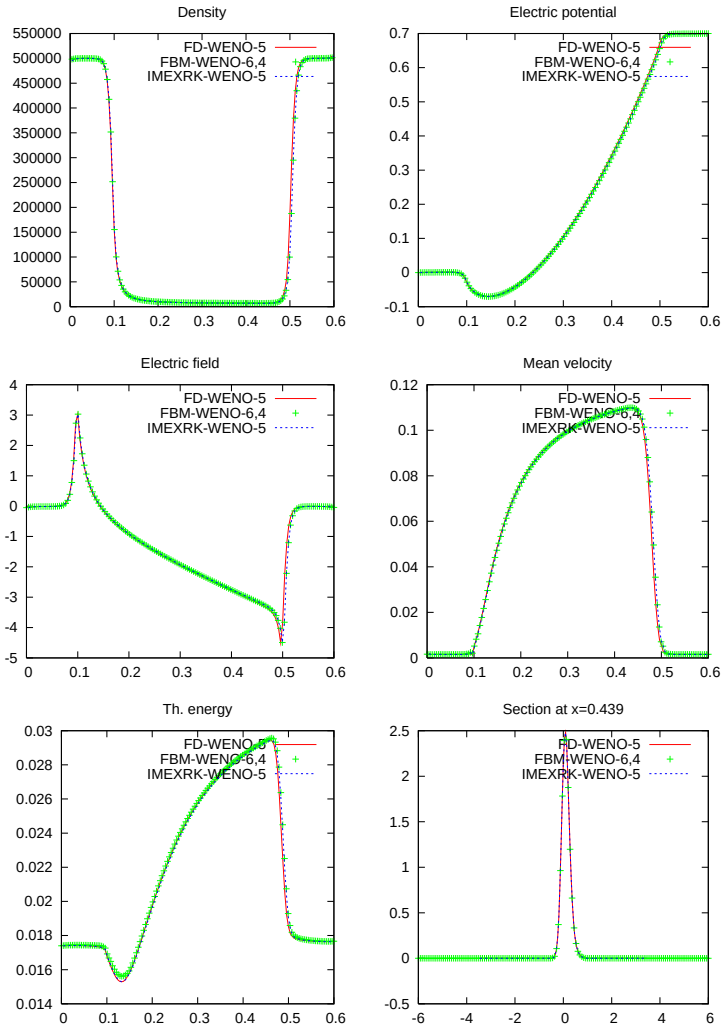


Figure 15: $1D$ $n^+ - n - n^+$ Silicon semiconductor. Evolution of several macroscopic magnitudes, for explicit Runge Kutta-3 WENO-5 Finite Differences method (CFL is set 0.6), WENO-6,4 Flux Balance method ($\Delta t = 0.01$), and an IMEX-LRR(3,2,2) FD method (CFL set 0.6). Grids are set 150×150 points.

Chapter 3

A semi-lagrangian deterministic solver for the semiconductor Boltzmann-Poisson system

This chapter corresponds to a work [25] in collaboration with J.A. Carrillo and A. Majorana whose reference is: "A Semi-lagrangian deterministic solver for the semiconductor Boltzmann-Poisson system", Commun. Comput. Phys. 2, 1027-1054, 2007.

3.1 Introduction

The semi-classical Boltzmann transport equation (BTE) is a mesoscopic description of the transport/collision of charged particles in an electronic device and is given by

$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_k \varepsilon \cdot \nabla_x f - \frac{q}{\hbar} E \cdot \nabla_k f = \mathcal{Q}[f] \tag{3.1}$$

where $f(t, x, k)$ measures the probability density of finding an electron at time t in position x with wave vector k. The parameter $\hbar$ is the Planck constant divided by 2π and q is the positive elementary charge.

The band structure of the semiconductor crystal is described by the energy-band function which can be approximated by a parabolic function given by

$$\varepsilon(k) = \frac{1}{2}\frac{\hbar^2}{m^*}|k|^2, \tag{3.2}$$

where m^* is the effective electron mass. In a first step, we shall consider the most important scattering mechanisms in Si: acoustic phonon scattering,

in its elastic approximation, and optical phonon scattering with a single frequency ω. Therefore, the structure of the collision operator [90, 100] is

$$\begin{aligned}\mathcal{Q}[f](t,x,k) &= \int_{\mathbb{R}^3} \left[S(k',k)f(t,x,k') - S(k,k')f(t,x,k)\right] dk' \\ &= \int_{\mathbb{R}^3} S(k',k)f(t,x,k')\,dk' - f(t,x,k)\int_{\mathbb{R}^3} S(k,k')\,dk' \\ &= \mathcal{Q}^+[f] - \mathcal{Q}^-[f] \end{aligned} \tag{3.3}$$

with

$$\begin{aligned}S(k,k') = K&\left[(n_q+1)\delta(\varepsilon(k')-\varepsilon(k)+\hbar\omega) + n_q\delta(\varepsilon(k')-\varepsilon(k)-\hbar\omega)\right] \\ &+ K_0\delta(\varepsilon(k')-\varepsilon(k))\end{aligned} \tag{3.4}$$

where n_q is the occupation number of phonons

$$n_q = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T_L}\right) - 1}, \tag{3.5}$$

k_B is the Boltzmann constant and T_L the lattice temperature. The kernel K_0 is

$$K_0 = \frac{k_B T_L E_{ac}^2}{4\pi^2\hbar {u_l}^2 \rho_0} \tag{3.6}$$

where E_{ac} is the deformation potential, u_l is the sound velocity and ρ_0 is the crystal density. The kernel K is

$$K = \frac{D_t k^2}{8\pi^2\hbar\rho_0\omega}, \tag{3.7}$$

where ω is the frequency and $D_t k$ is the optical coupling constant.

The self-consistent electrostatic field is computed through Poisson's equation

$$\Delta\Phi = \frac{q}{\epsilon}\left[\rho(t,x) - N_D(x)\right] \tag{3.8}$$

where ρ is the electron density

$$\rho(t,x) = \int_{\mathbb{R}^3} f(t,x,k)\,dk, \tag{3.9}$$

ϵ is the Silicon dielectric permittivity ($\epsilon = \epsilon_r\epsilon_0$ with ϵ_0 the vacuum dielectric permittivity and ϵ_r the Silicon relative dielectric permittivity), $N_D(x)$ represents the doping profile, which takes into account the injected impurities in the semiconductor lattice. The solution of the Poisson's equation $\Phi(t,x)$ gives the electrostatic potential, so that the electrostatic field is given by

$$E(t,x) = -\nabla_x\Phi(t,x). \tag{3.10}$$

Further information about semiconductor modelling and related mathematical issues can be found in [84].

This system has been traditionally solved by means of Direct Simulation Monte Carlo (DSMC) methods due to the easy incorporation of new physical effects by means of adding suitable scattering operators and its efficiency in two and three-dimensional devices [100]. Nevertheless, direct deterministic numerical methods have been recently proposed in the literature [91, 83, 28, 92, 29, 48] improving and complementing some features of the DSMC methods: noise-free results, detailed information of the distribution functions, transient description, different materials [22, 50, 51]... We refer to [28, 29, 8, 50, 49] for a complete discussion of these issues and to [21] for a review of the state of the art in the deterministic numerical simulation of the Boltzmann-Poisson system.

In this work, we propose a new deterministic numerical scheme for this system. In contrast with the approach in [83, 28], we work in the original coordinates (t, x, k) by using a splitting strategy decoupling transport from collision. For the transport part, we apply a semi-lagrangian numerical method based on a nonlinear local essentially non-oscillatory interpolation method recently developed in [26] for transport-like kinetic equations. The main objective of this choice is to avoid the potentially restrictive CFL condition emanating from the use of finite-differences WENO methods in energy and angular variables in [28, 29] but keeping a good control of possible oscillations during the transport steps.

The collisional step is performed by interpolation from computed values on a cartesian grid to obtain the missing values of the distribution function on the surfaces of equal energy needed for the evaluation of the collision operator $\mathcal{Q}(f)$. Different interpolation procedures have been tested from the simplest and less accurate direct linear interpolation to the most advanced nonlinear local essentially non-oscillatory interpolation method in [26] as above. Conservation of mass in the collision steps is inforced by redefining the loss operator as in [22]. The different choices for interpolation in the collision step and the splitting of the operators will be discussed and compared.

This new deterministic scheme is developed in Section 2 while Section 3 is devoted to show its performance to compute steady and transient states of 1d devices and comparisons to the numerical scheme introduced in [28]. The main advantage of this scheme being the smaller number of time steps needed and the much better definition of the distribution function in phase space. Moreover, more realistic collision operator for Si takes into account the different equivalent valleys in the conduction band of Si leading to several optical-phonon scattering operators with different frequencies $\hbar\omega$ and optical coupling constants $D_t k$, see for instance [82] and the references therein. Finally, we will show a comparison of our results in this case to multigroup WENO results as in [49]. This numerical scheme is based on the

use of the cell average method for treating the dependence of the electron distribution function on the three-dimensional wave vector and a fifth-order WENO solver for dealing with the physical space variables.

3.2 Pointwise WENO time splitting scheme for the BP equation

Let us first reduce the BP system to dimensionless cartesian coordinates. We assume that the doping profile, the potential and thus the force field are only x-dependent in space and thus, our device spans over the x-direction. Let us use the following adimensionalization of the BP system:

adim.	parameter	400 nm device	50 nm device
$\tilde{k} = k^* k$	$k^* = \frac{\sqrt{2m^* k_b T_L}}{\hbar}$	$4.65974 \times 10^8 m^{-1}$	$4.65974 \times 10^8 m^{-1}$
$\tilde{x} = l^* x$	$l^* =$ device length	$1\ \mu m$	$250\ nm$
$\tilde{t} = t^* t$	$t^* =$ typical time	$1\ ps = 10^{-12} s$	$1\ ps = 10^{-12} s$
$\tilde{V}(\tilde{x}) = V^* V(x)$	$V^* =$ typical Vbias	$1V$	$1V$
$\tilde{E}(\tilde{x}) = E^* E(x)$	$E^* = \frac{1}{10}\frac{V^*}{l^*}$	$100000 V m^{-1}$	$400000 V m^{-1}$
$\tilde{\varepsilon}(\tilde{k}) = \varepsilon^* \varepsilon(k)$	$\epsilon^* = \frac{\hbar^2 k^{*2}}{2m^*}$	$4.14195e-21$	$4.14195e-21$
$\tilde{\rho}(\tilde{x}) = \rho^* \rho(x)$	$\rho^* = \left(\frac{2m^* k_B T_L}{\hbar}\right)^{3/2}$	1.01178×10^{26}	1.01178×10^{26}
$\tilde{j}(\tilde{x}) = j^* j(x)$	$j^* = \frac{1}{l^{*2} t^*}$	10^{24}	1.6×10^{25}
$\tilde{u}(\tilde{x}) = u^* u(x)$	$u^* = \frac{l^*}{t^*}$	10^6	250000
$\tilde{W}(\tilde{x}) = W^* W(x)$	$W^* = (l^*/t^*)^2$	10^{12}	6.25×10^{10}

where tildes are written over dimensional magnitudes. Numerical values for all the parameters and the constants involved in the computations, as well as a resumé of all the dimensionless equations, can be found in the appendix. The BP equation transforms into

$$\frac{\partial f}{\partial t} + c_x \frac{\partial \varepsilon}{\partial k_1}\frac{\partial f}{\partial x} - c_k E_x \frac{\partial f}{\partial k_1} = \mathcal{Q}[f] \tag{3.11}$$

where the dimensionless parameters are

$$c_x = \frac{t^* \varepsilon^*}{\hbar k^* l^*}, \qquad c_k = \frac{q t^* E^*}{\hbar k^*}.$$

The electrostatic field is self-consistently computed by the rescaled Poisson's equation

$$\frac{\partial^2 \Phi}{\partial x^2} = c_p \left[\rho(t,x) - N_D(x)\right], \qquad c_p = \frac{q\rho^* l^{*2}}{\epsilon \Phi^*},$$

coupled with appropriate boundary values ($\Phi(0) = 0$, $\Phi(L) =$ Vbias).

The advantage of conserving the cartesian structure is that, thanks to the time splitting techniques [98, 35], we can apply the semi-lagrangian based Flux Balance Method [46, 26] to solve each transport step, which would be

more involved in the energy-band adapted coordinates [28], the energy flux resulting to be energy-dependent. On the other hand, we have to deal with a more complicated computation of the collisional part: instead of being a simple evaluation, like in [28] we shall need to reconstruct the values of the probability density f along a circle in the k-dimension.

In order to integrate the collisional part, we use that f only depends on k_1 and $k_{23} = \|(k_2, k_3)\| = \sqrt{k_2^2 + k_3^2}$, i.e. $f(k_1, k_2, k_3) = f(k_1, k_{23})$ due to symmetry considerations for this one-dimensional device. Using a change to polar coordinates in the (k_2, k_3)-plane, after straightforward computations, we obtain for the gain and the loss part of the collision operator $\mathcal{Q}(f)$ the following expressions:

$$\begin{aligned}\mathcal{Q}^+[f] = \; & c_0\,\pi \int_{-\sqrt{\gamma_0(k)}}^{\sqrt{\gamma_0(k)}} f\left(k'_1, \sqrt{\gamma_0(k) - k'^2_1}\right) dk'_1 \\ & + c_+\pi \int_{-\sqrt{\gamma_+(k)}}^{\sqrt{\gamma_+(k)}} f\left(k'_1, \sqrt{\gamma_+(k) - k'^2_1}\right) dk'_1 \\ & + \chi_{\{\gamma_-(k)>0\}} c_-\pi \int_{-\sqrt{\gamma_-(k)}}^{\sqrt{\gamma_-(k)}} f\left(k'_1, \sqrt{\gamma_-(k) - k'^2_1}\right) dk'_1 \qquad (3.12)\end{aligned}$$

with $\gamma_0(k) = \varepsilon(k)$, $\gamma_+(k) = \varepsilon(k) + \dfrac{h\omega}{\varepsilon^*}$, $\gamma_-(k) = \varepsilon(k) - \dfrac{\hbar\omega}{\varepsilon^*}$, and

$$\mathcal{Q}^-[f] = c_0 2\pi\sqrt{\gamma_0(k)} f(k) + \chi_{\{\gamma_-(k)>0\}} c_+ 2\pi\sqrt{\gamma_-(k)} f(k) + c_- 2\pi\sqrt{\gamma_+(k)} f(k),$$

with the dimensionless parameters

$$c_0 = \frac{K_0 t^* k^{*3}}{\varepsilon^*}, \qquad c_+ = \frac{K t^* (n_q + 1) k^{*3}}{\varepsilon^*}, \qquad c_- = \frac{K t^* n_q k^{*3}}{\varepsilon^*}.$$

In the next subsections, we shall explain in detail both the transport and the collision steps in this method. Let us finally comment that Poisson equation is solved through a standard centered finite differences leading to solving a linear system with a tridiagonal matrix.

3.2.1 Numerical scheme

Equation (3.11) is solved through a time splitting scheme dividing the system into the solution of transport steps and collision steps being the time stepping fixed. The computational domain is discretized into a tensor product mesh, and a uniform mesh is taken in each direction:

$$\begin{aligned}& x_i = i\Delta x, && i = 0, ..., N_x - 1, && \Delta x = \tfrac{1}{N_x - 1} \\ & (k_1)_j = -\varepsilon^{-1}(\alpha\bar{N}) + j\Delta k_1, && j = 0, ..., N_{k_1} - 1, && \Delta k_1 = \tfrac{2\varepsilon^{-1}(\alpha\bar{N})}{N_{k_1} - 1} \\ & (k_{23})_k = k\Delta k_{23}, && k = 0, ..., N_{k_{23}} - 1, && \Delta k_{23} = \tfrac{\varepsilon^{-1}(\alpha\bar{N})}{N_{k_{23}} - 1} \\ & t_n = n\Delta t, \end{aligned}$$

where α is the dimensionless energy $\alpha = \hbar\omega/k_B T_L$. Here, $\bar{N}$ is an integer number chosen as a maximal bound for the adimensionalized energy-band function

$$\varepsilon(k) = (k_1)^2 + (k_2)^2 + (k_3)^2 = (k_1)^2 + (k_{23})^2 = \varepsilon[k_1, k_{23}].$$

More precisely, at the value $\bar{N}\alpha$

$$\varepsilon\left[(k_1)_{N_{k_1}-1}, 0\right] = \bar{N}\alpha, \qquad \varepsilon\left[0, (k_{23})_{N_{k_{23}}-1}\right] = \bar{N}\alpha,$$

a magnitude which is related to the resolution in the (k_1, k_{23})-space.

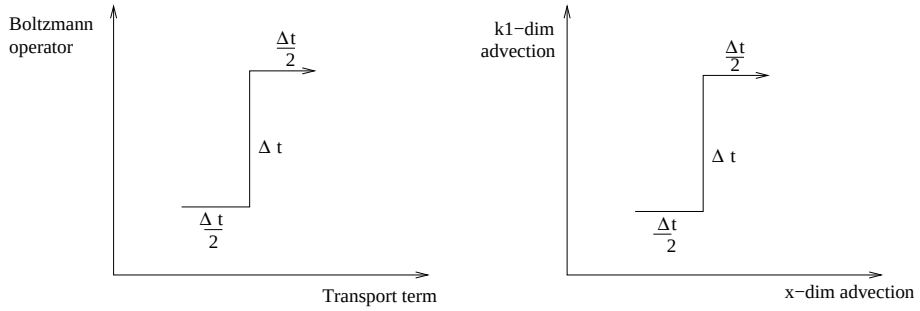


Figure 1: Time splitting scheme, see Appendix subsection 4.2 for a fully detailed splitting scheme.

The approximation denoted by $f^n_{i,j,k}$ to the point values of the solution $f\,[t_n, x_i, (k_1)_j, (k_{23})_k]$ are obtained through the second order time splitting scheme [35] subdividing the BP system (3.11):

- **Step 1.-** Solve $\dfrac{\partial f}{\partial t} + c_x \dfrac{\partial \varepsilon}{\partial k_1}\dfrac{\partial f}{\partial x} - c_k E_x \dfrac{\partial f}{\partial k_1} = 0$ for a $\frac{\Delta t}{2}$-time step;
- **Step 2.-** Solve $\dfrac{\partial f}{\partial t} = \mathcal{Q}[f]$ for a Δt-time step;
- **Step 3.-** Solve $\dfrac{\partial f}{\partial t} + c_x \dfrac{\partial \varepsilon}{\partial k_1}\dfrac{\partial f}{\partial x} - c_k E_x \dfrac{\partial f}{\partial k_1} = 0$ for a $\frac{\Delta t}{2}$-time step.

The same procedure is used for solving the two transport steps

$$\frac{\partial f}{\partial t} + c_x \frac{\partial \varepsilon}{\partial k_1}\frac{\partial f}{\partial x} - c_k E_x \frac{\partial f}{\partial k_1} = 0$$

by dimensional splitting. Therefore, we have subdivided the problem into the solution of the x-transport, the k-transport and the collision:

$$\frac{\partial f}{\partial t} + \underbrace{\overbrace{c_x \frac{\partial \varepsilon}{\partial k_1}\frac{\partial f}{\partial x}}^{\text{x-transport}} - \overbrace{c_k E_x \frac{\partial f}{\partial k_1}}^{\text{k-transport}}}_{\text{transport}} = \underbrace{\mathcal{Q}[f]}_{\text{collisions}}$$

as sketched in Figure 1 and fully specified in the Appendix, subsection 4.2.

Remark. Time Splittings In principle, the above time splitting procedure (TS) may seem unnecessary complicated due to the splitting of three different operators. Actually, the use of a direct first-order splitting, as reminded to the reader in subsection 4.3, would seem appropriate for simplicity. In section 4 we will compare the results for both splitting algorithms, and we will show that results given by the above splitting procedure improve the ones given by the first-order splitting.

Numerical Scheme: Transport Step.- Each transport block is solved by the Flux Balance Method [46, 26]: when solving the x-transport, k_1 and k_{23} act as parameters, as well as x and k_{23} when solving the k_1-transport. This method is based on the semi-lagrangian approach of following the characteristics backwards; the improvement is that we force the mass conservation, unlike the direct method, which gives no guarantee about this point. The solution of the x-transport gives

$$\begin{aligned} f^{**}_{i,j,k} = f^{*}_{i,j,k} + \frac{1}{\Delta x} \big\{ &\big[F(x_{i-1/2}) - F(x_{i-1/2} - c_x \nabla_k \varepsilon \Delta t)\big] \\ &- \big[F(x_{i+1/2}) - F(x_{i+1/2} - c_x \nabla_k \varepsilon \Delta t)\big] \big\} \\ F(x) = \int_0^x f^* \left[\xi, (k_1)_j, (k_{23})_k\right] d\xi \end{aligned}$$

and, as for the solution of the k_1-transport,

$$\begin{aligned} f^{**}_{i,j,k} = f^{*}_{i,j,k} + \frac{1}{\Delta k_1} \big\{ &\big[F((k_1)_{j-1/2}) - F((k_1)_{j-1/2} + c_k E \Delta t)\big] \\ &- \big[F((k_1)_{i+1/2}) - F((k_1)_{i+1/2} + c_k E \Delta t)\big] \big\} \\ F(k_1) = \int_0^{k_1} f^* \left[x_i, \xi, (k_{23})_k\right] d\xi. \end{aligned}$$

More details about the FBM method can be found in [46, 26]. In order to compute the fluxes, for instance,

$$F(x_{i+1/2}) - F(x_{i+1/2} - c_x \nabla_k \varepsilon)$$

we reconstruct the values $F(x_{i+1/2} - c_x \nabla_k \varepsilon)$, given the known values of the primitive at the grid points $F(x_{i+1/2})$, by the fifth order Pointwise WENO-6,4 interpolation summarized in next subsection.

Numerical Scheme: Collision Step.- In order to solve the collision step, we need to compute some integrals along semicircles of radius $\gamma_0(k)$, $\gamma_+(k)$ and $\gamma_-(k)$ in the (k_1, k_{23})-space. Section 1.4.1 explains the way we perform it.

We can now explain the implemented boundary conditions:

- at $x = 0$ and $x = L$ we use the following inflow/outflow condition:

$$f^n_{-i,j,k} = \begin{cases} f^n_{0,j,k} & k_1 < 0 \\ \frac{N_D(0)}{\rho(0)} f^n_{0,j,k} & k_1 \geq 0 \end{cases}$$

and

$$f^n_{N_x-1+i,j,k} = \begin{cases} f^n_{N_x-1,j,k} & k_1 > 0 \\ \frac{N_D(L)}{\rho(L)} f^n_{N_x-1,j,k} & k_1 \leq 0 \end{cases}$$

in order to have the ghost points we need for the PWENO interpolation and to preserve the correct values of the distribution function at the drain and the source of the diode;

- at $k_1 = -\varepsilon^{-1}(\alpha\bar{N})$ and $k_1 = \varepsilon^{-1}(\alpha\bar{N})$ a Neumann type boundary condition in used:

$$f^n_{i,-j,k} = f^n_{i,0,k} \qquad \text{and} \qquad f^n_{i,N_{k_1}-1+j,k} = f^n_{i,N_{k_1}-1,k}.$$

While in the transport steps the mass conservation is guaranteed by the Flux Balance Method being conservative, during the collision steps we numerically impose the mass conservation by redefining the collision operator by

$$\mathcal{Q}[f] = \mathcal{Q}^+[f] - \frac{\int_{\mathbb{R}^3} \mathcal{Q}^+[f]dk}{\int_{\mathbb{R}^3} \mathcal{Q}^-[f]dk} \mathcal{Q}^-[f].$$

3.3 Numerical Experiments

3.3.1 Steady-state results for the diodes

We consider two test examples: Si $n^+ - n - n^+$ diodes of total length of $1\mu m$ and $0.25\mu m$, with $400nm$ and $50nm$ channels located in the middle of the device respectively. For the $400nm$ device and the $50nm$ device the dimensional doping is, respectively,

$$N_D = \begin{cases} 5 \times 10^{17}\ cm^{-3} & n^+\text{-zone} \\ 2 \times 10^{15}\ cm^{-3} & n\text{-zone} \end{cases} \quad \text{and} \quad \begin{cases} 5 \times 10^{18}\ cm^{-3} & n^+\text{-zone} \\ 1 \times 10^{15}\ cm^{-3} & n\text{-zone} \end{cases}.$$

The results provided by the W5FD method [28] and our PW5TS are compared as for the macroscopic magnitudes (density, electrostatic potential, electrostatic field, mean velocity, energy and current), as we can see in Figures 2, 3, 4, 5, 6, 7 and 8. In the W5FD scheme the kinetic variable ω denotes the dimensionless electron energy and μ the cosine of the angle between the wave vector k and the x-axis.

The comparisons are set in such a way that we can infer how the choice of the different parameters of our scheme affects the results compared to

the W5FD results chosen as benchmarks. Five issues have been considered: order of the time splitting procedure, the energy cut-off $\bar{N}$, the type of interpolation chosen for computing the collision operator, the resolution in k and the time step Δt. Detailed comparisons of these parameters are shown for the 400 nm diode, whereas certain comparisons are drawn for the 50 nm diode.

One advantage of the proposed method is that we have no restriction for the time stepping, thus in both cases we can reach the equilibrium (5 ps for the 400 nm diode, 2 ps for the 50 nm diode) by largely less time-steps than in [28], where several thousands were needed. We have empirically searched for the largest time steppings by which no instabilities appear; for the 400 nm diode it seems to be around $\Delta t = 0.07$ ps (so by just about 70 steps we reach the equilibrium), for the 50 nm diode around $\Delta t = 0.04$ ps. As for the Finite Differences scheme, the adaptive time stepping situates between 10^{-3} and 10^{-4} ps.

The choice of shorter time steppings and better resolution in k due to finer grids or larger cut-off energy $\bar{N}$ improves the quality of the results, see Figures 4, 5, 6, 7 and 8; of course, the counterpart is that it increases the computational cost. The loss of reliability is evident when we increase the time stepping of the code if we look at the current and the mean velocity in Figure 4, where the oscillations are amplified, even if the density, the electric potential and the electric field remain very close. Thus, unfortunately the improvement in the time stepping does not always translate into a shorter computational time, a better resolution in the k-dimension with larger grids and better adapted cut-off energy $\bar{N}$ being needed in order to obtain reliable results.

All the above results have been obtained by using the direct linear interpolation for the collisional step and the time splitting procedure in subsection 2.1. Let us comment on this choice: we have tested and compared the direct linear interpolation and the PWENO-4,3 interpolations as discussed in subsection 2.1 together with the first-order splitting and the TS splitting procedure in subsection 2.1. In Figures 2 and 3 results are compared in terms of the current and the mean velocity at equilibrium. We first observe that first-order splitting results are in general worse than the results with time splitting procedure in subsection 2.1 and appendix 4.2. On the other hand, we observe that improving the accuracy of the interpolation procedure in the collision step from linear to PWENO-4,3 does not result in a marked gain of accuracy for these quantities. Even if results are not shown here, a simple two dimensional linear, in each variable not jointly, interpolation in each quadrangle of the cartesian grid has been performed. Again, this improvement in the interpolation accuracy does not yield a significant gain in the accuracy of the macroscopic quantities. This collisional step will need further improvements or alternative methods as spectral approaches [44] before being able to cope with two dimensional devices in comparison

to the efficiency of W5FD [29].

On the other hand, having a finer grid in the k-dimension, moreover in cartesian coordinates, permits a better resolution of the distribution function, as we can observe in Figure 9 and 10: like in [28], the pdf outside the channel is close to a Maxwellian distribution. Inside the channel it looks like a shifted Maxwellian in the large diode, while in the small one it assumes a very asymmetric shape. Moreover, in this small diode we observe the formation of a narrow ballistic pick. A good resolution of this narrow pick involves a very fine grid in k-space and a much larger computational cost. Its underresolution is the cause of the difference and oscillations in mean velocity and current between the W5FD method [28] and our PW5TS observed in Figure 8. Therefore a energy-based variables solver as the one in [28] gives better results in this case.

Finally, let us point out that previous results have been obtained looking at the stabilization in time of the macroscopic quantities. For instance, in the 400-nm diode case, runs have been performed till 5ps for which the density is stabilized up to 10^{-6}. A stabilization of the other macroscopic quantities: current, mean velocity and energy needs longer runs till 10ps approx. Typical problems in reaching numerical steady states occur for splitting in time numerical strategies. In our case, the main issue is that the results stabilize numerically to states in which the current is not constant as it should be for the stationary case. We can observe this problem in the current comparison of Figures 6 and 8. An improvement in the numerical approximation of the collisional step will certainly help to fix this problem.

3.3.2 Steady-state results in multifrequency phonons

With the method we have implemented it is easy to change the solver of the collision operator. Usually, phonons do not have a single frequency; in [28] this simplification was set in order to directly compute the collision operator without needing to perform interpolations. In this method, we just have to add as many interpolations as the frequencies are. In [82] they took into account six frequencies, for a diode of total length 600 nm, with a channel of 400 nm. The collision operator transforms into

$$\begin{aligned} S(k,k') = &\sum_{i=1}^{6} K_i \left[(n_{q_i}+1)\delta(\varepsilon(k') - \varepsilon(k) + \hbar\omega_i) + n_{q_i}\delta(\varepsilon(k') - \varepsilon(k) - \hbar\omega_i) \right] \\ &+ K_0 \delta(\varepsilon(k') - \varepsilon(k)) \end{aligned} \tag{3.13}$$

where n_{q_i} are the occupation numbers of phonons

$$n_{q_i} = \frac{1}{\exp\left(\frac{\hbar\omega_i}{k_B T_L}\right) - 1}, \tag{3.14}$$

and the kernels K_i are

$$K_i = \frac{Z_f D_t k_i^2}{8\pi^2 \hbar \rho_0 \omega_i} \tag{3.15}$$

where $\hbar\omega_i$ and $D_t k_i$ are the energy and the deformation potentials of the corresponding phonon type.

Results are shown in Figure 11 and Figure 12. The following doping profile is set

$$N_D = \begin{cases} 5 \times 10^{17}\ cm^{-3} & n^+\text{-zone} \\ 2 \times 10^{15}\ cm^{-3} & n\text{-zone} \end{cases}$$

and the quantities related to the phonon frequencies are set

freq.	Z_f	$\hbar\omega\ (meV)$	$D_t K(10^8 eV/cm)$
1	1	12	0.5
2	1	18.5	0.8
3	4	19	0.3
4	4	47.4	2
5	1	61.2	11
6	4	59	2

where Z_f is the number of equivalent valleys.

In this case, the use of the numerical scheme in [28] becomes much more involved leading to very fine grids in energy variables and interpolations like in our case. Moreover, we show in Figures 11 the comparison of our results to the ones obtained with the a simple application of the MultiGroup-WENO solver which are quite satisfactory. In [49] this technique was applied in case of a single phonon frequency; here, we consider multifrequency phonons. This requires only some simple modifications of the collision operator. Since the MultiGroup scheme is based on the cell average with respect to the wave vector, the presence of many delta distributions in the collsion operator does not pose new difficulties. Also in this case our method allows for a good resolution of the pdf's in k-space as shown in Figure 12.

3.4 Appendix

3.4.1 Adimensionalization Summary

The BP system reads

$$\begin{cases} \dfrac{\partial f}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f - c_k E \cdot \nabla_k f = \mathcal{Q}[f] \\ f_0(x,k) = c_{init} N_D(x) M(k). \end{cases}$$

where the energy-band function becomes $\varepsilon(k) = c_\varepsilon |k|^2$, and the Maxwellian

$$M(k) = \left(\frac{\pi}{C_M}\right)^{-3/2} e^{-C_M k^2}.$$

The electrostatic field is self-consistently computed through the Poisson's equation

$$\Delta_x \Phi = c_p \left[\rho(t,x) - N_D(x)\right], \qquad E = -c_e \nabla_x \Phi$$

where the density is given by

$$\rho(t,x) = c_d \int_{\mathbb{R}^3} f(t,x,k) dk.$$

The gain and loss parts of the collision operator are

$$\mathcal{Q}^+[f] = \int_{\mathbb{R}^3} f(t,x,k) \left[c_0 \delta(\varepsilon(k) - \varepsilon(k')) + c_+ \delta\left(\varepsilon(k) - \varepsilon(k') + \frac{\hbar\omega}{\varepsilon^*}\right) + c_- \delta\left(\varepsilon(k) - \varepsilon(k') - \frac{\hbar\omega}{\varepsilon^*}\right)\right] dk'$$

and

$$\mathcal{Q}^-[f] = f(t,x,k) \int_{\mathbb{R}^3} \left[c_0 \delta(\varepsilon(k) - \varepsilon(k')) + c_+ \delta\left(\varepsilon(k) - \varepsilon(k') - \frac{\hbar\omega}{\varepsilon^*}\right) + c_- \delta\left(\varepsilon(k') - \varepsilon(k) + \frac{\hbar\omega}{\varepsilon^*}\right)\right] dk'$$

The current is the first momentum in the k_1 direction,

$$j(x) = c_j \int_{\mathbb{R}^3} k_1 f(k) dk,$$

the mean velocity is

$$u(x) = c_u \frac{j(x)}{\rho(x)},$$

and the energy is

$$W(x) = c_W \frac{1}{\rho(x)} \int_{\mathbb{R}^3} \varepsilon(k) f(k) dk.$$

The dimensionless parameters are derived from physical constants, the problem data and the adimensionalization parameters:

parameter	400 nm diode	50 nm diode
$c_\varepsilon = \frac{\hbar^2 k^{*2}}{2m^*\varepsilon^*}$	1	1
$c_x = \frac{t^*\varepsilon^*}{\hbar k^* l^*}$	0.0842885	0.337154
$c_k = \frac{qt^*E^*}{\hbar k^*}$	0.326042	1.30417
$c_d = \frac{k^{*3}}{\rho^*}$	1	1
$c_p = \frac{q\rho^* l^{*2}}{\epsilon V^*}$	156480	9780.02
$c_e = \frac{V^*}{l^* E^*}$	10	10
$c_{init} = \frac{1}{c_d}$	1	1
$C_M = \frac{\hbar^2 k^{*2}}{2m^* k_B T_L}$	1	1
$c_0 = \frac{K_0 t^* k^{*3}}{\varepsilon^*}$	0.265376	0.265376
$c_+ = \frac{Kt^*(n_q+1)k^{*3}}{\varepsilon^*}$	0.507132	0.507132
$c_- = \frac{Kt^* n_q k^{*3}}{\varepsilon^*}$	0.0443372	0.0443372
$c_j = \frac{\hbar k^{*4}}{m^* j^*}$	1.70562×10^7	1.06601×10^6
$c_u = \frac{j^*}{u^*\rho^*}$	9.88362×10^{-9}	6.32552×10^{-7}
$c_W = \frac{k^{*3}\varepsilon^*}{\rho^* W^*}$	4.14195×10^{-33}	6.62712×10^{-32}

Physical constants involved in the solution of the BP problem are:

name	meaning	value
$\hbar$	Dirac's constant	$\frac{6.626068\times10^{-34}}{2\pi}$ $= 1.05456 \times 10^{-34}\frac{m^2 Kg}{s}$
q	elementary charge	1.60217646
m^*	effective electron mass = $= 0.32 \times$ electron mass	$0.32 \times 9.10938188 \times 10^{-31} Kg$ $= 2.915 \times 10^{-31} Kg$
k_B	Boltzmann's constant	$1.3806503 \times 10^{-23}\frac{m^2 Kg}{s^2 K}$
u_l	sound velocity	$9040\frac{m}{s}$
ρ_0	Si crystal density	$2330\frac{Kg}{m^3}$
ϵ_0	vacuum dielectric permittivity	$8.85419 \times 10^{-12}\frac{F}{m}$
ϵ_r	Si relative permittivity	11.7
ϵ	Si dielectric permittivity	$1.0359402 \times 10^{-10}\frac{F}{m}$
F	Farad	$F = \frac{s^4 A^2}{m^2 Kg} = \frac{C}{V}$

Finally, the problem data are:

$$\begin{cases} \omega = \text{frequency} = \frac{0.063eV}{\hbar} \\ D_t k = \text{optical coupling frequency} = 11.4 \times 10^{10}\frac{eV}{m} \\ K_0 = \frac{k_B T_L {E_{ac}}^2}{4\pi^2 h u_l^2 \rho_0} = 1.08638 \times 10^{-35} \\ E_{ac} = \text{deformation potential} = 9\ eV \\ K = \frac{D_t k^2}{8\pi^2 \rho_0 \omega} = 1.89456 \times 10^{-35} \\ T_L = \text{lattice temperature} = 300K \end{cases}$$

3.4.2 Time Splitting Scheme

Combining all the time splittings, we get the following scheme: given the distribution function $f^n = f(t = t^n)$, we update f up to f^{n+1} by the following successive steps:

1. perform a $\frac{\Delta t}{2}$ step for transport

$$\frac{\partial f^n}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^n - c_E \cdot \nabla_k f^n = 0 :$$

   1.1 perform a $\frac{\Delta t}{4}$ step for x-transport

$$\frac{\partial f^n}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^n = 0;$$

$f^{n+1/7} =$

x-transport	k-transport	collision
$\Delta t/4$	0	0

   1.2 perform a $\frac{\Delta t}{2}$ step for k-transport

$$\frac{\partial f^{n+1/7}}{\partial t} - c_E \cdot \nabla_k f^{n+1/7} = 0;$$

$f^{n+2/7} =$

x-transport	k-transport	collision
$\Delta t/4$	$\Delta t/2$	0

   1.3 perform a $\frac{\Delta t}{4}$ step for x-transport

$$\frac{\partial f^{n+2/7}}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^{n+2/7} = 0$$

$f^{n+3/7} =$

x-transport	k-transport	collision
$\Delta t/2$	$\Delta t/2$	0

2. perform a Δt step for collisions

$$\frac{\partial f^{n+3/7}}{\partial t} = \mathcal{Q}[f^{n+3/7}]$$

$f^{n+4/7} =$

x-transport	k-transport	collision
$\Delta t/2$	$\Delta t/2$	Δt

3. perform a $\frac{\Delta t}{2}$ step for transport

$$\frac{\partial f^{n+4/7}}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^{n+4/7} - c_E \cdot \nabla_k f^{n+4/7} = 0 :$$

3.1 perform a $\frac{\Delta t}{4}$ step for x-transport

$$\frac{\partial f^{n+4/7}}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^{n+4/7} = 0;$$

$f^{n+5/7} =$

x-transport	k-transport	collision
$\frac{3}{4}\Delta t$	$\Delta t/2$	Δt

3.2 perform a $\frac{\Delta t}{2}$ step for k-transport

$$\frac{\partial f^{n+5/7}}{\partial t} - c_E \cdot \nabla_k f^{n+5/7} = 0;$$

$f^{n+6/7} =$

x-transport	k-transport	collision
$\frac{3}{4}\Delta t$	Δt	Δt

3.3 perform a $\frac{\Delta t}{4}$ step for x-transport

$$\frac{\partial f^{n+6/7}}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^{n+6/7} = 0$$

$f^{n+1} =$

x-transport	k-transport	collision
Δt	Δt	Δt

3.4.3 First order time Splitting Scheme

The first order time splitting scheme reads:

- **Step 1.-** Solve $\frac{\partial f}{\partial t} + c_x \frac{\partial \varepsilon}{\partial k_1} \frac{\partial f}{\partial x} - c_k E_x \frac{\partial f}{\partial k_1} = 0$ for a Δt-time step;
- **Step 2.-** Solve $\frac{\partial f}{\partial t} = \mathcal{Q}[f]$ for a Δt-time step,

combined with a splitting of the same order for the solution of the transport part.

The scheme we obtain is:

1. perform a Δt step for transport

$$\frac{\partial f^n}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^n - c_E \cdot \nabla_k f^n = 0 :$$

1.1 perform a Δt step for x-transport

$$\frac{\partial f^n}{\partial t} + c_x \nabla_x \varepsilon \cdot \nabla_x f^n = 0;$$

$f^{n+1/3} =$

x-transport	k-transport	collision
Δt	0	0

1.2 perform a Δt step for k-transport

$$\frac{\partial f^{n+1/3}}{\partial t} - c_E \cdot \nabla_k f^{n+1/3} = 0;$$

$f^{n+2/3} =$

x-transport	k-transport	collision
Δt	Δt	0

2. perform a Δt step for collisions

$$\frac{\partial f^{n+2/3}}{\partial t} = \mathcal{Q}[f^{n+2/3}]$$

$f^{n+1} =$

x-transport	k-transport	collision
Δt	Δt	Δt

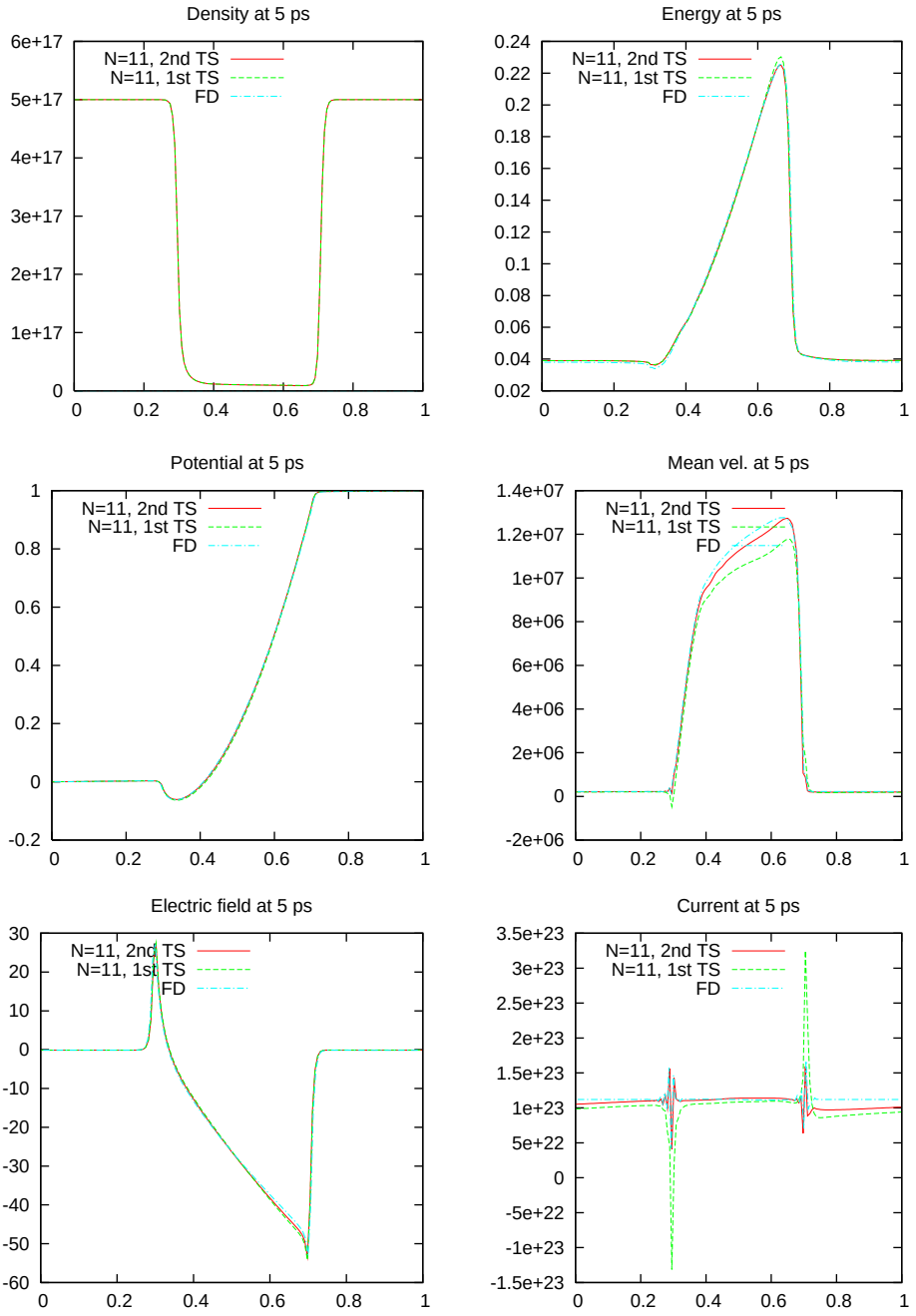


Figure 2: Comparison between some macroscopic quantities of the 400 nm diode at equilibrium (5 ps) given by splitting schemes of order 1 and 2. Top left: density in cm^{-3}; top right: energy in eV; center left: potential in V; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 V m^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 40 \times 16$ for (x, ω, μ) for the W5FD method, $150 \times 71 \times 71$ for $(x, k_1, k_{23} = \|(k_2, k_3)\|)$, $\bar{N} = 11$, $\Delta t = 0.01\ ps$, linear interpolation for collisions, for the PW5TS method.

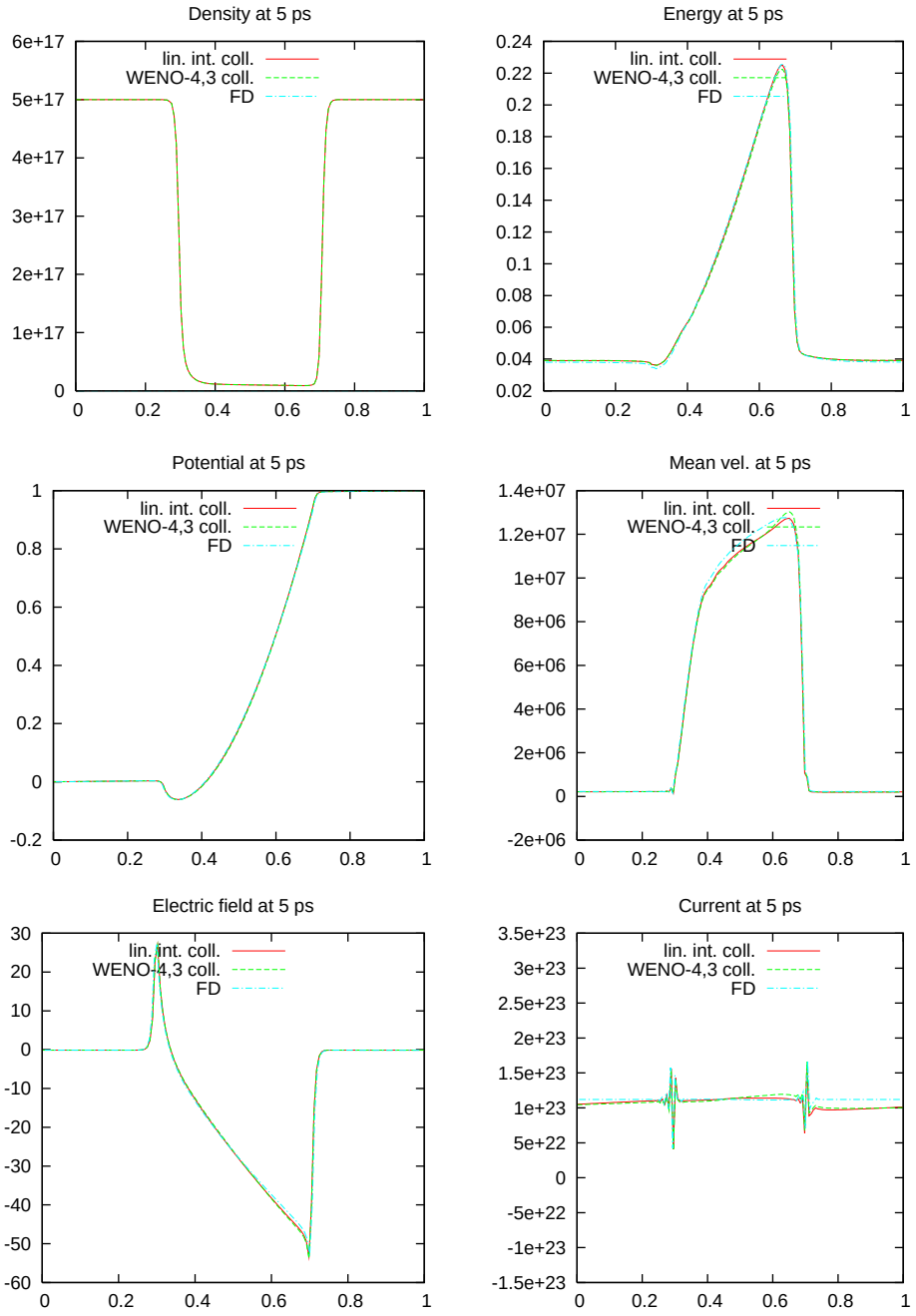


Figure 3: Comparison between some macroscopic quantities of the 400 nm diode at equilibrium (5 ps) given by different integrations of the collisions, wither by linear interpolation or by PWENO-4,3 interpolation. Top left: density in cm^{-3}; top right: energy in eV; center left: potential in V; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 Vm^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 40 \times 16$ for (x, ω, μ) for the W5FD method, $150 \times 71 \times 71$ for $(x, k_1, k_{23} = \|(k_2, k_3)\|)$, $\bar{N} = 11$, $\Delta t = 0.01\ ps$, 2nd order TS, for the PW5TS method.

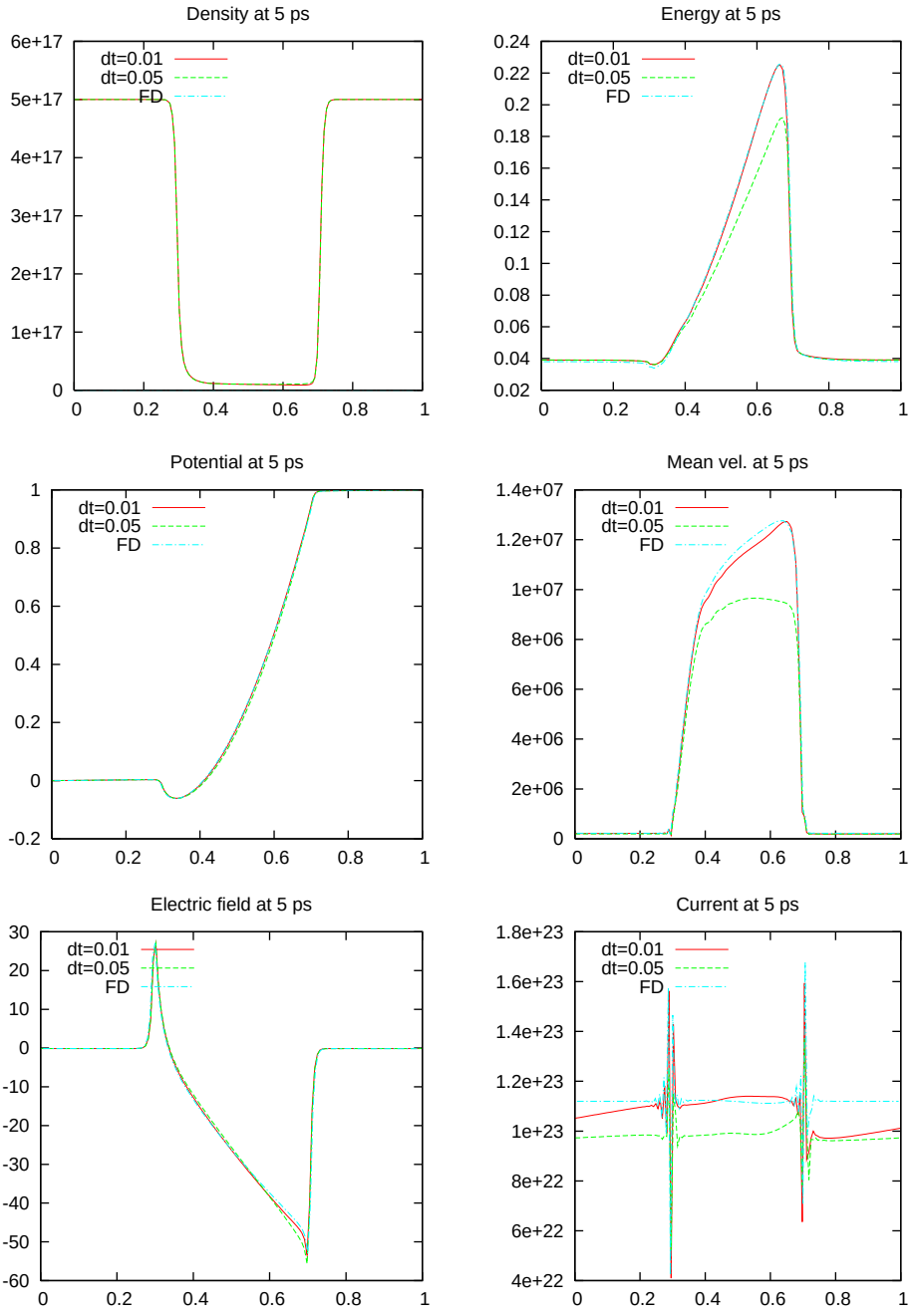


Figure 4: Comparison between some macroscopic quantities of the 400 nm diode at equilibrium (5 ps) given by different time steppings. Top left: density in cm^{-3}; top right: energy in eV; center left: potential in V; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 V m^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 40 \times 16$ for (x, ω, μ) for the W5FD method, $150 \times 71 \times 71$ for $(x, k_1, k_{23} = \|(k_2, k_3)\|)$, $\bar{N} = 11$, linear interpolation for collisions, 2nd order TS, for the PW5TS method.

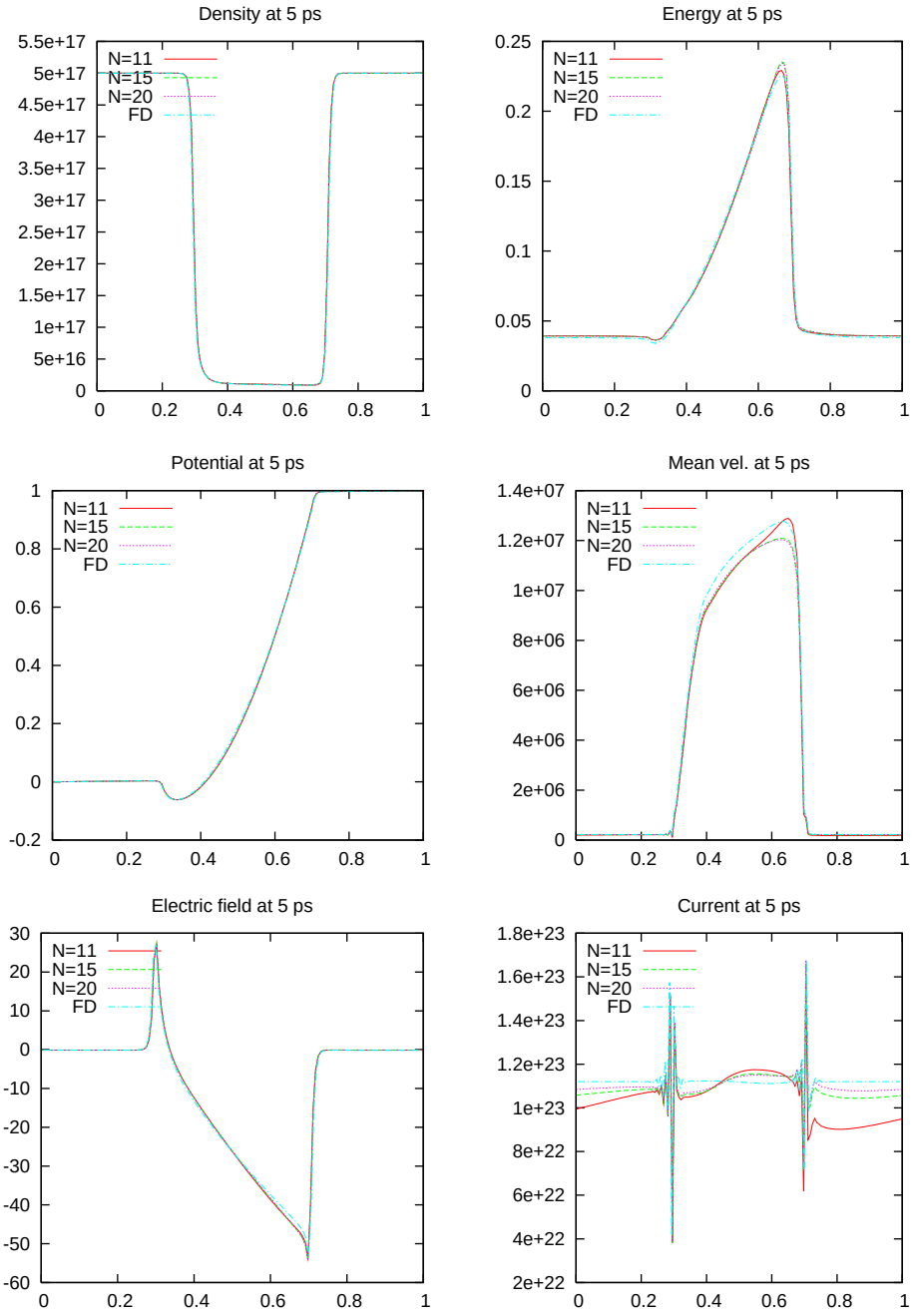


Figure 5: Comparison between some macroscopic quantities of the 400 nm diode at equilibrium (5 ps) given by different $\bar{N}$. Top left: density in cm^{-3}; top right: energy in eV; center left: potential in V; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 V m^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 40 \times 16$ for (x, ω, μ) for the W5FD method, $150 \times 64 \times 64$ for $(x, k_1, k_{23} = \|(k_2, k_3)\|)$ (when $\bar{N} = 10$), $\Delta t = 0.01\ ps$, linear interpolation for collisions, 2nd order TS, for the PW5TS method.

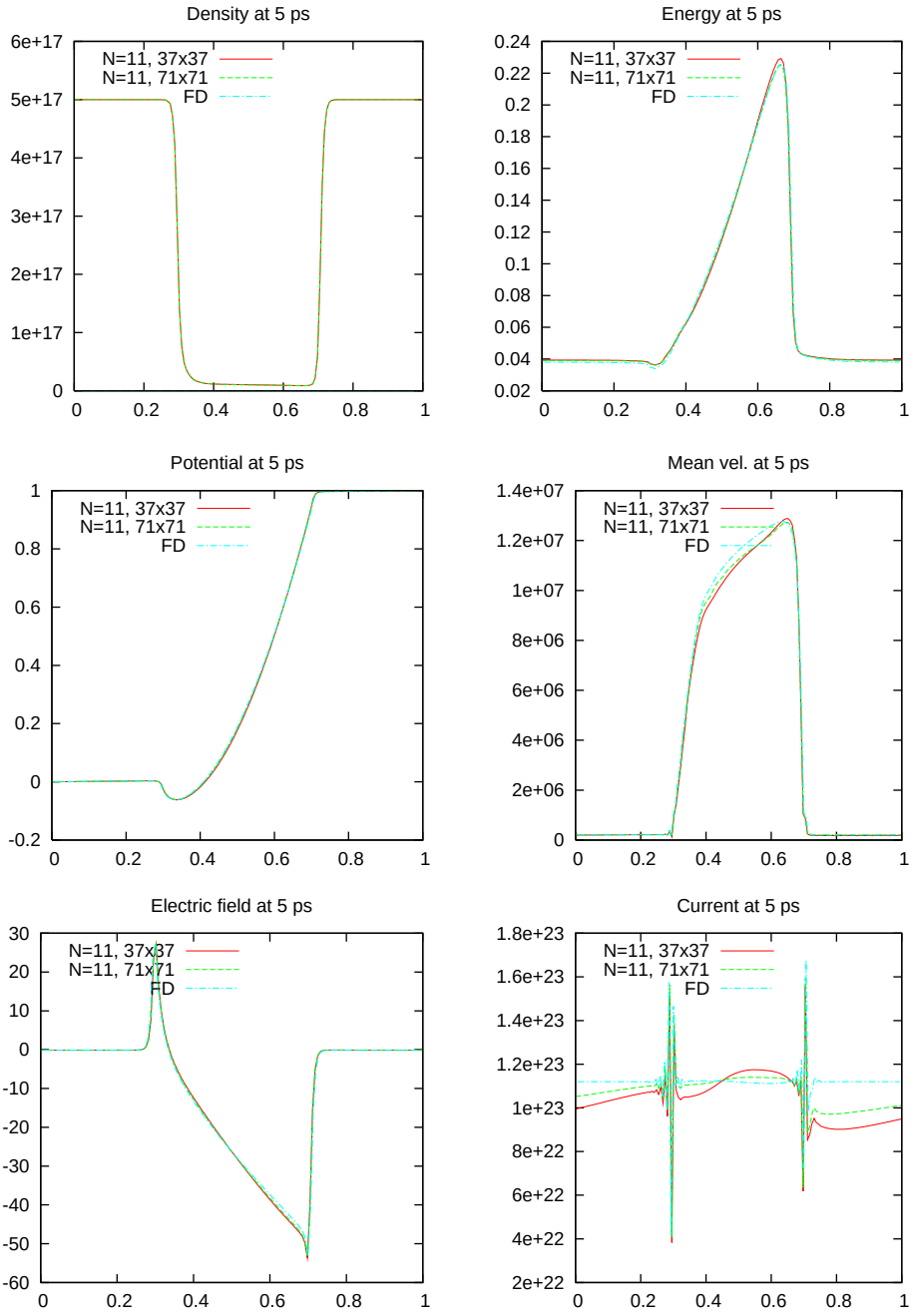


Figure 6: Comparison between some macroscopic quantities of the 400 nm diode at equilibrium (5 ps) given by different resolutions of the (k_1, k_{23})-grid. Top left: density in cm^{-3}; top right: energy in eV; center left: potential in V; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 V m^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 40 \times 16$ for (x, ω, μ) for the W5FD method, $\bar{N} = 11$, $\Delta t = 0.01\ ps$, linear interpolation for collisions, 2nd order TS, for the PW5TS method.

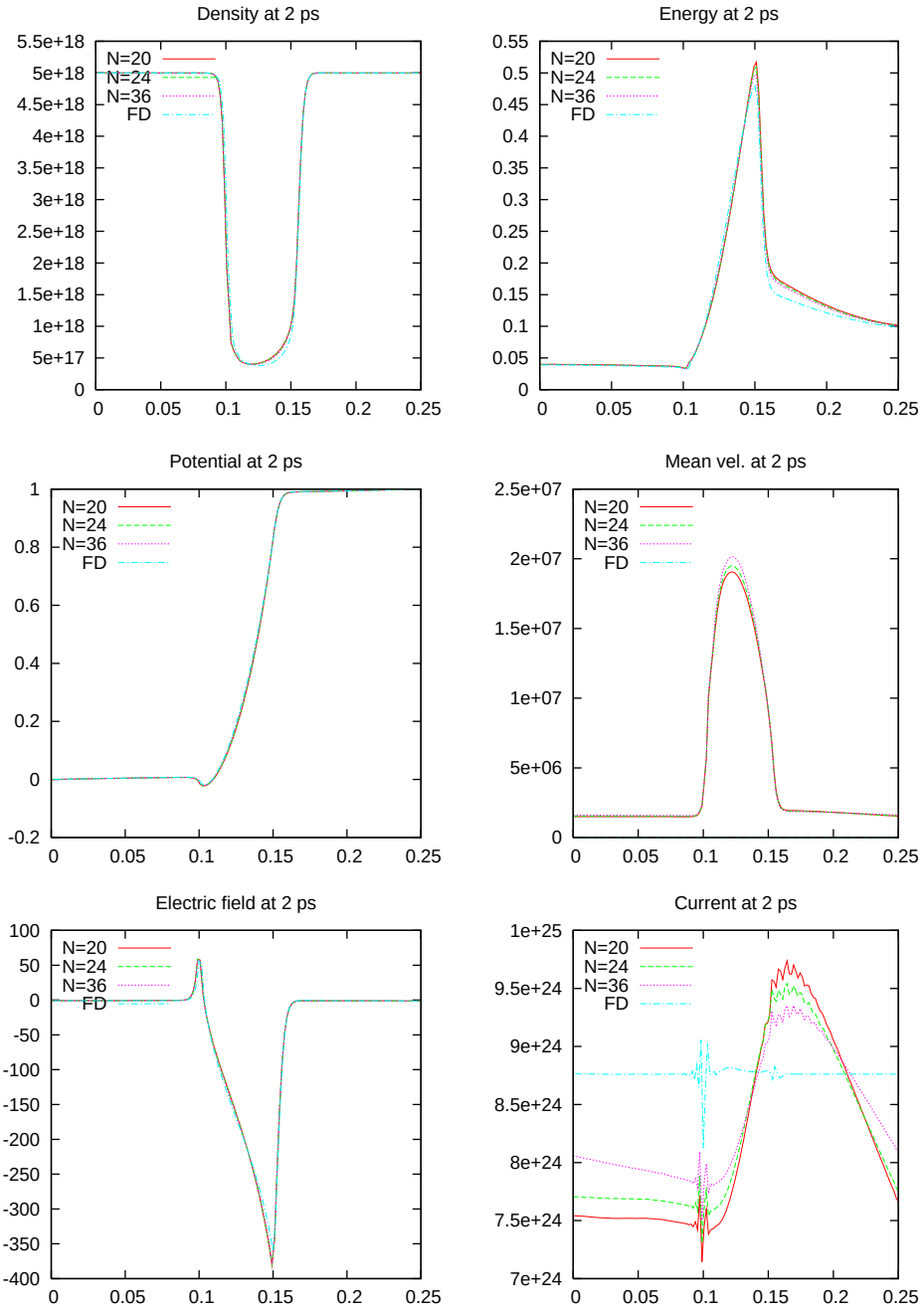


Figure 7: Comparison between some macroscopic quantities of the 50 nm diode at equilibrium (2 ps) given by different $\bar{N}$. Top left: density in cm^{-3}; center left: potential in V; top right: energy in eV; center right: mean velocity in $cm\ s^{-1}$; bottom left: electric field in $10^3 V m^{-1}$; bottom right: current in $cm^{-2}s^{-1}$. Grids are set $150 \times 144 \times 16$ for (x, ω, μ) for the W5FD method, $150 \times 32 \times 32$ for $(x, k_1, k_{23} = \|(k_2, k_3)\|)$ (when $\bar{N} = 18$), $\Delta t = 0.01\ ps$, linear interpolation for the collisions, 2nd order TS, for the PW5TS method.

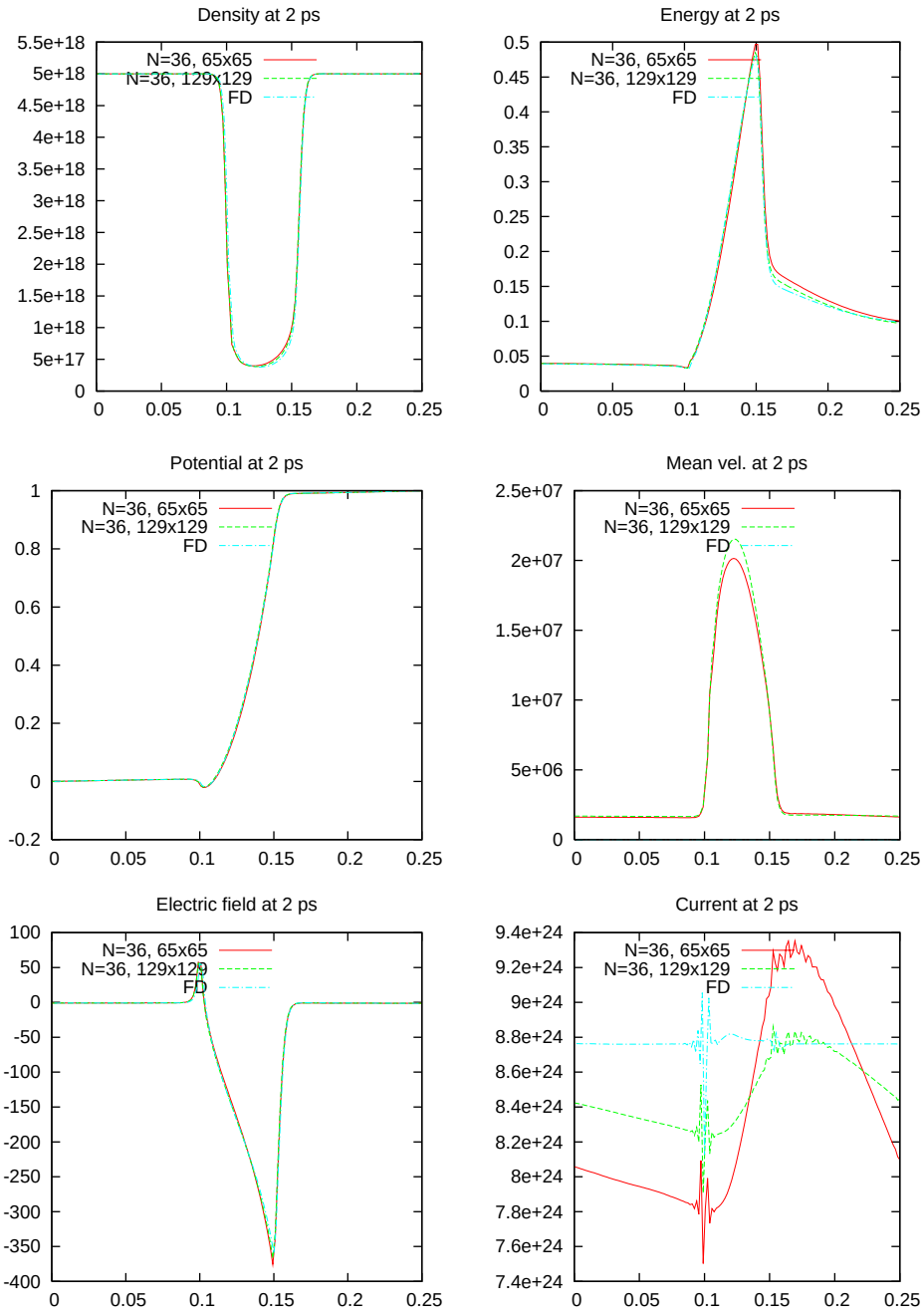


Figure 8: Comparison between some macroscopic quantities of the 50 nm diode at equilibrium (2 ps) given by different resolutions of the (k_1, k_{23})-grid. The grid is set $150 \times 144 \times 16$ for (x, ω, μ) for the W5FD method. $\bar{N} = 20$, $\Delta t = 0.01$ ps, linear interpolation for the collisions, 2nd order TS, for the PW5TS method.

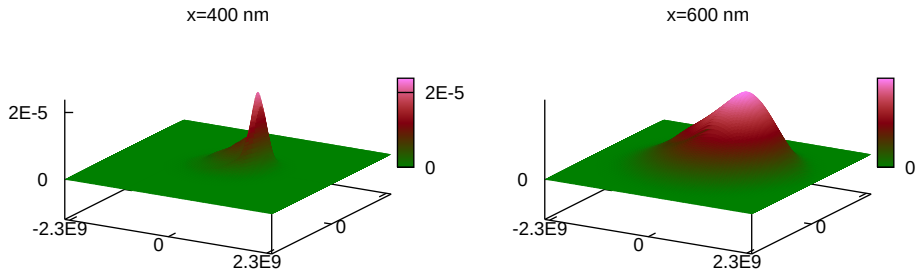


Figure 9: Distribution function of the 400 nm diode at time 5 ps given by the PW5TS method at different points of the device, for a $150 \times 71 \times 71$ grid, $\bar{N} = 11$, $\Delta t = 0.01$ ps, 2nd order TS, linear interpolation for collisions.

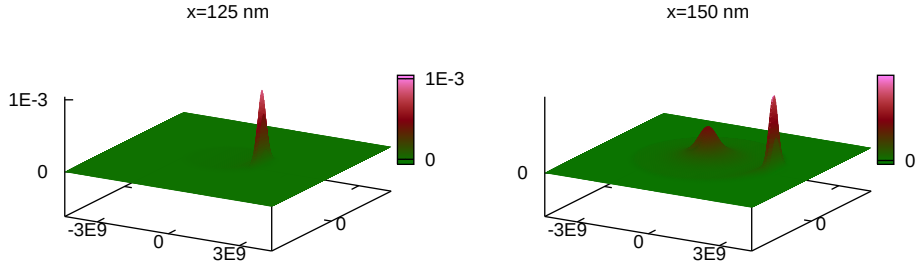


Figure 10: Distribution function of the 50 nm diode at time 2 ps given by the PW5TS method at different points of the device, for a $150 \times 129 \times 129$ grid, $\bar{N} = 36$, $\Delta t = 0.01$ ps, 2nd order TS, linear interpolation for collisions.

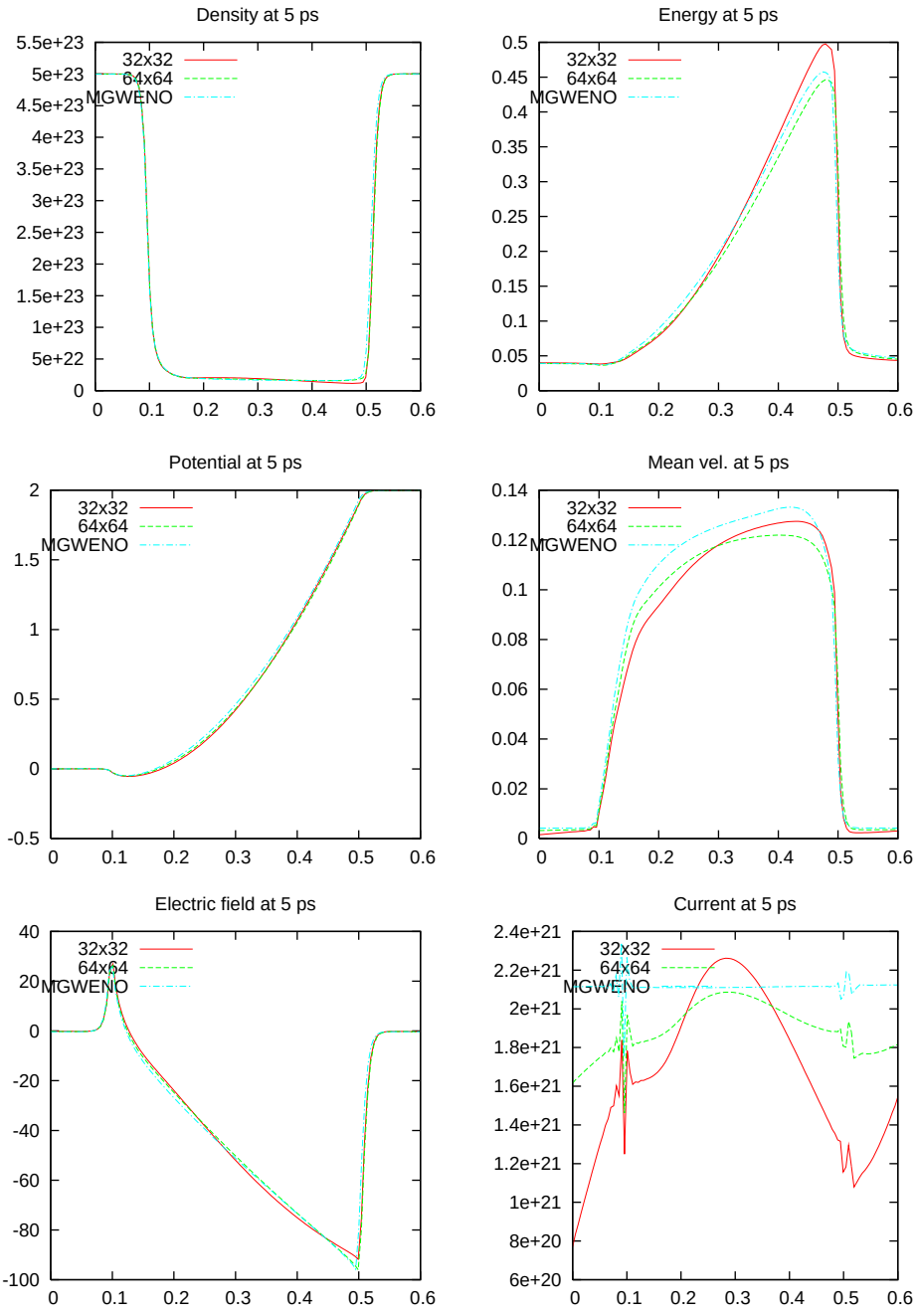


Figure 11: Macroscopic magnitudes given by the PW5TS method and a reference result obtained through a Multi-Group WENO scheme. For the PW5TS simulation The k-resolution is set $\bar{N} = 27$, the time stepping is set $\Delta t = 0.01$ ps, 2nd order TS, linear interpolation for collisions.

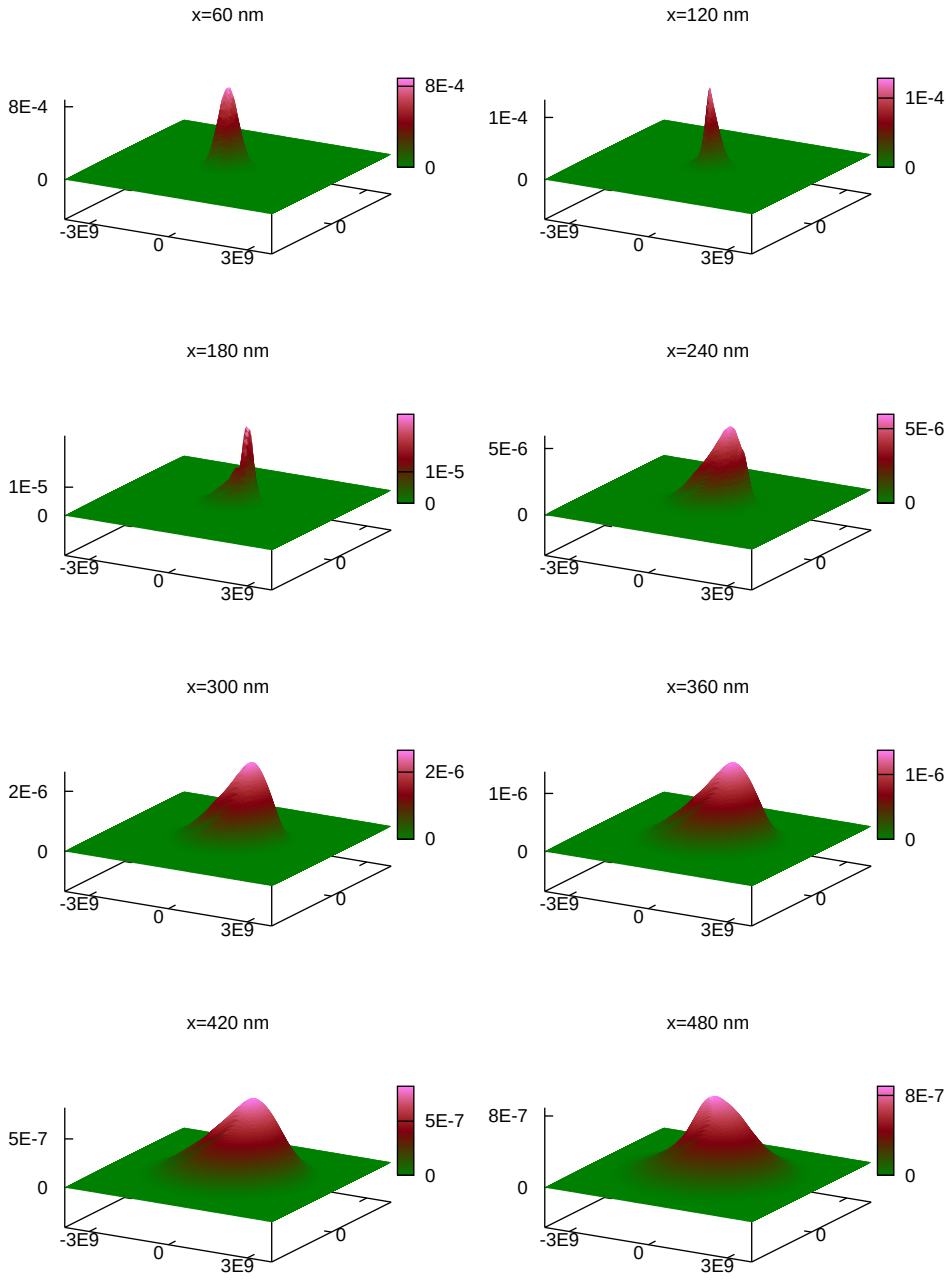


Figure 12: Snapshots given by the PW5TS method at different points of the multifrequency device, at time 5 ps. The grid is set $120 \times 64 \times 64$, $\bar{N} = 27$, $\Delta t = 0.01$ ps, 2nd order TS, linear interpolation for collisions.

Chapter 4

Numerical Schemes of Diffusion Asymptotics and Moment Closures for Kinetic Equations

This article corresponds to a work [23] in collaboration with J.A. Carrillo, P. Lafitte and Th. Goudon whose reference is: "Numerical schemes for Diffusion Asymptotics and Moment Closures for Kinetic Equations", accepted for publication in Journal of Scientific Computing.

4.1 Introduction

We are interested in numerical simulations of "intermediate" models for kinetic equations in diffusion regimes. Such questions arise in many application fields where we adopt a statistical description of a large set of "particles": neutron transport in nuclear engineering, radiative transfer, rarefied gas dynamics... The unknown is the particle distribution function that gives the number of particles being at time t and position x in a certain physical state described by the variable v. In most of the applications, v is nothing but the translational velocity, or the direction of flight of the particles and, assuming that v belongs to a certain measured set $(V, \mathrm{d}\mu)$, the quantities of interest are essentially averages over v of the unknown. The evolution of the particles obeys the following equation

$$\varepsilon \partial_t f_\varepsilon + v \partial_x f_\varepsilon = \frac{1}{\varepsilon} Q(f_\varepsilon). \tag{4.1}$$

In the right hand side, the operator Q is intended to describe the interactions that particles are subject to; the dimensionless parameter $\varepsilon > 0$ is related to the mean free path, that is the average distance travelled by the particles

without being subject to any interaction. As $\varepsilon \to 0$ the unknown f_ε relaxes to an equilibrium the dependence of which with respect to v is fixed and the dynamics is described by the evolution of only macroscopic quantities. As we shall see below, it turns out that under some suitable assumptions on the collision operator Q, the limit equation reduces to a mere diffusion equation. However, for applications, one is interested in preserving more information concerning the microscopic setting that motivates the derivation of reduced $\varepsilon-$dependent models. Then, it is legitimate to address the following two-fold question: Is the reduced model consistent to the diffusion approximation? How accurate is the obtained approximation and in which sense is it better than the solution of the limit diffusion equation?

In this paper we investigate numerically these questions, restricting to the simplest situation. Namely, we only deal with the one-dimensional framework ($x \in \mathbb{R}$, $v \in V \subset \mathbb{R}$) and the collision operator is a mere relaxation operator

$$Q(f) = \int_V f \, \mathrm{d}\mu(v) - f. \tag{4.2}$$

Throughout the paper $(V, \mathrm{d}\mu)$ is required to satisfy

$$\begin{cases} \displaystyle\int_V \mathrm{d}\mu(v) = 1, \\ \text{for any odd integrable function } h : V \to \mathbb{R}, \quad \displaystyle\int_V h(v)\,\mathrm{d}\mu(v) = 0, \\ \displaystyle\int_V v^2 \,\mathrm{d}\mu(v) = d \quad \text{is positive.} \end{cases} \tag{4.3}$$

Typical examples are therefore:

- $V = (-1, +1)$ endowed with the normalized Lebesgue measure,
- $V = \{v^1, \dots, v^M\}$ where the v^i's are well-chosen points in $(-1, +1)$, endowed with the discrete velocity measure, $\mathrm{d}\mu(v) = \frac{1}{M}\sum_{i=1}^{M} \delta(v = v^i)$,
- $V = \mathbb{R}$ endowed with the Gaussian measure $\mathrm{d}\mu(v) = (2\pi)^{-1/2}\, e^{-v^2/2}\, \mathrm{d}v$.

Under these hypotheses, as we shall recall below, the behavior of f_ε for small ε's is given by the heat equation

$$\partial_t \rho - d\partial^2_{xx}\rho = 0. \tag{4.4}$$

Looking at such a simple situation makes a direct computation of the solution f_ε affordable, including for small values of ε. Therefore we have data at hand to compare with the solutions of reduced models. However, evaluating

f_ε when ε is small has a high numerical cost. Thus, it does not make sense to extend it to several dimensions for our purposes. Nevertheless, we can take advantage of our understanding of the limit process to design a numerical method that is well-suited to the asymptotic regime. The scheme we analyze is based on a splitting strategy with a convective-like step involving $\mathcal{O}(1)$ speeds and an explicitly solvable ODE step containing stiff sources. Hence, the scheme, which is naturally asymptotic preserving, is amenable to a fully explicit treatment, free of any $\varepsilon-$dependent restriction, and provides accurate results for a quite cheap numerical cost. Another viewpoint consists in using reduced macroscopic models which are intended to reproduce the main features of the original equation (4.1). Usually these models are derived either by using some truncated Chapman-Enskog expansion or by imposing a closure to the system that is satisfied by some moments of f_ε. A crucial requirement that is usually addressed to the model is to satisfy the so-called limited-flux property. In what follows a particular attention will be paid to the model the derivation of which relies on the Entropy Minimization Principle. In itself the numerical simulation of the reduced models is an issue, due to the presence of stiff terms and large speeds of propagation, that depend on ε. Nevertheless, we introduce original specific schemes for these models using relaxation techniques that we treat following the numerical philosophy evoked above, and interpreting the relaxing system as a discrete kinetic equation. This approach allows to compute efficiently the solutions of the macroscopic models.

The paper is organized as follows, postponing references to the existing literature to the following Sections. In Section 4.2, we recall some basic facts on the diffusion asymptotics and we present the reduced models we are interested in. In Section 4.3, we detail the derivation of the asymptotically-induced scheme for (4.1)-(4.2). We discuss the splitting strategy as well as the numerical boundary conditions which are designed to satisfy the mass conservation. Section 4.4 is devoted to adapting the method to the macroscopic models. This relies on the interpretation of the models through a relaxation limit. We end with the discussion of the numerical results in Section 4.5, with in particular simulations of the traditional Su-Olson benchmark.

4.2 A Brief Overview on Diffusion Asymptotics and Moment Closures

4.2.1 Diffusion Limit

We check readily that Assumption (4.3) has the following remarkable consequences.

Lemma 4.2.1 (Dissipation Properties of the Collision Operator)**.** *Assume* (4.3)*. Then the operator defined by* (4.2) *satisfies*

i) *Q is a bounded operator on $L^p(V)$, $1 \leq p \leq \infty$ spaces;*

ii) *Q is conservative which means that for any $f \in L^1(V, \mathrm{d}\mu)$,*

$$\int_V Q(f)\,\mathrm{d}\mu(v) = 0.$$

iii) *Q satisfies the dissipation property*

$$-\int_V Q(f) f\,\mathrm{d}\mu(v) = \int_V \left|f - \langle f\rangle\right|^2 \mathrm{d}\mu(v) \geq 0,$$

for any $f \in L^2(V, \mathrm{d}\mu)$, where the bracket is a shortcut notation for the average over V;

iv) *The elements of the kernel of Q are independent of the microscopic variable v: $\mathrm{Ker}(Q) = \mathrm{Span}(\mathbb{1})$;*

v) *The following Fredholm alternative holds: for any $h \in L^2(V)$ satisfying $\langle h\rangle = 0$, there exists a unique $f \in L^2(V)$ such that $Q(f) = h$ and $\langle f\rangle = 0$.*

The Fredholm alternative follows from a direct application of the Lax-Milgram theorem applied to the variational formula $\int_V Q(f)g\,\mathrm{d}\mu(v) = \int_V hg\,\mathrm{d}\mu(v)$ on the closed subspace $\{f \in L^2(V),\ \langle f\rangle = 0\}$.

As ε tends to 0, the number of interactions or "collisions" events per time unit increases. Accordingly, we can expect for small ε's that f_ε resembles an element of the kernel of the operator Q:

$$f_\varepsilon(t, x, v) \simeq \rho(t, x),$$

and it remains to describe the evolution of the macroscopic quantity ρ. The asymptotics can be readily understood by inserting the following Hilbert expansion

$$f_\varepsilon = F_0 + \varepsilon F_1 + \varepsilon^2 F_2 + \dots$$

into (4.1). Identifying terms arising with the same power of ε, we obtain

- At the leading order $Q(F_0) = 0$ that confirms $F_0 = \rho(t, x)$,
- Next, we have $Q(F_1) = v\partial_x F_0$. Then, we appeal to the second condition in (4.3) (applied with $h(v) = v$) which allows to make use of the Fredholm alternative. Accordingly, for the simple operator (4.2), we get $F_1(t, x, v) = -v\partial_x \rho(t, x)$.

- Then, we obtain a closed equation for ρ by integrating over v the relation: $Q(F_2) = \partial_t F_0 + v\partial_x F_1$. We obtain

$$\partial_t \rho + \partial_x \left(\int_V (-v^2 \partial_x \rho) \, \mathrm{d}\mu(v) \right) = 0$$

  that is the diffusion equation (4.4) for ρ.

Remark 4.2.1 (Time Scaling)**.** *The time scaling in* (4.1) *is motivated by the fact, embodied into* (4.3)*, that the equilibrium functions, i.e. the elements of* $\mathrm{Ker}(Q)$*, have a vanishing flux: considering only the penalization of the collision term, we would be led to the uninspiring equation* $\partial_t \rho = 0$*.*

The convergence of f_ε, solution of (4.1), to ρ, solution of (4.4), has been widely investigated under various and general assumptions, including non linear situations motivated by physical applications; we refer among others to [10, 11, 40, 59, 15, 80, 60]. Under suitable regularity assumptions, we can make the Hilbert expansion approach rigorous, estimate the remainder and justify the convergence with a rate. We refer to [11] for the following statement, which is part of the folklore in kinetic theory.

Theorem 4.2.1 (Asymptotic Convergence Rate)**.** *Assume that* (4.3) *hold. Let* $\overline{\rho} > 0$ *be a constant. Let* $f_0 : \mathbb{R} \times V \to \mathbb{R}$ *such that* $f_0 - \overline{\rho} \in L^2(\mathbb{R} \times V)$*.*

i) Then, as ε *goes to 0,* f_ε *and* ρ_ε *converge to* ρ *strongly in* $L^2_{\mathrm{loc}}(\mathbb{R}^+ \times \mathbb{R})$*, and* ρ_ε *converges to* ρ *in* $\mathcal{C}([0,T]; L^2(\mathbb{R}) - weak)$*, where* ρ *is the solution to the heat equation* (4.4) *with initial datum* $\rho_{|t=0} = \int_V f_0(x,v) \, d\mu(v)$*.*

ii) If the initial datum is close to a smooth enough macroscopic state, say e.g. $\|f_0 - \rho_0\|_{L^2(\mathbb{R}\times V)} \leq \varepsilon$*, with* $(\rho_0 - \overline{\rho}) \in H^3(\mathbb{R})$*, then, for any* $0 < T < \infty$*, there exists* $C_T > 0$ *such that one has*

$$\|f_\varepsilon - \rho\|_{L^2((0,T)\times\mathbb{R}\times V)} \leq C_T \, \varepsilon. \tag{4.5}$$

4.2.2 Approximate Models

We are interested in intermediate models, which are intended to be in between the full kinetic equation (4.1) and the heat equation (4.4). Such models are expected to provide "better" approximations of f_ε for moderate values of ε, that are small, but possibly not so small. We also expect that such a model retains more information from the microscopic modelling and we address the question of "how close" to the original unknown f_ε the approximate solution is. Finally, from a practical viewpoint, one should expect that the solution of the intermediate model can be computed with a reduced computational cost. Of course, the solution ρ of (4.4) already provides an approximation of order $\mathcal{O}(\varepsilon)$ in L^2 norm, but it has the drawback of loosing completely any microscopic feature since it does not depend on v. It

could also be tempting to use as an approximation the Hilbert expansion truncated at first order, the so-called $\mathbb{P}1$ approximation

$$f_\varepsilon(t,x,v) \simeq \rho(t,x) - \varepsilon\ v\ \partial_x \rho(t,x)$$

with ρ still the solution of (4.4). However, such an approximation is not non negative for any t, x, v. Furthermore, the heat equation propagates information at infinite speed while in (4.1) characteristic speeds are of order $\mathcal{O}(1/\varepsilon)$, at least if the set of velocities is bounded. Actually, the finite speed of propagation and preservation of non-negativeness are related; indeed, since $f_\varepsilon \geq 0$, we have the following relation between the macroscopic current and density

$$\Big| \int_V \frac{v}{\varepsilon} f_\varepsilon \,\mathrm{d}\mu(v) \Big| \leq \int_V \frac{|v|}{\varepsilon} f_\varepsilon \,\mathrm{d}\mu(v) \leq \frac{\|v\|_{L^\infty(V)}}{\varepsilon} \int_V f_\varepsilon \,\mathrm{d}\mu(v).$$

Therefore, we can require that a suitable approximation fulfills this so called "limited flux condition", which is thus guaranteed for free if the approximation is non negative.

To obtain intermediate models, a general strategy consists in writing a system of equations defined by the evolution of moments of f_ε. The system is not closed since the convection term makes the $(k+1)$-th moment appear in the evolution equation of the kth moment. Thus, we impose a relation between the higher moment involved in the system and the previous ones. We expect that this closure provides a suitable approximation of the evolution of the kinetic density. For (4.1), it is enough to consider the evolution of the zeroth and first order moments. Let us set

$$\begin{pmatrix} \rho_\varepsilon \\ J_\varepsilon \\ \mathbb{P}_\varepsilon \end{pmatrix} = \int_V \begin{pmatrix} 1 \\ v/\varepsilon \\ v^2 \end{pmatrix} f_\varepsilon \,\mathrm{d}\mu(v).$$

We get the mass conservation

$$\partial_t \rho_\varepsilon + \partial_x J_\varepsilon = 0, \tag{4.6}$$

completed by

$$\varepsilon^2 \partial_t J_\varepsilon + \partial_x \mathbb{P}_\varepsilon = -J_\varepsilon. \tag{4.7}$$

According to [36], we are interested in two possible closure strategies:

(C1) Either we define an approximation, formally close to the $\mathbb{P}1$ formula, but which preserves non negativity. By using this approximation into the conservation law (4.6), we obtain a possibly nonlinear equation, that, in some sense, interpolates between transport and diffusion.

(C2) Or we close the moment system (4.6)-(4.7), so that we obtain a hyperbolic system that restores the finite speeds of propagation.

We refer to [36] and the references therein for further detail. Let us introduce the following notation

$$\mathbb{F}(\beta) = \int_V e^{\beta v}\,\mathrm{d}\mu(v), \qquad \mathbb{G}(\beta) = \frac{\mathbb{F}'}{\mathbb{F}}(\beta)$$

and

$$\psi(u) = \frac{\mathbb{F}"}{\mathbb{F}}\Big(\mathbb{G}^{(-1)}(u)\Big).$$

The zeroth order closure **(C1)** is based on the modified Hilbert expansion

$$f_\varepsilon = \exp(a_0 + \varepsilon a_1 + \varepsilon^2 a_2 + \dots)$$

Truncating at first order, we get the approximation

$$\widetilde{f}_\varepsilon(t,x,v) = \frac{\varrho(t,x)}{Z(t,x)} \exp\Big(-\varepsilon v\ \frac{\partial_x \varrho}{\varrho}(t,x)\Big),$$

with $Z(t,x)$ to normalize the density to $\varrho(t,x)$. Plugging this expression into the moment equation, ϱ satisfies

$$\partial_t \varrho - \partial_x\Big(\frac{\varrho}{\varepsilon}\mathbb{G}\big(\varepsilon\frac{\partial_x\varrho}{\varrho}\big)\Big) = 0. \tag{4.8}$$

The first order closure **(C2)** follows from a Entropy Minimization Principle. This idea is due to Levermore [73, 74, 77, 75, 76], but it also appears in various physical applications [34, 47]. It works as follows. For given ϱ, J, let

$$\widetilde{f} = \operatorname{argmin}\Big\{ \int_V f\ln(f)\,\mathrm{d}\mu(v), \quad \int_V (1, v/\varepsilon) f\,\mathrm{d}\mu(v) = (\varrho, J)\Big\}.$$

We obtain

$$\widetilde{f}(v) = e^{\lambda_0 + \lambda_1 v/\varepsilon}$$

where the Lagrange multipliers $\lambda_{0,1}$ are defined by the constraints

$$\begin{aligned}\varrho &= \int_V e^{\lambda_0+\lambda_1 v/\varepsilon}\,\mathrm{d}\mu(v) = e^{\lambda_0}\mathbb{F}(\lambda_1/\varepsilon)\\ J &= \int_V \frac{v}{\varepsilon} e^{\lambda_0+\lambda_1 v/\varepsilon}\,\mathrm{d}\mu(v) = \frac{\rho}{\varepsilon}\mathbb{G}(\lambda_1/\varepsilon).\end{aligned}$$

Then, we use $\widetilde{f}$ to define the second moment that closes the system (4.6)-(4.7). Namely, we set

$$\mathbb{P} = \int_V v^2 \widetilde{f}(v)\,\mathrm{d}\mu(v) = \varrho\frac{\mathbb{F}"}{\mathbb{F}}(\lambda_1/\varepsilon) = \varrho\psi(\varepsilon J/\varrho),$$

and we are thus led to the system

$$\begin{cases} \partial_t\varrho + \partial_x J = 0,\\ \varepsilon^2\partial_t J + \partial_x\big(\varrho\psi(\varepsilon J/\varrho)\big) = -J. \end{cases} \tag{4.9}$$

The microscopic approximation is defined by

$$\widetilde{f}_\varepsilon(t,x,v) = \varrho(t,x)\ \frac{\exp\left[v\,\mathbb{G}^{(-1)}\big(\varepsilon J/\varrho(t,x)\big)\right]}{\mathbb{F}\circ\mathbb{G}^{(-1)}\big(\varepsilon J/\varrho(t,x)\big)}. \tag{4.10}$$

Of course, Equations (4.8) and (4.9) highly depend on the considered measure $\mathrm{d}\mu$ through the functions $\mathbb{F}$, $\mathbb{G}$ and ψ:

- For the Lebesgue measure, we have $\mathbb{F}(\beta) = \sinh(\beta)/\beta$, $\mathbb{G}(\beta) = \coth(\beta) - 1/\beta$.
- For the discrete 2-velocity measure, we have $\mathbb{F}(\beta) = \cosh(\beta)$. The first order closure (4.9) is in this case completely equivalent to the original kinetic model and there is no approximation at all.
- For the Gaussian measure, we have $\psi(u) = 1 + u^2$. The zeroth order closure actually leads to the heat equation, and the first order closure gives the isothermal Euler system.

In [36], the well-posedness of (4.8) and (4.9) is justified, at least for small and smooth initial data, but, hopefully, with an $\varepsilon-$free smallness condition. (We also refer to [37] for preliminary discussions on weak solutions.) Furthermore it is shown that $\|f_\varepsilon - \widetilde{f}_\varepsilon\|_{L^2}$ is of order $\mathcal{O}(\varepsilon)$. This estimate is a bit disappointing since it is not better than those evaluating the distance to the solution of the heat equation. Our aim in this paper is to investigate numerically (4.1)-(4.2) and its approximation (4.8) or (4.9)-(4.10), compared to the heat equation and the $\mathbb{P}1$ approximation. It is indeed interesting to check numerically whether we can expect sharper estimates or not. It is also important in view of applications to discuss how the quality of the approximation is degraded as ε increases and to know if one of the approximation strategies has some decisive advantages. Let us mention that there exist a huge variety of possible closure methods, based either on mathematical arguments or physical grounds, and we mention among others [77, 31].

4.3 Asymptotic Preserving Explicit Kinetic Scheme

On the numerical viewpoint, the computation of (4.1)-(4.2) is also a challenging question due to the presence of large, say $\mathcal{O}(1/\varepsilon)$, speeds of propagation and stiff terms. An attempt to solve (4.1)-(4.2) by integrating the equation along the characteristics following a splitting strategy between collisions and transport through lines $x+tv/\varepsilon$ fails for small ε. Since in general the characteristics do not end at a point of the discrete mesh, this approach needs to be completed by a suitable interpolation procedure. It gives rise to semi-lagrangian numerical methods that have been used successfully for Vlasov's like equations [45, 46]. Proceeding naively, such a procedure can

produce unacceptable numerical diffusion. One can repair this drawback by using interpolation procedures based on the WENO approach. We refer to [96, 94] for the basis of the WENO method, and to [26] for a description of the adaptation to design an accurate interpolation method. Of course, for small ε's these computations become unbearably time consuming with large meshes and small time step due to the large velocities that are involved.

Asymptotic schemes working in the stiffness regime have to be developed. We propose an alternative approach using a splitting scheme inspired by the Hilbert expansion that treats the stiffness of (4.1). The method is well fitted and much less costly than the previous approach to the diffusion regime while remaining fully explicit. This numerical method, which improves the scheme already proposed in [53], is a fully explicit variation of the methods introduced in [69, 70], and it has successfully been used in other contexts [53, 24, 58]. It is also relevant to compare our method to those of [65, 66], based on odd/even flux decomposition or, concerning relaxation problems, those of [81] for methods based on central schemes. Here, the scheme is based on the expansion

$$f_\varepsilon = \rho_\varepsilon + \varepsilon g_\varepsilon, \qquad \rho_\varepsilon(t,x) = \int_V f_\varepsilon \, \mathrm{d}\mu(v),$$

where the dissipation properties of the operator Q imply that the "fluctuations" g_ε are indeed bounded in $L^2(\mathbb{R}^+ \times \mathbb{R} \times V)$. We rewrite (4.1) as

$$\partial_t f_\varepsilon + v\partial_x g_\varepsilon = \frac{1}{\varepsilon^2}(\rho_\varepsilon - f_\varepsilon) - \frac{v}{\varepsilon}\partial_x \rho_\varepsilon,$$

which motivates the following two step splitting scheme:

Given a uniform subdivision of step Δt of $[0,\infty)$ and knowing f^n, which is expected to approximate $f^\varepsilon(n\Delta t, x, v)$, $n \in \mathbb{N}$

Step 1.- Solve on the time interval $[n\Delta t, (n+1)\Delta t)$ the stiff ODE

$$\partial_t f = \frac{1}{\varepsilon^2}(\rho - f) - \frac{1}{\varepsilon} v\partial_x \rho. \tag{4.11}$$

Since the average over V of the right hand side vanishes, the macroscopic density is not modified during this time step, that is,

$$\rho^{n+1/2} = \int_V f^{n+1/2} \, \mathrm{d}\mu(v) = \rho^n.$$

Moreover, (4.11) also defines the evolution of the fluctuation

$$\partial_t g = -\frac{1}{\varepsilon^2} g - \frac{1}{\varepsilon^2} v\partial_x \rho. \tag{4.12}$$

Step 2.- Solve on the time interval $[n\Delta t, (n+1)\Delta t)$:

$$\partial_t f + v\partial_x g = 0 \qquad \text{and} \qquad \partial_t g = 0, \tag{4.13}$$

with initial data $f^0 = f^{n+1/2}$ and $g^0 = g^{n+1/2}$ This defines, f^{n+1} and

$$\rho^{n+1} = \int_V f^{n+1}\,\mathrm{d}\mu(v).$$

We emphasize that the index ε has been dropped for notational convenience. Note that, in the second step, the convective term involves a characteristic speed of order $\mathcal{O}(1)$ only and that we will not force any update on g as $g^{n+1/2} = (f^{n+1/2} - \rho^{n+1/2})/\varepsilon$. This update might make the relation between f and g consistent at the end of the second step but it leads to undesirable numerical divisions by the small parameter ε; but, for well-prepared initial data, this consistency can be imposed at the beginning. Similar arguments were already given in [69, 70] to avoid this update of the fluctuations g.

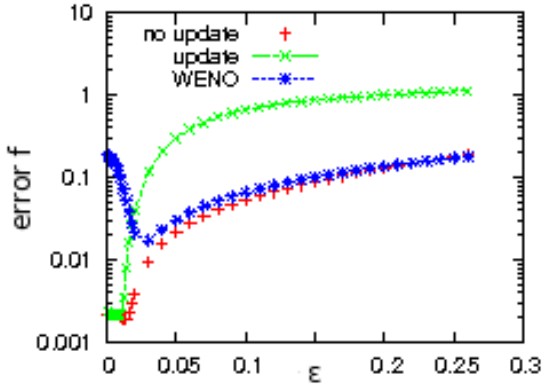


Figure 1: $L^2_{t,x,v}$-error of the distribution function f with respect to the solution of the heat equation with a symmetric initial data as in Section 4.5 with a mesh of 100x100 with respect to ε.

Before proceeding further with the analysis of this kinetic method, we show in Figure 1 a comparison between the results of the three discussed kinetic methods: a semi-lagrangian PWENO6,4-interpolation scheme [26] (SL-WENO), the asymptotic preserving method without update of g proposed above and the asymptotic preserving method with update of g. In all cases, we show the $L^2_{t,x,v}$-error between the kinetic results and the solutions of the heat equation, its $\varepsilon \to 0$ asymptotic limit, in a log-plot depending on ε. The results show that the kinetic scheme proposed in this paper works perfectly in the $\varepsilon \to 0$ regime while both the updated scheme and the SL-WENO scheme do not describe well the asymptotic limit. It is important

to point out that all the schemes are computed with the parabolic CFL condition corresponding to the limiting heat equation. We also note that for larger values of ε the difference between the SL-WENO method and the asymptotic preserving scheme becomes small which shows the ability of the proposed scheme to capture the behavior of the kinetic equation, for moderately small value of the mean free path as well, with a considerable gain of CPU time. It is also worthy to emphasize that the results are given for a fixed mesh 100x100, so that as $\varepsilon \to 0$, the SL-WENO method cannot work, the velocities being of order $\mathcal{O}(1/\varepsilon)$. In order to get accurate results comparable to those obtained with the asymptotic-induced scheme, the SL-WENO method would require larger and larger meshes in velocity and smaller and smaller time steps as $\varepsilon \to 0$. This leads to an unbearable computational cost for such a simple equation.

We can simplify the first step by keeping only the leading contribution in ε and, by explicitly solving Equations (4.11) and (4.12) that define $f^{n+1/2}$, $\rho^{n+1/2}$ and $g^{n+1/2}$, leading to

$$g^{n+1/2} = e^{-\Delta t/\varepsilon^2} g^n - (1 - e^{-\Delta t/\varepsilon^2}) v \partial_x \rho^n \tag{4.14}$$

and

$$f^{n+1/2} = e^{-\Delta t/\varepsilon^2} f^n + (1 - e^{-\Delta t/\varepsilon^2}) \rho^n, \tag{4.15}$$

keeping in mind that $\rho^{n+1/2} = \rho^n = \int_V f^n \, \mathrm{d}\mu(v)$. The final semi-discrete scheme is summarized as:

Step 1.- Compute

$$\begin{cases} g^{n+1/2} = e^{-\Delta t/\varepsilon^2} g^n - (1 - e^{-\Delta t/\varepsilon^2}) v \partial_x \rho^n, \\ f^{n+1/2} = e^{-\Delta t/\varepsilon^2} f^n + (1 - e^{-\Delta t/\varepsilon^2}) \rho^n. \end{cases} \tag{4.16}$$

Remember that $\rho^{n+1/2} = \rho^n$.

Step 2.- Solve for time Δt the convection equation:

$$\partial_t f + v \partial_x g = 0$$

to compute the values of f^{n+1} and ρ^{n+1} while $g^{n+1} = g^{n+1/2}$.

Remark 4.3.1 (Asymptotic Preserving)**.** *It is worthwhile mentioning that the scheme is "asymptotic preserving": using* (4.15) *and* (4.14) *for the completely relaxed model, i.e., $\varepsilon = 0$, yields*

$$f^{n+1/2} = \rho^{n+1/2}, \qquad g^{n+1/2} = -v \partial_x \rho^n = -v \partial_x \rho^{n+1/2},$$

which coincides with the first order term in the Hilbert expansion. Thus, the first step becomes

$$\partial_t f - v^2 \partial_{xx}^2 \rho = 0$$

Integrating over the mesh of velocities leads to the expected heat equation, up to a suitable v-mesh definition in order to guarantee $\int_V v^2 \, \mathrm{d}v = d$.

Remark 4.3.2 (Spatial Derivatives Discretization)**.** *One has to take care of the treatment of the space derivative: if one uses the same upwind discretization for evaluating both $-v\partial_x\rho$ in the first step and $-v\partial_x g$ in the second one, it leads to an unstable scheme for the heat equation. The usual $3-$point scheme is obtained by choosing opposite upwind discretization in the successive time steps. Accordingly, for $\varepsilon = 0$, the stability of the scheme is guaranteed by the CFL condition $d\Delta t/(\Delta x)^2 \leq 1/2$.*

Remark 4.3.3 (Stability)**.** *The stability condition for the scheme used with $\varepsilon > 0$ is less clear, even if a CFL condition close to the parabolic one can be reasonably expected. We refer to [71] for a discussion on a semi-implicit version of the proposed scheme. This difficulty has motivated the development of implicit methods, as in [56, 57].*

Remark 4.3.4 (Current and Distribution Computation)**.** *Due precisely to the separation between fluctuations and relaxation towards the homogeneous density we impose in the scheme and taking into account the comments above regarding asymptotic preservation, we need to compute and reconstruct J and f to compare to other methods. In fact, since currents appear due to fluctuations, it is intuitive to reconstruct it as*

$$J^{n+1} = \int_V v\ g^{n+1}\,\mathrm{d}\mu(v).$$

Due to the Hilbert expansion approach, we will consider the reconstructed distribution given by $\rho^{n+1} + \varepsilon g^{n+1}$.

Let us restrict from now on to the case of the normalized Lebesgue measure $\mathrm{d}\mu(v)$ on the velocity space $[-1, 1]$. The space interval $[X_{min}, X_{max}]$ is uniformly discretized in $N_x - 1$ intervals with points $x_i = i\Delta x$ from $i = 0, \ldots, N_x - 1$ and the velocity interval $[-1, 1]$ is discretized analogously in $N_v - 1$ intervals with points $v_j = -1 + j\Delta v$ from $j = 0, \ldots, N_v - 1$. For further purposes, it is convenient to introduce the sets

$$V_+ = \big\{j \in \{0, N_v - 2\} \text{ such that } v_j > 0\big\},$$
$$V_- = \{j \in \{0, N_v - 2\} \text{ such that } v_j < 0\}.$$

Let us specify our discrete scheme to the case of simple upwind discretization $\mathbb{D}_j$ of the spatial differential operator $-v_j\partial_x$ with $\bar{\mathbb{D}}_j$ being its alternate direction: for a given sequence $(\varphi_i)_{i\in\mathbb{N}}$, we set

$$\big[\mathbb{D}_j\varphi\big]_i = \begin{cases} -v_j(\varphi_i - \varphi_{i-1}) & \text{if } v_j \in V_+, \\ -v_j(\varphi_{i+1} - \varphi_i) & \text{if } v_j \in V_-, \end{cases} \quad \big[\bar{\mathbb{D}}_j\varphi\big]_i = \begin{cases} -v_j(\varphi_{i+1} - \varphi_i) & \text{if } v_j \in V_+, \\ -v_j(\varphi_i - \varphi_{i-1}) & \text{if } v_j \in V_-. \end{cases} \tag{4.17}$$

More advanced non-centered non linear distinct numerical fluxes for $-v_j\partial_x$, such as flux limiting ones, may be chosen. Similarly, we could use a non uniform time mesh. However, this might complicate boundary conditions below to preserve mass and it will certainly change the relaxed asymptotic scheme. The fully discrete scheme summarizes as

Step 1.- Compute

$$\begin{cases} g_{i,j}^{n+1/2} = e^{-\Delta t/\varepsilon^2} g_{i,j}^n + (1 - e^{-\Delta t/\varepsilon^2}) \bar{\mathbb{D}}_j \rho_i^n \\ f_{i,j}^{n+1/2} = e^{-\Delta t/\varepsilon^2} f_{i,j}^n + (1 - e^{-\Delta t/\varepsilon^2}) \rho_i^n \end{cases}, \tag{4.18}$$

with

$$\rho_i^{n+1/2} = \rho_i^n = \frac{\Delta v}{2} \sum_{j=0}^{N_v-2} f_{i,j}^n$$

since a left rectangular rule has been chosen.

Step 2.- Solve for time Δt the convection-like equation:

$$f_{i,j}^{n+1} = f_{i,j}^{n+1/2} + \Delta t \mathbb{D}_j g_{i,j}^{n+1/2} \tag{4.19}$$

to compute the values of f^{n+1} and ρ^{n+1} while $g^{n+1} = g^{n+1/2}$.

Remark 4.3.5 (Maximum Principle)**.** *We point out again that the scheme is specifically designed for the small ε regime, and there is no guarantee about the accuracy of the results when ε becomes large. In particular difficulties might arise with the maximum principle. Indeed, in Step 1, given a non negative f^n,* (4.18) *returns a non negative $f^{n+1/2}$, but this property is not naturally preserved in Step 2, see* (4.19)*.*

Finally, we need to impose boundary conditions on the advection step ensuring the total mass conservation. With this aim, we need

$$\sum_{i=1}^{N_x-2} \sum_{j=0}^{N_v-2} \mathbb{D}_j g_{i,j}^{n+1/2} = 0$$

which is equivalent, by summing the telescopic series appearing due to the definition of the upwinding operators, to

$$\sum_{v_j \in V_+} v_j (g_{N_x-2,j}^{n+1/2} - g_{0,j}^{n+1/2}) + \sum_{v_j \in V_-} v_j (g_{N_x-1,j}^{n+1/2} - g_{1,j}^{n+1/2}) = 0$$

where $V_+ = \{j \in \{0, \dots, N_v - 2\} \text{ such that } v_j > 0\}$ and $V_+ = \{j \in \{0, \dots, N_v - 2\} \text{ such that } v_j < 0\}$. From this, we will impose as boundary conditions for the fluctuations:

$$g_{0,k}^{n+1/2} = \frac{-1}{v_k \, \#[V_+]} \sum_{v_j \in V_-} v_j g_{1,j}^{n+1/2}$$

for $k \in V_+$ and

$$g_{N_x-1,k}^{n+1/2} = \frac{-1}{v_k \, \#[V_-]} \sum_{v_j \in V_+} v_j g_{N_x-2,j}^{n+1/2}$$

for $k \in V_-$, where $\#[B]$ is the cardinal of the set B. Let us remark that the previous boundary condition in the complete relaxed scheme, $\varepsilon \to 0$, coincides with the Neumann boundary condition for the density, i.e.,

$$\rho_0^n = \rho_1^n \qquad \text{and} \qquad \rho_{N_x-1}^n = \rho_{N_x-2}^n.$$

The scheme described above gives a simple way to compute the solution of (4.1)-(4.2), and the associated macroscopic density, that has to be compared, both in terms of accuracy and computational cost, to the direct evaluation, see Fig. 1, and computation of the solution of the heat equation (4.4) and the different approximations by the closure strategies.

The method adapts easily to more complicated models: gas dynamics [69, 70, 65, 66], radiative transfer [53, 58], fluid-particles flows [24]. It can be also incorporated in a domain decomposition method to deal with space varying mean free path, in the spirit of [54, 102].

4.4 Numerical Schemes for Closure Approximations

Next, the idea to treat the hyperbolic system (4.9) or the conservation equation (4.8) is two-fold:

1. We introduce additional unknowns and parameters and the equations are seen as the relaxation limit of an extended system, in the spirit of general methods described in [85],

2. The relaxation system is interpreted itself as a kinetic equation with a discrete set of velocities to which we apply the splitting algorithm described above.

4.4.1 Relaxation Method for the First-Order Closure

We will at first focus on developing a numerical scheme for the first order closure (4.9). The nonlinear system (4.9) can be seen as the limit, as α tends to 0, of

$$\partial_t \rho + \partial_x J = 0, \tag{4.20}$$

$$\varepsilon^2 \partial_t J + \partial_x z = -J, \tag{4.21}$$

$$\partial_t z + \varepsilon^2 \lambda^2 \partial_x J = \frac{1}{\alpha}\big(\rho\psi(\varepsilon J/\rho) - z\big). \tag{4.22}$$

Let us define $u := \varepsilon J/\rho$. Recall that u should be small of order $\mathcal{O}(\varepsilon)$, see [36]. This system involves an additional unknown $z(t,x)$ and the parameters λ (convection speed) and α (relaxation parameter). Actually we relax on the

quantity $\varepsilon^2 J$ so that we consider the velocity in (4.22) rescaled by ε (that fits dimensional considerations). The advantage in considering (4.20)-(4.22) is that now we have to deal with simple convection equations, the convection part being linear, and all nonlinearities only appear in the (zeroth order) source terms. This idea is reminiscent to the introduction of kinetic schemes in [17, 52, 89, 79, 78], and relaxation methods for conservation laws [67]. We refer to [7, 88] for further details and references. This approach can be used also to treat degenerate diffusion equations [85].

Let us find the constraints on the additional velocities $\pm\lambda$ that should be large enough to propagate enough information to reconstruct the behavior of (4.9). It is important to check whether the condition becomes more constrained as ε tends to 0. To this end, let us perform the Chapman-Enskog reasoning; we expand (4.22) with respect to α considering ε to be small. We have

$$z = \rho\psi(u) - \alpha(\partial_t z + \varepsilon^2\lambda^2\partial_x J) = \rho\psi(u) - \alpha(\partial_t(\rho\psi(u)) + \varepsilon^2\lambda^2\partial_x J) + \mathcal{O}(\alpha^2), \tag{4.23}$$

by (4.22)-(4.21). Thus, (4.21) can be recast as

$$\varepsilon^2\partial_t J + \partial_x\big(\rho\psi(u)\big) + J = \alpha\varepsilon^2\lambda^2\partial^2_{xx}J + \alpha\partial^2_{xt}\big(\rho\psi(u)\big) + \mathcal{O}(\alpha^2).$$

Let us compute the leading contribution in the last term; by using (4.20) and (4.21), we get

$$\partial_t\big(\rho\psi(u)\big) = -\psi(u)\partial_x J + \rho\psi'(u)\partial_t u.$$

But

$$\partial_t(\rho\psi(u)) = (u\psi'(u) - \psi(u))\partial_x J - \frac{\psi'(u)}{\varepsilon}(J + \partial_x z).$$

Now, considering that formally $J + \partial_x z$ is of order $\mathcal{O}(\varepsilon)$ at least, that ψ is an even function and using the approximation $\psi(u) = \psi(0) + \mathcal{O}(\varepsilon^2)$, with $\psi(0) > 0$, we get

$$\partial_t\big(\rho\psi(u)\big) + \psi(0)\partial_x J = \mathcal{O}(\varepsilon),$$

so that

$$\varepsilon^2\partial_t J + \partial_x\big(\rho\psi(u)\big) + J = \alpha\partial_x((\varepsilon^2\lambda^2 - \psi(0))\partial_x J) + \mathcal{O}(\alpha^2, \varepsilon).$$

Consequently, as soon as $\varepsilon|\lambda| > \sqrt{\psi(0)}$, the parabolicity is ensured. It is certainly natural to find that the speeds tend to infinity as ε tends to 0 since we want to approximate the heat equation. Now, we need to diagonalize System (4.20)-(4.22). Since the quantity $\varepsilon\lambda$ remains bounded from below, we denote it by μ. We define

$$f_0 = \mu^2\rho - z, \tag{4.24}$$

$$f_\pm = \frac{1}{2}(z \pm \varepsilon\mu J). \tag{4.25}$$

Of course, we have

$$z = f_+ + f_- \quad \text{and} \quad J = \frac{f_+ - f_-}{\varepsilon\mu} \quad \text{and} \quad \rho = \frac{f_0 + f_+ + f_-}{\mu^2}.$$

Noting that

$$J = \pm\frac{2f_\pm - z}{\varepsilon\mu},$$

the new system we are interested in is

$$\partial_t f_0 = -\frac{1}{\alpha}(\rho\psi(u) - z), \tag{4.26}$$

$$\partial_t f_\pm \pm \frac{\mu}{\varepsilon}\partial_x f_\pm = -\frac{f_\pm}{\varepsilon^2} + \frac{z}{2\varepsilon^2} + \frac{1}{2\alpha}(\rho\psi(u) - z). \tag{4.27}$$

The system shares some structures with the kinetic equation analyzed in the previous section. This similarity will be used to design a new scheme that will be expressed only in terms of the macroscopic quantities ρ and J. Since we have two small parameters, we can use a double splitting method, i.e., by splitting with respect to ε inside the splitting with respect to α, that is:

Step 1.- Solve

$$\partial_t f_0 = 0, \tag{4.28}$$

$$\partial_t f_\pm \pm \frac{\mu}{\varepsilon}\partial_x f_\pm = -\frac{f_\pm}{\varepsilon^2} + \frac{z}{2\varepsilon^2}. \tag{4.29}$$

This system is again stiff as ε tends to 0. Let us solve it with the splitting method described in Section 4.3: we introduce the intermediate variables

$$g_\pm := \frac{2f_\pm - z}{2\varepsilon} = \pm\frac{\mu J}{2}$$

and rewrite (4.29) as

$$\partial_t f_\pm \pm \mu\partial_x g_\pm = -\frac{g_\pm}{\varepsilon} \mp \frac{\mu}{2\varepsilon}\partial_x z. \tag{4.30}$$

Solve

Step 1.1.-
$$\partial_t f_\pm = -\frac{f_\pm}{\varepsilon^2} + \frac{z}{2\varepsilon^2} \mp \frac{\mu}{2\varepsilon}\partial_x z,$$
$$\partial_t g_\pm = -\frac{g_\pm}{\varepsilon^2} \mp \frac{\mu}{2\varepsilon^2}\partial_x z,$$

where the initial condition for the ODEs are the values computed in the previous step and solve

Step 1.2.-
$$\partial_t f_\pm \pm \mu\partial_x g_\pm = 0,$$
$$\partial_t g_\pm = 0,$$

where the initial conditions are, for $f_\pm$, the ones obtained by Step 1.1. For $g_\pm$, we update them in terms of the flux $g_\pm(0) = \pm\mu J/2 = \pm(f_+ - f_-)/2\varepsilon$. Note that here the update is necessary since the goal of the scheme is actually to compute the macroscopic flux and the microscopic quantities $f_\pm$, f_0 and $g_\pm$ are only auxiliary devices.

Note that, during Step 1.1, $\partial_t z = 0$. Let us now specify our fully discrete kinetic scheme. As in the previous section, let us choose $\mathbb{D}_\pm$ an upwind discretization of the spatial differential operator $\mp\mu\partial_x$ and $\bar{\mathbb{D}}_\pm$ its alternate direction version, see (4.17). The fully discrete kinetic scheme in this step reads as

Step 1.1.-(Micro)

$$\begin{aligned}
g_\pm^{n+1/4} &= e^{-\Delta t/\varepsilon^2} g_\pm^n + (1 - e^{-\Delta t/\varepsilon^2})\, \frac{1}{2}\, \bar{\mathbb{D}}_\pm (f_+^n + f_-^n), \\
f_\pm^{n+1/4} &= e^{-\Delta t/\varepsilon^2} f_\pm^n + (1 - e^{-\Delta t/\varepsilon^2}) \left(\frac{f_+^n + f_-^n + \varepsilon \bar{\mathbb{D}}_\pm (f_+^n + f_-^n)}{2} \right), \\
f_0^{n+1/4} &= f_0^n.
\end{aligned}$$

At the microscopic level, after Step 1.1, $g_+^{n+1/4} \neq -g_-^{n+1/4}$.

Step 1.2.-(Micro)

$$\begin{aligned}
g_\pm^{n+1/2} &= \pm \frac{f_+^{n+1/4} - f_-^{n+1/4}}{2\varepsilon}, \\
f_\pm^{n+1/2} &= f_\pm^{n+1/4} + \Delta t \mathbb{D}_\pm (g_\pm^{n+1/4}), \\
f_0^{n+1/2} &= f_0^{n+1/4}, \\
(\rho\psi)^{n+1/2} &= \frac{1}{\mu^2} (f_+^{n+1/2} + f_-^{n+1/2} + f_0^{n+1/2}) \\
&\qquad \times \psi \left(\frac{\mu (f_+^{n+1/2} - f_-^{n+1/2})}{f_+^{n+1/2} + f_-^{n+1/2} + f_0^{n+1/2}} \right)
\end{aligned}$$

Note that, after Step 1.2, we have $g_+^{n+1/2} = -g_-^{n+1/2}$, as it is the case in the continuous setting. That is the reason why the macroscopic quantities can only be expressed at the end of Step 1 as a whole, and not at the end of

Step 1.1. The **macroscopic scheme** summarizes in this step as:

Step 1.-(Macro)

$$\begin{aligned}
z^{n+1/2} &= z^n + \frac{\varepsilon(1-e^{-\Delta t/\varepsilon^2})}{2}\left(\bar{\mathbb{D}}_+(z^n)+\bar{\mathbb{D}}_-(z^n)\right)\\
&\quad +\Delta t\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\
&\quad \left.+\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right],\\
J^{n+1/2} &= e^{-\Delta t/\varepsilon^2}J^n + \frac{1-e^{-\Delta t/\varepsilon^2}}{2\mu}\left(\bar{\mathbb{D}}_+(z^n)-\bar{\mathbb{D}}_-(z^n)\right)\\
&\quad +\frac{\Delta t}{\varepsilon\mu}\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\
&\quad \left.-\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right],\\
\rho^{n+1/2} &= \rho^n + \frac{\Delta t}{\mu^2}\left(\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\
&\quad \left.+\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right).
\end{aligned}$$

Let us remark that the second term in $z^{n+1/2}$ is of order ε and thus, it will be omitted in the computations below. Now, we can write the relaxation step with respect to the parameter α:

Step 2.- Solve the ODE

$$\partial_t f_0 = -\frac{1}{\alpha}(\rho\psi(u)-z), \tag{4.31}$$

$$\partial_t f_\pm = \frac{1}{2\alpha}(\rho\psi(u)-z), \tag{4.32}$$

that is, since $\partial_t J = 0$ and $z = 2f_\pm \mp \varepsilon\mu J$ by virtue of (4.25),

$$\partial_t f_0 = -\frac{1}{\alpha}(\rho\psi(u)-z), \tag{4.33}$$

$$\partial_t f_\pm = -\frac{1}{\alpha}f_\pm + \frac{1}{2\alpha}(\rho\psi(u)\pm\varepsilon\mu J), \tag{4.34}$$

with as initial conditions the values computed from Step 1. In this last step, we also note that $\partial_t\rho = 0$, see (4.24) and (4.25), and that, consequently, u is constant. The fully discrete kinetic scheme summarizes in this step as

Step 2.-(Micro)

$$\begin{aligned}
f_\pm^{n+1} &= e^{-\Delta t/\alpha}f_\pm^{n+1/2} + \frac{1}{2}(1-e^{-\Delta t/\alpha})((\rho\psi)^{n+1/2}+2\varepsilon g_\pm^{n+1/2}),\\
f_0^{n+1} &= f_0^n + (1-e^{-\Delta t/\alpha})\left(f_+^n+f_-^n-(\rho\psi)^{n+1/2}\right)
\end{aligned}$$

In this last step, we do not update g, since $\partial_t J = 0$. Denoting by $\psi^{n+1/2}$ the quantity $\psi(\varepsilon J^{n+1/2}/\rho^{n+1/2})$, we deduce the following macroscopic scheme for the second step:

Step 2.-(Macro)

$$\begin{aligned} z^{n+1} &= e^{-\Delta t/\alpha} z^{n+1/2} + (1 - e^{-\Delta t/\alpha})\rho^{n+1/2}\psi^{n+1/2}, \\ J^{n+1} &= J^{n+1/2}, \\ \rho^{n+1} &= \rho^{n+1/2}. \end{aligned}$$

Remark 4.4.1 (Splitting Order)**.** *We choose a semi-linear relaxation method to have to deal only with transport-like equations, that is, move the non-linearities to the right-hand side, as source terms. So it is natural to take this precise order of splitting, since α must be the first one to tend to 0, so that we can keep a non-zero ε, even if it is small.*

Remark 4.4.2 (Initial Conditions)**.** *In order to prevent an initial layer from appearing [86] in the $\alpha-$splitting, we need to prescribe well-prepared initial conditions taking into account both splittings as:*

$$\begin{aligned} \rho(0,x) &= \rho^0(x), \\ J(0,x) &= J^0(x), \\ z(0,x) &= \rho^0(x)\psi(\varepsilon J^0(x)/\rho^0(x)) =: z^0(x), \end{aligned}$$

corresponding to the choice of the equilibrium state for the (hyperbolic) α-splitting, as in the classical relaxation approach.

Remark 4.4.3 (Boundary Conditions)**.** *We consider, for a computation domain $[X_{min}, X_{max}]$, Neumann conditions for ρ and z,*

$$\rho_0^n = \rho_1^n \qquad \rho_{N_x-1}^n = \rho_{N_x-2}^n \qquad \text{and} \qquad z_0^n = z_1^n \qquad z_{N_x-1}^n = z_{N_x-2}^n.$$

and

$$J_0^n = -J_1^n \qquad J_{N_x-1}^n = -J_{N_x-2}^n.$$

These conditions guarantee the conservation of the total mass.

Let us now have a look at the limits as α tends to 0 :

$$\begin{aligned} J^{n+1} &= e^{-\Delta t/\varepsilon^2} J^n + \frac{1 - e^{-\Delta t/\varepsilon^2}}{2\mu} \left(\bar{\mathbb{D}}_+(\rho^n\psi^n) - \bar{\mathbb{D}}_-(\rho^n\psi^n)\right) \\ &\quad + \frac{\Delta t}{\varepsilon\mu}\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2} + (1 - e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(\rho^n\psi^n)}{2}\right)\right. \\ &\quad \left. - \mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2} + (1 - e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(\rho^n\psi^n)}{2}\right)\right], \\ \rho^{n+1} &= \rho^n + \frac{\Delta t}{\mu^2}\left(\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2} + (1 - e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(\rho^n\psi^n)}{2}\right)\right. \\ &\quad \left. + \mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2} + (1 - e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(\rho^n\psi^n)}{2}\right)\right). \end{aligned}$$

We thus obtain a pure transport-projection scheme [7]. Note that, since $\Delta t e^{-\Delta t/\varepsilon^2} = \mathcal{O}(\varepsilon^2)$ and

$$\left(\mathbb{D}_+\left(\frac{\bar{\mathbb{D}}_+(\rho^n\psi(0))}{2}\right) - \mathbb{D}_-\left(\frac{\bar{\mathbb{D}}_-(\rho^n\psi(0))}{2}\right)\right) = \mathcal{O}(\varepsilon),$$

the scheme is still reasonable for small ε. The convergence in α of the schemes has been checked numerically. In Figure 2, we show the $L^2_{t,x}$-error in densities of the macroscopic method for $\alpha > 0$ with respect to the completely relaxed method above $\alpha = 0$ for a fixed valued of $\varepsilon = 0.01$.

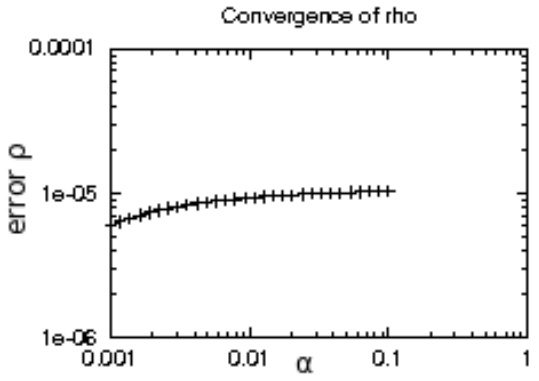


Figure 2: $L^2_{t,x}$-error of the densities ρ for the $\alpha > 0$ method with respect to the completely relaxed scheme $\alpha = 0$ for $\varepsilon = 0.01$.

Remark 4.4.4 (Well-Balanced Scheme)**.** *Note that the obtained scheme is well-balanced which means that the stationary states are preserved, if we choose linear discretizations $\mathbb{D}$: if we take some initial conditions ρ^0 and J^0 that satisfy*

$$\partial_x J^0 = 0,$$
$$\partial_x(\rho^0\psi(\varepsilon J^0/\rho^0)) = -J^0,$$

so that, in particular, $\partial^2_{xx}(\rho^0\psi(\varepsilon J^0/\rho^0) = 0$, a direct induction implies that the discrete solution $(\rho^n, J^n)_n$ is stationary, since $\mathbb{D}_\pm(J^0) = 0 = \ldots = \mathbb{D}_\pm(J^n)$ and $\mathbb{D}_\pm(\bar{\mathbb{D}}_\pm(\rho^0\psi(\varepsilon J^0/\rho^0)) = 0 = \ldots = \mathbb{D}_\pm(\bar{\mathbb{D}}_\pm(\rho^n\psi(\varepsilon J^n/\rho^n))$, for all $n \in \mathbb{N}$; see (4.36)*.*

In turn, the limit $\varepsilon \to 0$ gives

$$\rho^{n+1} = \rho^n + \frac{\Delta t}{\mu^2}\left(\mathbb{D}_+\left(\frac{\bar{\mathbb{D}}_+(\rho^n\psi(0))}{2}\right) + \mathbb{D}_-\left(\frac{\bar{\mathbb{D}}_-(\rho^n\psi(0))}{2}\right)\right).$$

Let us detail the upwind case: since we have, for any sequence $(v_j)_{j\in\mathbb{Z}}$ and for $j \in \mathbb{Z}$,

$$\begin{aligned}\mathbb{D}_+(\bar{\mathbb{D}}_+(v))_j &= \frac{-\mu}{\Delta x}\left(\left(\frac{-\mu}{\Delta x}(v_{k+1}-v_k)_k\right)_j - \left(\frac{-\mu}{\Delta x}(v_{k+1}-v_k)_k\right)_{j-1}\right),\\ &= \frac{\mu^2}{(\Delta x)^2}(v_{j+1}-2v_j+v_{j-1}), \qquad (4.35)\end{aligned}$$

$$\mathbb{D}_-(\bar{\mathbb{D}}_-(v))_j = \mathbb{D}_+(\bar{\mathbb{D}}_+(v))_j, \qquad (4.36)$$

we get the standard classical 3-point finite difference scheme for the heat equation with conduction $\psi(0)$:

$$\rho^{n+1} = \rho^n + \psi(0)\frac{\Delta t}{(\Delta x)^2}(\rho^n_{j+1} - 2\rho^n_j + \rho^n_{j-1}).$$

Remark 4.4.5 (Comparison to existing literature)**.** *We point out that the strategy differs from the one used in [18, 20] where the adopted method, based on well-balanced schemes as introduced in [56, 57], is implicit (see also [41]). The main advantage in the latter is the control on the stability condition. Note however that our method works under the parabolic CFL condition (this has to be compared to the stability analysis in [66]). This could be seen as too restrictive when the kinetic equation or the reduced model is coupled to hydrodynamics, like in applications in radiative transfer [19, 20, 41, 53], but, it is possible in such a context to appeal to a sub-cycling method where several "parabolic" time steps are performed within a "hyperbolic" time step, see [58].*

Besides, the scope of this scheme differs from that of the method described in [85] in the sense that we are interested in computations for a positive value of the parameter ε, not only for the fully relaxed situation.

4.4.2 Relaxation Method for the Zeroth-Order Closure

Here, we use again analogous ideas to propose a relaxation numerical scheme to solve the zeroth-order closure in (4.8). The nonlinear equation (4.8) can be seen as the limit, as α tends to 0, of

$$\partial_t \rho + \partial_x J = 0, \qquad (4.37)$$

$$\partial_t J + \frac{\mu^2}{\varepsilon^2}\partial_x \rho = -\frac{1}{\alpha}\left[J + \frac{\rho}{\varepsilon}\mathbb{G}\left(\varepsilon\frac{\partial_x\rho}{\rho}\right)\right]. \qquad (4.38)$$

Defining now

$$f_\pm = \frac{\rho}{2} \pm \frac{\varepsilon J}{2\mu}, \qquad (4.39)$$

we have

$$\rho = f_+ + f_- \quad \text{and} \quad J = \frac{\mu}{\varepsilon}(f_+ - f_-).$$

The new system we are interested in is

$$\partial_t f_\pm \pm \frac{\mu}{\varepsilon}\partial_x f_\pm = \frac{1}{\alpha}\left[\frac{\rho}{2} - f_\pm \mp \frac{\rho}{2\mu}\mathbb{G}\left(\varepsilon\frac{\partial_x \rho}{\rho}\right)\right]. \tag{4.40}$$

The relaxation scheme follows the same ideas as above. We define the fluctuations as

$$g_\pm = \frac{1}{\varepsilon}f_\pm - \frac{1}{2\varepsilon}\rho$$

and then, the equation rewrites as

$$\partial_t f_\pm \pm \mu\partial_x g_\pm = \frac{1}{\alpha}\left[\frac{\rho}{2} - f_\pm \mp \frac{\rho}{2\mu}\mathbb{G}\left(\varepsilon\frac{\partial_x \rho}{\rho}\right)\right] \mp \frac{\mu}{2\varepsilon}\partial_x\rho.$$

The steps of the method are:

Step 1.- Solve the ODE

$$\partial_t f_\pm = \frac{1}{\alpha}\left[\frac{\rho}{2} - f_\pm \mp \frac{\rho}{2\mu}\mathbb{G}\left(\varepsilon\frac{\partial_x \rho}{\rho}\right)\right] \mp \frac{\mu}{2\varepsilon}\partial_x\rho. \tag{4.41}$$

Step 2.- Solve the transport equation

$$\partial_t f_\pm \pm \frac{\mu}{\varepsilon}\partial_x f_\pm = 0, \tag{4.42}$$

$$\partial_t g_\pm = 0. \tag{4.43}$$

Here the initial value for the fluctuations for the second step are computed from the values of the first step by:

$$g_\pm(0) = \frac{1}{\varepsilon}f_\pm - \frac{1}{2\varepsilon}\rho$$

The first step of the scheme results into the fully discrete scheme

$$\begin{aligned} f_\pm^{n+1/2} &= e^{-\Delta t/\alpha} f_\pm^n + \frac{\rho^n}{2}(1 - e^{-\Delta t/\alpha})\left[1 \mp \frac{1}{\mu}\mathbb{G}\left(\mp\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_\pm \rho^n}{\rho^n}\right)\right] \\ &+ \alpha(1 - e^{-\Delta t/\alpha})\frac{1}{2\varepsilon}\bar{\mathbb{D}}_\pm \rho^n. \end{aligned}$$

The completely relaxed scheme, $\alpha \to 0$ is

$$f_\pm^{n+1/2} = \frac{\rho^n}{2}\left[1 + \frac{1}{\mu}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_\pm \rho^n}{\rho^n}\right)\right]$$

where the odd character of $\mathbb{G}$ was used. Taking into account the initialization of the fluctuations above, we get

$$g_{\pm}^{n+1/2} = \frac{\rho^n}{2\mu\varepsilon}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_{\pm}\rho^n}{\rho^n}\right)$$

where a term of order Δx^2 was neglected. Now, the values of the solutions in the second step are

$$f_{\pm}^{n+1} = f_{\pm}^{n+1/2} + \Delta t \mathbb{D}_{\pm} g_{\pm}^{n+1/2}$$

respectively. The use of alternate approximations of the spatial derivatives $\mp\mu\partial_x$ is again needed since for all $A \in \mathbb{R}$

$$\lim_{\varepsilon\to 0}\frac{\rho}{\varepsilon}\mathbb{G}\left(\varepsilon\frac{A}{\rho}\right) = \frac{1}{3}A.$$

The complete relaxed scheme in terms of the macroscopic variable ρ reads as

$$\rho^{n+1} = \rho^n + \Delta t\left\{\mathbb{D}_+\left[\frac{\rho^n}{2\varepsilon\mu}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_+\rho^n}{\rho^n}\right)\right] + \mathbb{D}_-\left[\frac{\rho^n}{2\varepsilon\mu}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_-\rho^n}{\rho^n}\right)\right]\right\}. \tag{4.44}$$

Due to the diffusive character of the approximation, a parabolic CFL condition for small ε-values, $d\Delta t/(\Delta x)^2 \leq 1/2$, has to be imposed. In this case, following a Chapman-Enskog approach there is no restriction in principle on the value of $\mu > 0$ but being of order 1 with respect to ε. The boundary conditions are standard discrete Neumann conditions for ρ.

4.5 Numerical Results

4.5.1 Comparisons between Closures

To start with, we compare the solution of the kinetic equation (4.1)-(4.2), computed with the method described in Section 4.3, and the solutions of the heat equation (4.4), the zeroth order closure (4.8), and the first order closure (4.9); the two last models are evaluated by using the method described in Section 4.4. Figure 3 shows the error in a log-log plot with respect to ε, for the symmetric initial data

$$f_0(x,v) = \begin{cases} 2. & \text{for } -0.5 \leq x \leq 0.5 \text{ and } -0.5 \leq v \leq 0.5 \\ 1. & \text{otherwise} \end{cases}$$

with mesh $N_x = N_v = 100$ and up to time 5. We used the completely $\alpha-$relaxed version ($\alpha = 0$) of the schemes in Section 4.4.

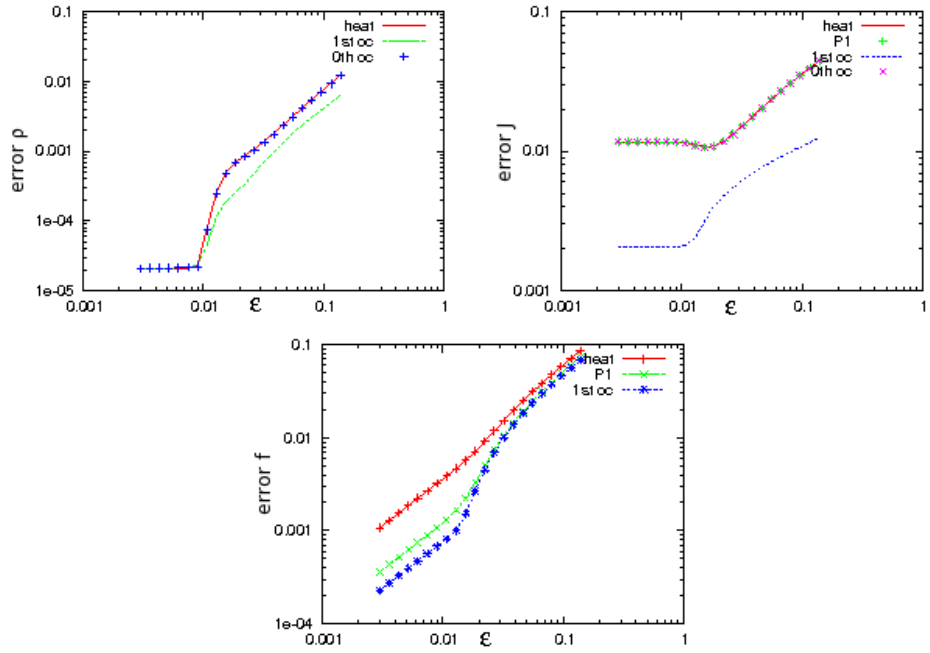


Figure 3: Top left: $L^2_{t,x}$ density error, top right: $L^2_{t,x}$ current error, bottom: $L^2_{t,x,v}$ distribution function error between the kinetic result and the corresponding approximations with respect to ε for the symmetric initial data.

As expected the convergence rates are of order $\mathcal{O}(\varepsilon)$ for all models, confirming the results in [36]. Note however that the macroscopic density ρ is better reproduced by the first order closure and the behavior of the current J is even better captured by this model. For very small values of ε, the density error becomes constant: it is actually dominated by the consistency error, with an error of order $\mathcal{O}((\Delta x)^2)$ (confirmed by changing the mesh size). This is not surprising when thinking of the 3−point scheme and is due to the splitting method.

Next, we consider a data which is not symmetric with respect to velocity. Figure 4 shows the error in a log-log plot with respect to ε, for the initial data

$$f_0(x,v) = \begin{cases} 2. & \text{for } -0.5 \le x \le 0.5 \text{ and } -0.75 \le v \le 0.25 \\ 1. & \text{otherwise} \end{cases}$$

with mesh $N_x = N_v = 100$ and up to time 5. We still use the completely relaxed framework $\alpha = 0$ The previous conclusions are amplified and the advantage of the first order closure appears more strongly, in agreement with conclusions already given in [41].

This is confirmed again by looking at the time evolution of the density

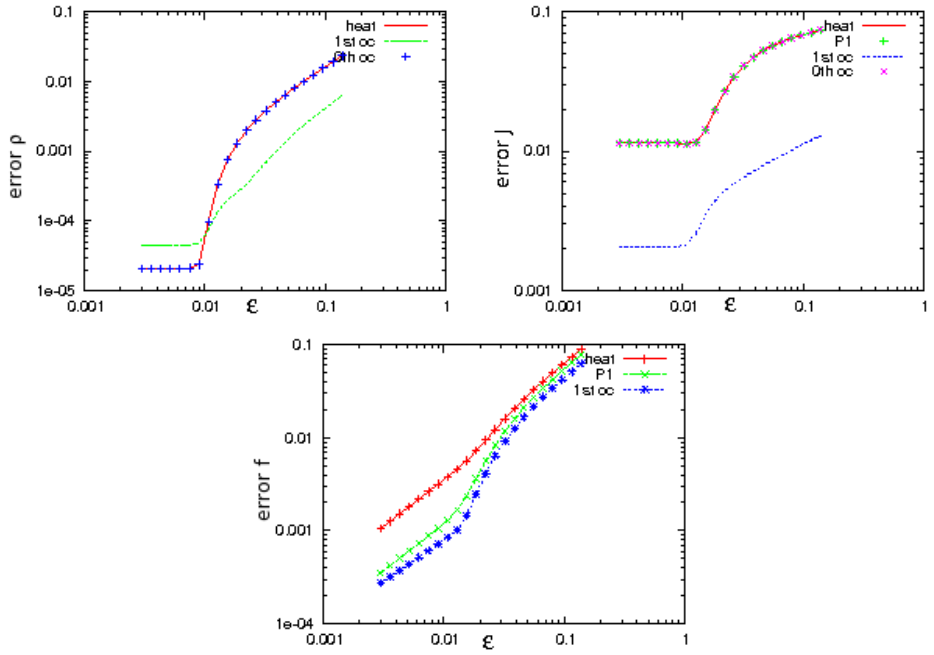


Figure 4: Top left: $L^2_{t,x}$ density error, top right: $L^2_{t,x}$ current error, bottom: $L^2_{t,x,v}$ distribution function error between the kinetic result and the corresponding approximations with respect to ε for the asymmetric initial data.

and current computed by the different models. Figure 5 corresponds to the evolution of the macroscopic density for the asymmetric initial data with a mesh of $N_x = N_v = 100$ with $\varepsilon = 0.1$ and completely relaxed $\alpha = 0$ and up to time 5. In Figure 6 we show the corresponding evolution for the first moment J. These results favor on the one hand the kinetic and the first order simulation which remain very close, even in this situation where ε is not particularly small, and on the other hand the zeroth order model which behaves like the heat equation, far from the profiles obtained by the kinetic computations.

4.5.2 The Su-Olson Test

This test is a standard benchmark for radiative transfer problems [87, 99, 19, 20, 18]. In such problems, the unknown f is the specific intensity of radiations, which interact with the matter through energy exchanges, see e.g. [19, 53]. Therefore, the complete models couples a kinetic equation with the Euler system describing the evolution of the matter. In the Su-Olson test, the coupling with hydrodynamics is replaced by a simple ODE describing the evolution of the material temperature. More precisely, we

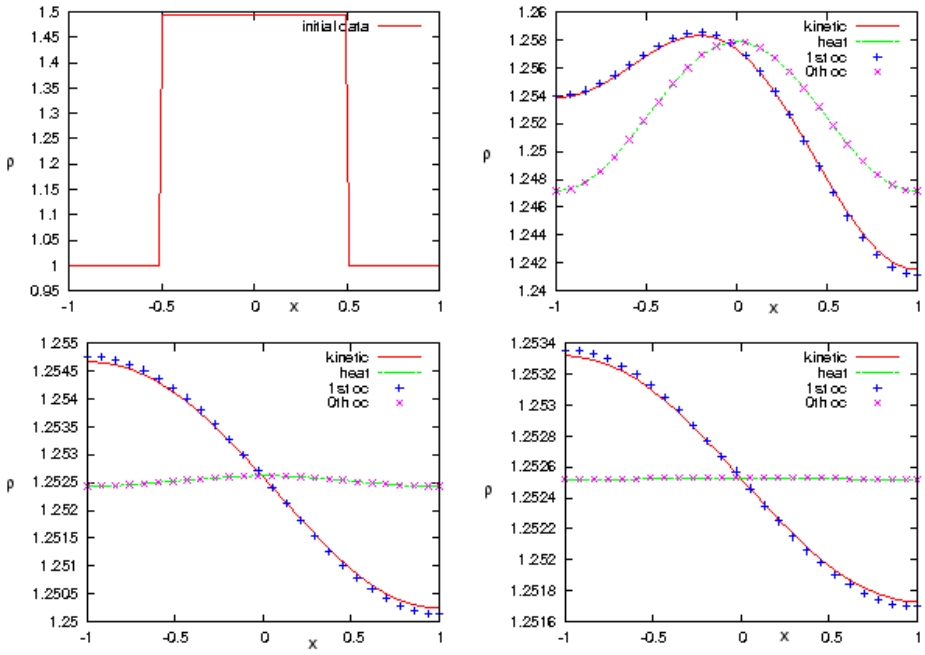


Figure 5: Evolution of the density for the different methods with the asymmetric initial data: top left: initial data, top right: 1.25 time units, bottom left: 2.5 time units, bottom right: 3.75 time units.

have the kinetic equation

$$\partial_t f_\varepsilon + \frac{v}{\varepsilon}\partial_x f_\varepsilon = \frac{1}{\varepsilon^2} Q(f_\varepsilon) + \sigma_a(\Theta - \rho) + S \tag{4.45}$$

coupled with

$$\partial_t \Theta = \sigma_a(\rho - \Theta) \tag{4.46}$$

where $\Theta = T^4$ with $T > 0$ the temperature of matter and $S = S(t,x)$ a given source. We propose to solve this stiff coupled problem (4.45)-(4.46) with the same approach as in the previous Subsections. We shall also adapt the scheme for replacing the kinetic equation by the zeroth or first order closures.

We solve the temperature equation at the steps in which the density ρ is constant to have an explicit formula for its solution. We start with the kinetic scheme and we follow the same notation as in Subsection 2.1 skipping some detail. The semi-discrete numerical scheme will summarize as follows:

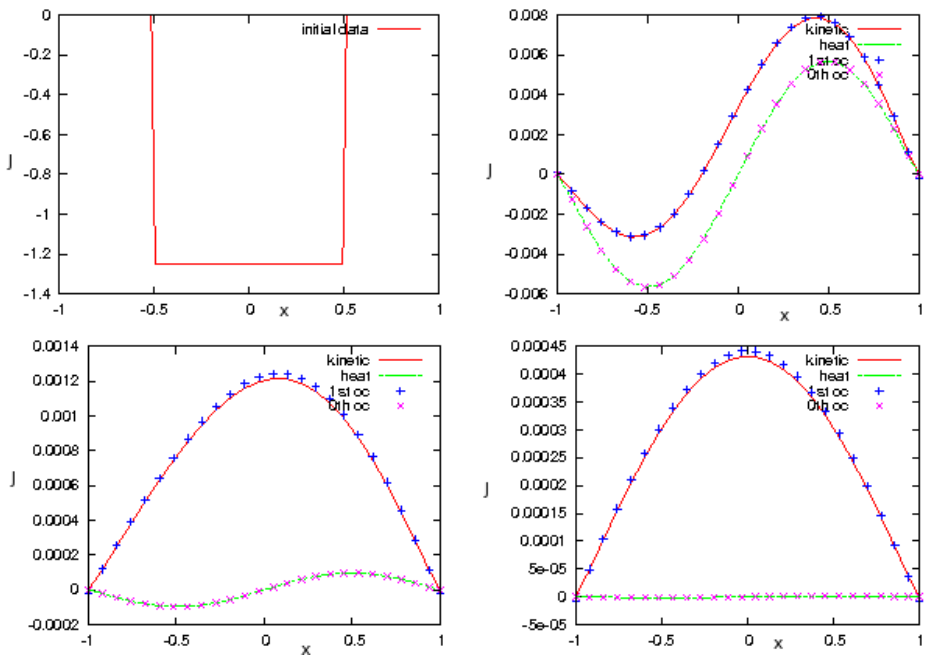


Figure 6: Evolution of the current for the different methods with the asymmetric initial data: top left: initial data, top right: 1.25 time units, bottom left: 2.5 time units, bottom right: 3.75 time units.

Step 1.- Compute

$$\begin{cases} g^{n+1/2} = e^{-\Delta t/\varepsilon^2} g^n - (1 - e^{-\Delta t/\varepsilon^2}) v \partial_x \rho^n, \\ f^{n+1/2} = e^{-\Delta t/\varepsilon^2} f^n + (1 - e^{-\Delta t/\varepsilon^2}) \rho^n, \\ \Theta^{n+1/2} = e^{-\sigma_a \Delta t} \Theta^n + \sigma_a (1 - e^{-\sigma_a \Delta t}) \rho^n \end{cases} \tag{4.47}$$

Remember that $\rho^{n+1/2} = \rho^n$.

Step 2.- Solve on a time interval of length Δt the convection equation:

$$\partial_t f + v \partial_x g = \sigma_a (\Theta - \rho) + S$$

to compute the values of f^{n+1} and ρ^{n+1} while $g^{n+1} = g^{n+1/2}$ and $\Theta^{n+1} = \Theta^{n+1/2}$. The right-hand side uses the final value provided by Step 1.

Similar schemes have to be written for the zeroth and first order closures of the Su-Olson test. We start with the first order closure in Subsection 2.2,

keeping the notation used therein. The system to solve reads

$$\begin{cases} \partial_t \varrho + \partial_x J = \sigma_a(\Theta - \rho) + S, \\ \varepsilon^2 \partial_t J + \partial_x \big(\varrho \psi(\varepsilon J/\varrho)\big) = -J \\ \partial_t \Theta = \sigma_a(\rho - \Theta). \end{cases} \tag{4.48}$$

The nonlinear system (4.48) can be seen as the limit, as α tends to 0, of

$$\begin{cases} \partial_t \rho + \partial_x J = \sigma_a(\Theta - \rho) + S, \\ \varepsilon^2 \partial_t J + \partial_x z = -J, \\ \partial_t z + \varepsilon^2 \lambda^2 \partial_x J = \dfrac{1}{\alpha}\big(\rho\psi(\varepsilon J/\rho) - z\big) \\ \partial_t \Theta = \sigma_a(\rho - \Theta). \end{cases} \tag{4.49}$$

The kinetic scheme will be summarized as

Step 1.- Solve

$$\begin{aligned} \partial_t f_0 &= \mu^2 \left(\sigma_a(\Theta - \rho) + S\right), \\ \partial_t f_\pm \pm \frac{\mu}{\varepsilon}\partial_x f_\pm &= -\frac{f_\pm}{\varepsilon^2} + \frac{z}{2\varepsilon^2}, \\ \partial_t \Theta &= \sigma_a(\rho - \Theta), \end{aligned}$$

that can be computed as

Step 1.1.-
$$\begin{aligned} &\partial_t f_0 = 0, \\ &\partial_t f_\pm = -\frac{f_\pm}{\varepsilon^2} + \frac{z}{2\varepsilon^2} \mp \frac{\mu}{2\varepsilon}\partial_x z, \\ &\partial_t g_\pm = -\frac{g_\pm}{\varepsilon^2} \mp \frac{\mu}{2\varepsilon^2}\partial_x z, \\ &\partial_t \Theta = \sigma_a(\rho - \Theta), \end{aligned}$$

where the initial condition for the ODEs are the values computed in the previous step and

Step 1.2.-
$$\begin{aligned} &\partial_t f_0 = \mu^2 \left(\sigma_a(\Theta - \rho) + S\right) \\ &\partial_t f_\pm \pm \mu \partial_x g_\pm = 0, \\ &\partial_t g_\pm = 0, \\ &\partial_t \Theta = 0 \end{aligned}$$

where the initial conditions are, for $f_\pm$, the ones obtained in Step 1.1. For $g_\pm$, we update them in terms of the flux $g_\pm(0) = \pm\mu J/2 = \pm(f_+ - f_-)/2\varepsilon$.

Step 2.- Solve the ODE

$$\begin{aligned}\partial_t f_0 &= -\frac{1}{\alpha}(\rho\psi(u)-z),\\ \partial_t f_\pm &= \frac{1}{2\alpha}(\rho\psi(u)-z),\\ \partial_t \Theta &= 0.\end{aligned}$$

This kinetic scheme in macroscopic variables is

$$\begin{aligned}z^{n+1/2} &= z^n + \frac{\varepsilon(1-e^{-\Delta t/\varepsilon^2})}{2}\left(\bar{\mathbb{D}}_+(z^n)+\bar{\mathbb{D}}_-(z^n)\right)\\ &\quad +\Delta t\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\ &\quad \left.+\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right],\end{aligned}$$

$$\begin{aligned}J^{n+1/2} &= e^{-\Delta t/\varepsilon^2}J^n + \frac{1-e^{-\Delta t/\varepsilon^2}}{2\mu}\left(\bar{\mathbb{D}}_+(z^n)-\bar{\mathbb{D}}_-(z^n)\right)\\ &\quad +\frac{\Delta t}{\varepsilon\mu}\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\ &\quad \left.-\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right],\end{aligned}$$

$$\Theta^{n+1/2} = e^{-\sigma_a\Delta t}\Theta^n + \sigma_a(1-e^{-\sigma_a\Delta t})\rho^n,$$

$$\begin{aligned}\rho^{n+1/2} &= \rho^n + \frac{\Delta t}{\mu^2}\left(\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(z^n)}{2}\right)\right.\\ &\quad \left.+\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(z^n)}{2}\right)\right)\\ &\quad +\Delta t\left(\sigma_a(\Theta^{n+1/2}-\rho^n)+S^n\right).\end{aligned}$$

while the second step will coincide with Step 2 of Subsection 2.2 together with $\Theta^{n+1} = \Theta^{n+1/2}$. From here, we can write the completely relaxed scheme

$$\begin{aligned}J^{n+1} &= e^{-\Delta t/\varepsilon^2}J^n + \frac{1-e^{-\Delta t/\varepsilon^2}}{2\mu}\left(\bar{\mathbb{D}}_+(\rho^n\psi^n)-\bar{\mathbb{D}}_-(\rho^n\psi^n)\right)\\ &\quad +\frac{\Delta t}{\varepsilon\mu}\left[\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(\rho^n\psi^n)}{2}\right)\right.\\ &\quad \left.-\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2}+(1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(\rho^n\psi^n)}{2}\right)\right],\end{aligned}$$

$$\begin{aligned}
\Theta^{n+1} &= e^{-\sigma_a \Delta t}\Theta^n + \sigma_a(1-e^{-\sigma_a\Delta t})\rho^n,\\
\rho^{n+1} &= \rho^n + \frac{\Delta t}{\mu^2}\left(\mathbb{D}_+\left(e^{-\Delta t/\varepsilon^2}\frac{\mu J^n}{2} + (1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_+(\rho^n\psi^n)}{2}\right)\right.\\
&\qquad \left. +\mathbb{D}_-\left(e^{-\Delta t/\varepsilon^2}\frac{(-\mu J^n)}{2} + (1-e^{-\Delta t/\varepsilon^2})\frac{\bar{\mathbb{D}}_-(\rho^n\psi^n)}{2}\right)\right)\\
&\quad +\Delta t\left(\sigma_a(\Theta^{n+1}-\rho^n)+S^n\right).
\end{aligned}$$

Concerning the zeroth order closure for the Su-Olson test, we have the system

$$\begin{cases} \partial_t \varrho - \partial_x\Big(\frac{\varrho}{\varepsilon}\mathbb{G}\big(\varepsilon\frac{\partial_x\varrho}{\varrho}\big)\Big) = \sigma_a(\Theta-\rho)+S,\\ \partial_t\Theta = \sigma_a(\rho-\Theta), \end{cases} \tag{4.50}$$

that can be seen as the relaxation, when α tends to 0, of

$$\partial_t\rho + \partial_x J = \sigma_a(\Theta-\rho)+S,$$

$$\partial_t J + \frac{\mu^2}{\varepsilon^2}\partial_x\rho = -\frac{1}{\alpha}\left[J + \frac{\rho}{\varepsilon}\mathbb{G}\left(\varepsilon\frac{\partial_x\rho}{\rho}\right)\right],$$

$$\partial_t\Theta = \sigma_a(\rho-\Theta).$$

Proceeding similarly to Subsection 2.4 and as above for the first order closure, we conclude the completely relaxed scheme for the density and temperature is

$$\begin{aligned}
\Theta^{n+1} =& e^{-\sigma_a\Delta t}\Theta^n + \sigma_a(1-e^{-\sigma_a\Delta t})\rho^n.\\
\rho^{n+1} =& \rho^n + \Delta t\left\{\mathbb{D}_+\left[\frac{\rho^n}{2\varepsilon\mu}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_+\rho^n}{\rho^n}\right)\right] + \mathbb{D}_-\left[\frac{\rho^n}{2\varepsilon\mu}\mathbb{G}\left(\frac{\varepsilon}{\mu}\frac{\bar{\mathbb{D}}_-\rho^n}{\rho^n}\right)\right]\right\}\\
&+\Delta t\left(\sigma_a(\Theta^{n+1}-\rho^n)+S^n\right),
\end{aligned} \tag{4.51}$$

For intermediate values of the parameter ε, it is worth comparing the results obtained with the models described above to the simulations based on the semi-lagrangian SL-WENO scheme already discussed in Section 4.3. Indeed, remind that the asymptotic kinetic scheme was developed for the asymptotic limit $\varepsilon \to 0$. Similarly, the range of validity of the macroscopic models is also restricted to small ε's; furthermore, the theoretical results in [36] prove the validity of these models for density values close to constant and far from vacuum. Hence, it is worthy to compare its results to those of the previous scheme particularly for moderate values of ε. The SL-WENO scheme for the Su-Olson test summarizes as follows:

Step 1.- Relax f

$$\partial_t f = \frac{1}{\varepsilon^2}Q(f) + \sigma_a(\Theta-\rho)+S$$

Step 2.- Compute advection and relax the temperature

$$\partial_t f + \frac{v}{\varepsilon}\partial_x f_{=}0$$

$$\partial_t \Theta = \sigma_a(\rho - \Theta)$$

which gives the following numerical method:

Step 1.- Relax f

$$f^{n+1/2} = e^{-\Delta t/\varepsilon^2} f^n + (1 - e^{-\Delta t/\varepsilon^2}) \left[\rho^n + \varepsilon^2 \left(\sigma_a(\Theta^n - \rho^n) + S^n\right)\right]$$

$$\Theta^{n+1/2} = \Theta^n$$

Step 2.- Compute advection by an interpolation method and relax the temperature

$$f^{n+1}(x_i, v_j) = f_\varepsilon^n \left(x_i - \Delta t \frac{v_j}{\varepsilon}\right),$$

$$\Theta^{n+1} = e^{-\sigma_a \Delta t}\Theta^{n+1} + \sigma_a(1 - e^{-\sigma_a \Delta t})\rho^{n+1}.$$

For the simulations, the source term $S(x)$ has been chosen as the characteristic function of the interval $[0, 1]$ inside the total interval $[0, 30]$ with $\varepsilon = 0.01$ and $\varepsilon = 0.26$ respectively. We refer to the results in [87, 99, 18, 20] for comparison. The solutions of the macroscopic models are computed with the complete relaxed methods $\alpha = 0$ with mesh $N_x = 256$ and $N_v = 256$. The traditional test considers as initial data the constant equilibrium value 10^{-10} for $f_0 = \rho_0 = \Theta_0$. The smallness of this value makes the simulation particularly tough; hence, we also perform the computations with $f_0 = \rho_0 = \Theta_0 = 1$. We make different runs, with ε varying from 0.026 to 0.26. The numerical results are displayed in Figures 7 to 12.

A first conclusion is that the SL-WENO code is highly sensitive to the changes of ε, see Figure 7-(i) and(j), as already seen above, see Figure 1; we believe that the results become relevant only for the largest values of ε ($\varepsilon = 0.26$, $\varepsilon = 1$), see Figures 10, 11, 12. The result in the case $\varepsilon = 0.26$ is surprisingly close to the solution of the heat equation. This is a bit misleading since in this regime there is no reason why the heat equation can describe the dynamics of the kinetic equation well. Other tests with direct finite-differences WENO schemes as the ones used in [30] may be interesting to clarify this point, although not directly linked to the asymptotic discussion in this paper, and thus it will be treated elsewhere.

We observe that the results given by the heat equation, the two closure models and the kinetic scheme are almost undistinguishable from each other up to final time 10, for small ε's, see Figures 8, 9. Differences appear as ε grows and correspond to the results in [18, 20]. There are discrepancies

between the diffusion model, the other macroscopic models and the kinetic equation, especially for earlier times. These discrepancies reduce as time grows. It is also worth pointing out that, as in [18, 20] and contrarily to [87, 99], the results are oscillation free for the first order closure, both for the density ρ and the reduced flux $\varepsilon J/\rho$, which remains bounded by 1, as expected.

The kinetic scheme is also sensitive to the variations of ε, particularly for the almost vanishing initial data, see Figure 7-(g) and (h). We observe that the results differ from the ones given by SL-WENO for $\varepsilon = 0.26$ and almost vanishing initial data: the main errors appear in the regions of large gradient of the density, see Figure 10. Clearly the kinetic scheme is not well adapted for this regime for such a small initial data. However, the performances are better considering a larger initial data, since in such a case the slopes are less steep see Figures 11 and 12. Note that this test also shows the limitation of the asymptotic-induced method since when ε grows we are faced with difficulties related to the maximum principle, see Remark 4.3.5. In particular, the scheme for the first order closure is not positive in the sense of [20], the computed $\varepsilon J/\rho$ can violate the limited flux condition and we are in trouble to evaluate the flux (again, increasing the initial data makes things easier). Finally, it is remarkable to observe that the first order closure results are satisfactory in all regimes. This makes this closure model really valuable. More figures are available on the URL `http://diffnum.gforge.inria.fr/SU-OLSON/`.

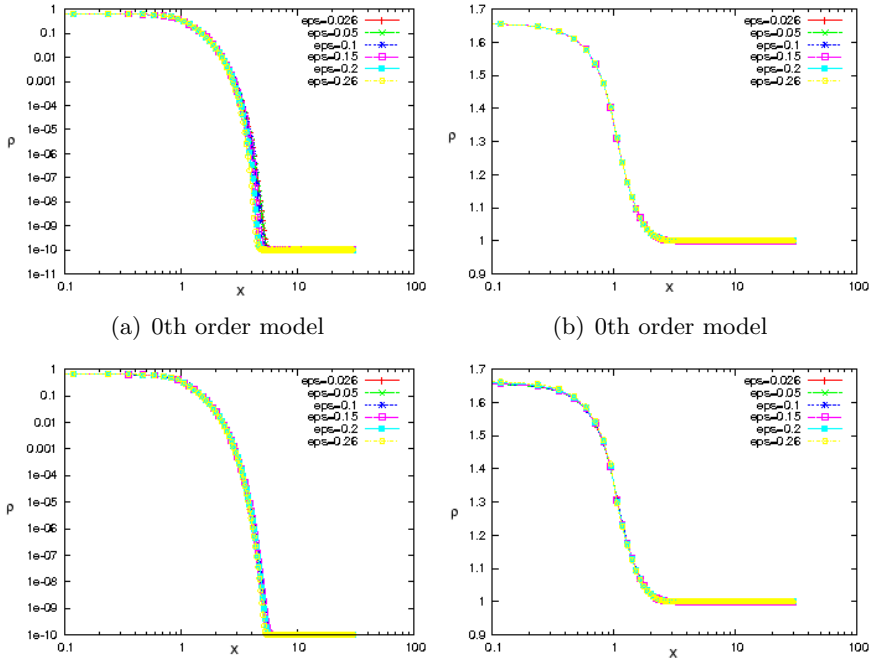
4.6 Conclusion

We have proposed new numerical schemes based on splitting techniques specifically adapted to diffusion regimes. The main idea behind this strategy is the separation between the hydrodynamic quantities and the fluctuations. Hence, the method we design is explicit, asymptotic preserving, well balanced and mass preserving thanks to a suitable treatment of the numerical boundary conditions. This approach applies equally well to the original kinetic equation and to the macroscopic models coming from closure approximations.

The numerical experiments demonstrate the abilities of the scheme to give accurate quantitative estimates of the errors made by the approximations to the kinetic equation. The first order closure is shown to be the most accurate approximation, among those we chose, for the kinetic equation in the diffusive limit. This confirms that the choice of the closure by entropy minimization principle is certainly appropriate for applications where the kinetic equation is coupled with more complex systems.

Acknowledgements

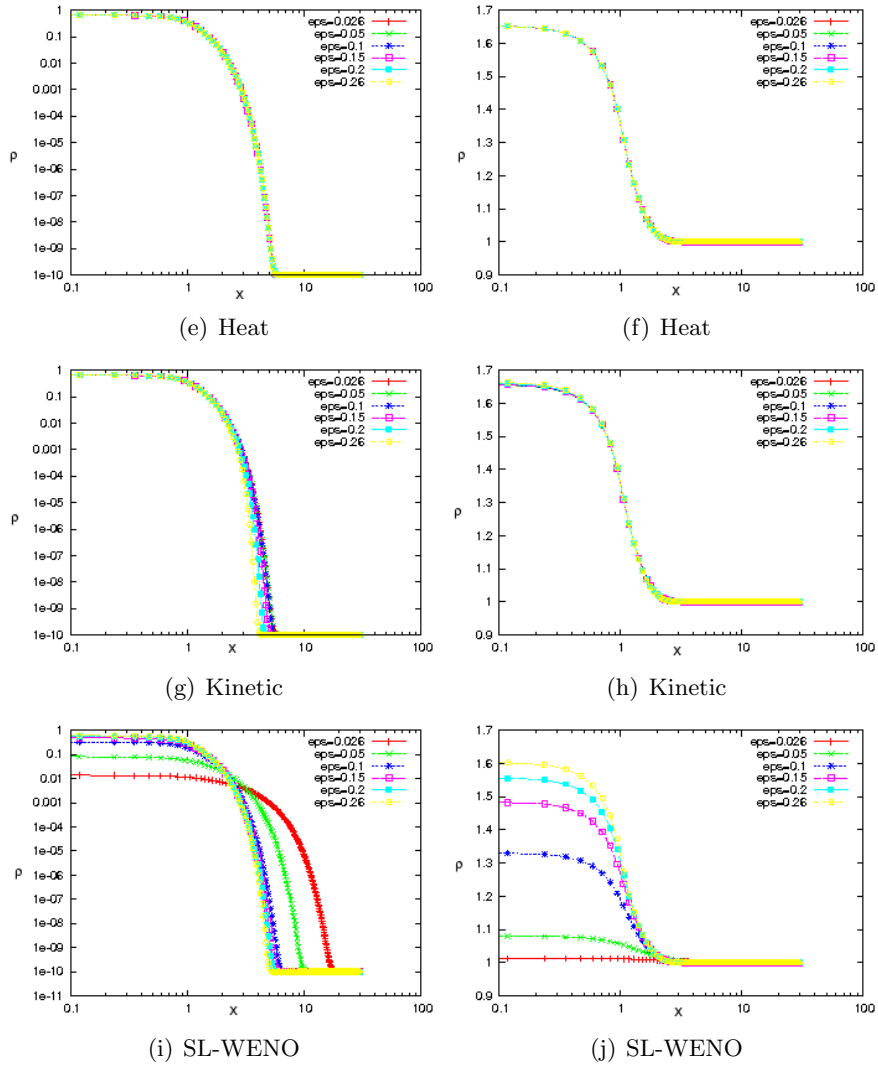
The authors acknowledge for the computational resources of the Grid'5000 project which made the simulations possible. TG and PL thank the Centre de Recerca Matemàtica, Barcelona, Catalonia for its warm hospitality. JAC and FV thank INRIA for the invitation in Lille and the support from the Spanish DGI-MCYT/FEDER project MTM2005-08024. The authors acknowledge partial support of the bilateral project France-Spain HF2006-0198.



(a) 0th order model (b) 0th order model

(c) 1st order model (d) 1st order model

Figure 7: Su-Olson test: Comparison of the density ρ computed by the different models as ε varies at time $t = 1$. From top to bottom: 0th order model, 1st order model, heat equation, kinetic asymptotic-induced model, SL-WENO scheme. Left column: results in log-log scale for the initial data $f_0 = \rho_0 = \Theta_0 = 10^{-10}$; Right column: results in semi-log scale initial data $f_0 = \rho_0 = \Theta_0 = 1$.



(e) Heat (f) Heat

(g) Kinetic (h) Kinetic

(i) SL-WENO (j) SL-WENO

Figure 7: Su-Olson test: Comparison of the density ρ computed by the different models as ε varies at time $t = 1$ (continued).

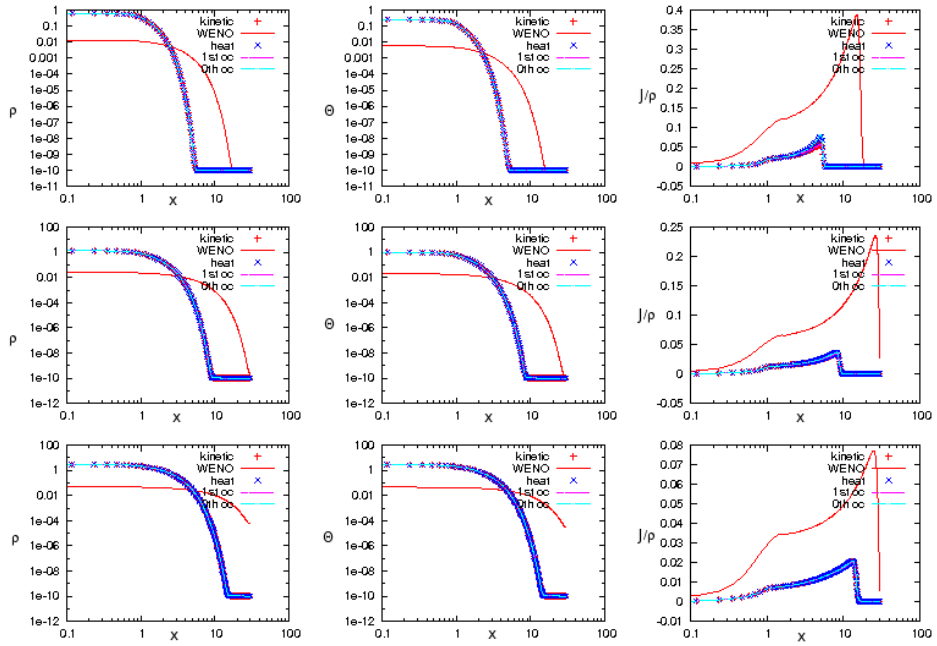


Figure 8: Su-Olson test: Left column: comparison of densities ρ; middle column: comparison of temperatures Θ ; right column : comparison of reduced fluxes $\varepsilon J/\rho$ in log-log scales for the solutions after time 1, 3 and 10 time units respectively (from top to bottom) computed with the kinetic, the heat equation, the first and the zeroth order closure methods for $\varepsilon = 0.026$. The initial data is $f_0 = \rho_0 = \Theta_0 = 10^{-10}$.

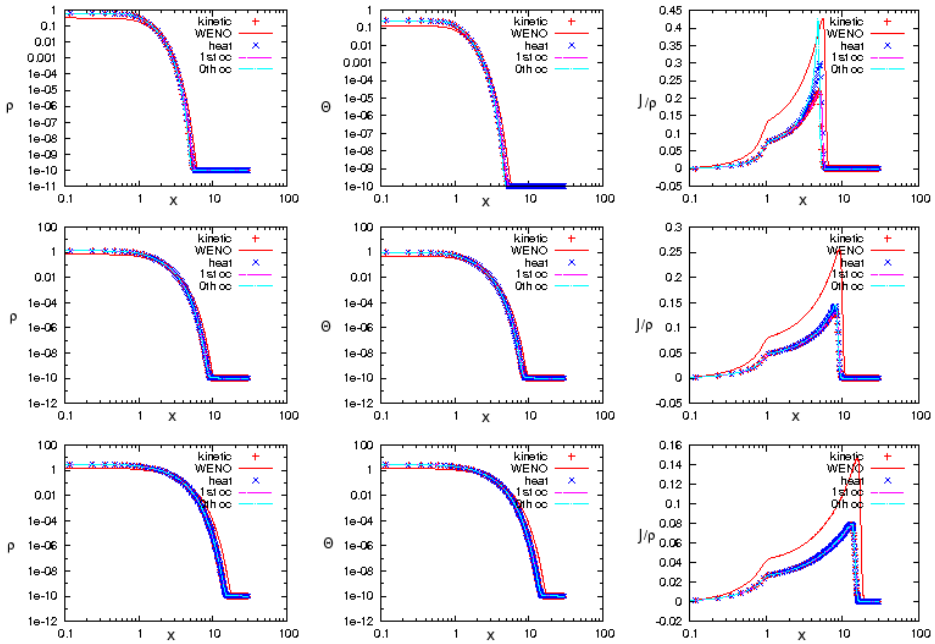


Figure 9: Su-Olson test: Left column: comparison of densities ρ; middle column: comparison of temperatures Θ ; right column : comparison of reduced fluxes $\varepsilon J/\rho$ in log-log scales for the solutions after time 1, 3 and 10 time units respectively (from top to bottom) computed with the kinetic, the heat equation, the first and the zeroth order closure methods for $\varepsilon = 0.1$. The initial data is $f_0 = \rho_0 = \Theta_0 = 10^{-10}$.

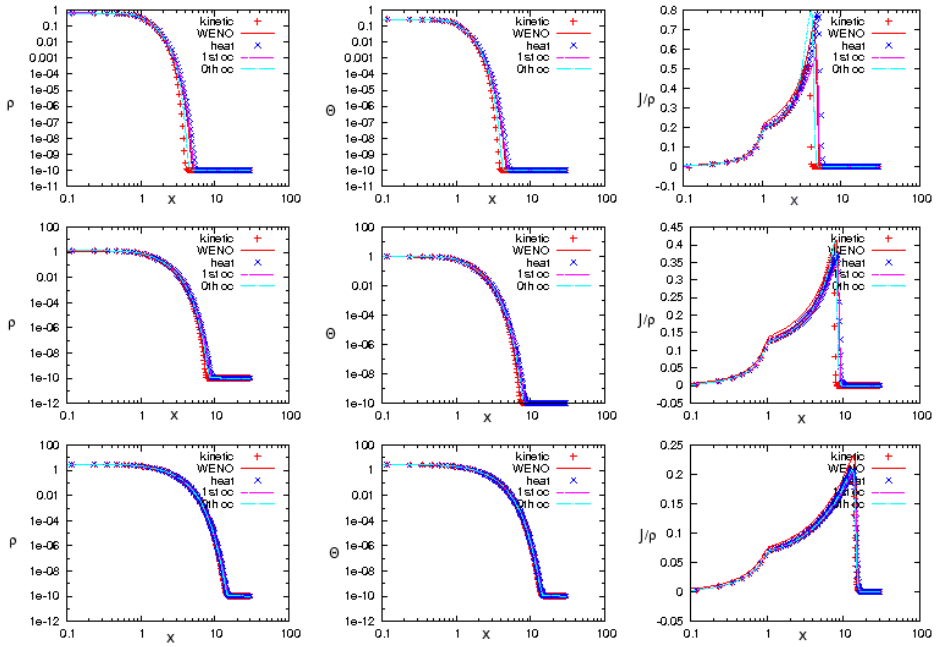


Figure 10: Su-Olson test: Left column: comparison of densities ρ; middle column: comparison of temperatures Θ ; right column : comparison of reduced fluxes $\varepsilon J/\rho$ in log-log scales for the solutions after time 1, 3 and 10 time units respectively (from top to bottom) computed with the kinetic, the heat equation, the first and the zeroth order closure methods for $\varepsilon = 0.26$.The initial data is $f_0 = \rho_0 = \Theta_0 = 10^{-10}$.

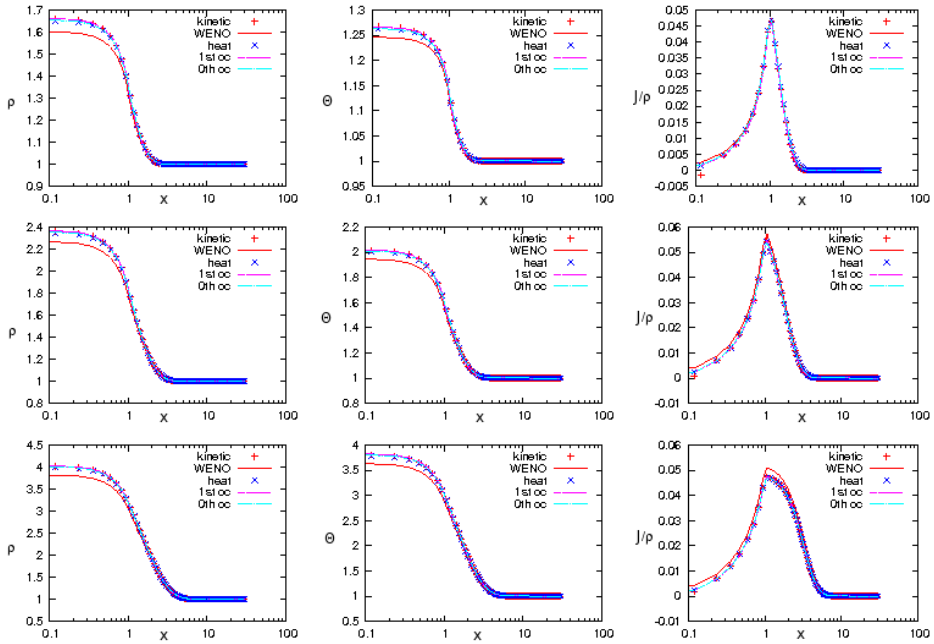


Figure 11: Su-Olson test: Left column: comparison of densities ρ; middle column: comparison of temperatures Θ ; right column : comparison of reduced fluxes $\varepsilon J/\rho$ in log-log scales for the solutions after time 1, 3 and 10 time units respectively (from top to bottom) of with the kinetic, the heat equation, the first and the zeroth order closure methods for $\varepsilon = 0.26$. The initial data is $f_0 = \rho_0 = \Theta_0 = 1$.

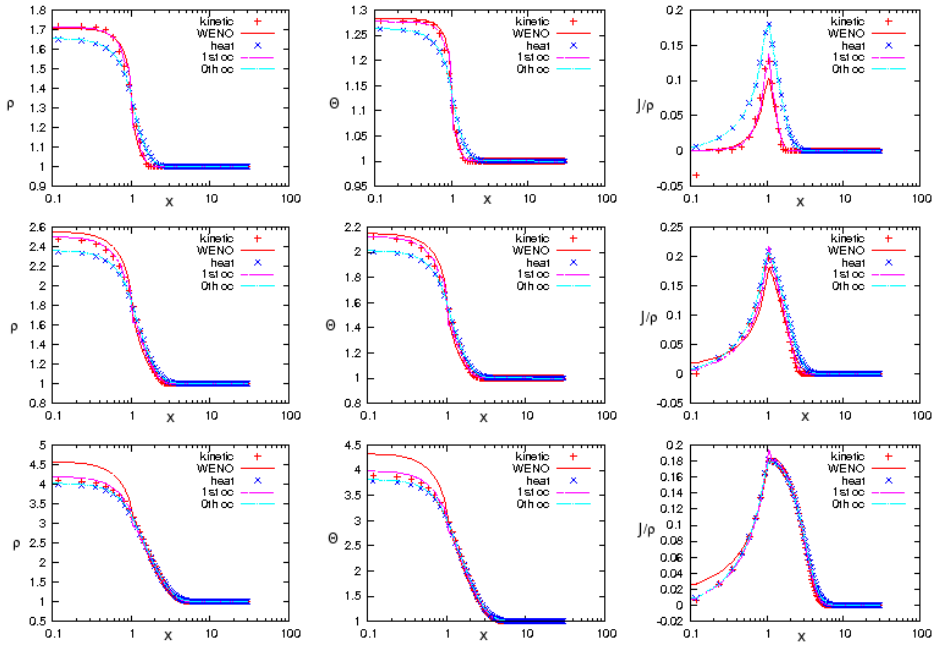


Figure 12: Su-Olson test: Left column: comparison of densities ρ; middle column: comparison of temperatures Θ ; right column : comparison of reduced fluxes $\varepsilon J/\rho$ in log-log scales for the solutions after time 1, 3 and 10 time units respectively (from top to bottom) computed with the kinetic, the heat equation, the first and the zeroth order closure methods for $\varepsilon = 1$. The initial data is $f_0 = \rho_0 = \Theta_0 = 1$.

Chapter 5

A Semi-lagrangian deterministic solver for a hybrid quantum-classical nanoMOSFET

This Chapter refers to a work with Naoufel Ben Abdallah from Toulouse and José Antonio Carrillo from Barcelona. It is still in progress, and no paper has been submitted yet, but the first results we have got are encouraging for going on on this path.

5.1 Introduction

The deterministic simulation of double gate MOSFETs is an important scientific computing problem in electrical engineering. The typical size of current MOSFETs in nowadays research has decreased dramatically and quantum effects can no longer be neglected. Very Large Scaled Integrated (VLSI) chips are designed using these basic MOSFETs as bricks of circuits with complicated topologies.

Here, we follow an extended approach in the electrical engineering community in this kind of devices consisting in a different description depending on the dimension. A typical MOSFET geometry can be seen in Figure 1. Usually, the transport in the longer dimension, x-space variable, is considered as classical while the confinement in the shorter dimension, z-space variable, is quantum-mechanically modeled.

Therefore, the PDE models considered in each dimension are different, see Figure 2. In the x-dimension, the electrons behave like particles, so a classical description is satisfactory. In the z-dimension, as its length is comparable with the Abbye length, the energy levels become quantized and the

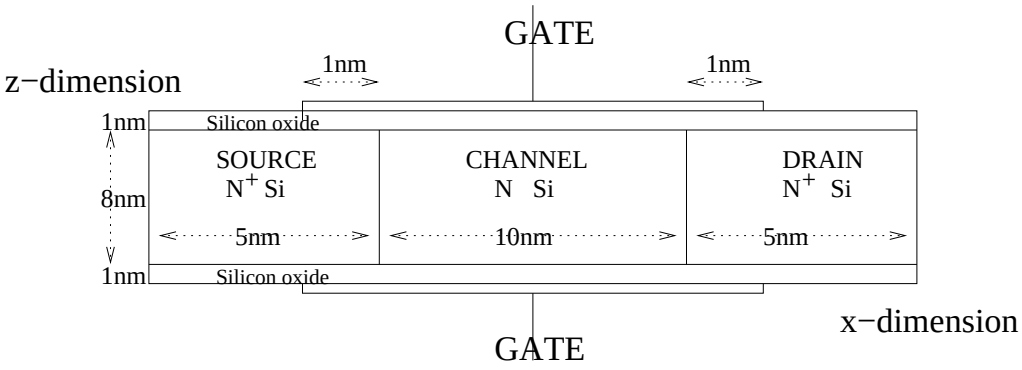


Figure 1: A typical double gate MOSFET.

electrons behave like waves, so a quantum description via a wave-function is adopted. Since electrons in different energy levels or bands have to be considered as independent populations, we have to use a structured population model in energy-bands, i.e., an equation for each of these single energy-band populations. We refer to [3, 4, 5, 101] for a detailed description of these models and some properties of these models in case a drift-diffusion approximation is chosen for the classical transport.

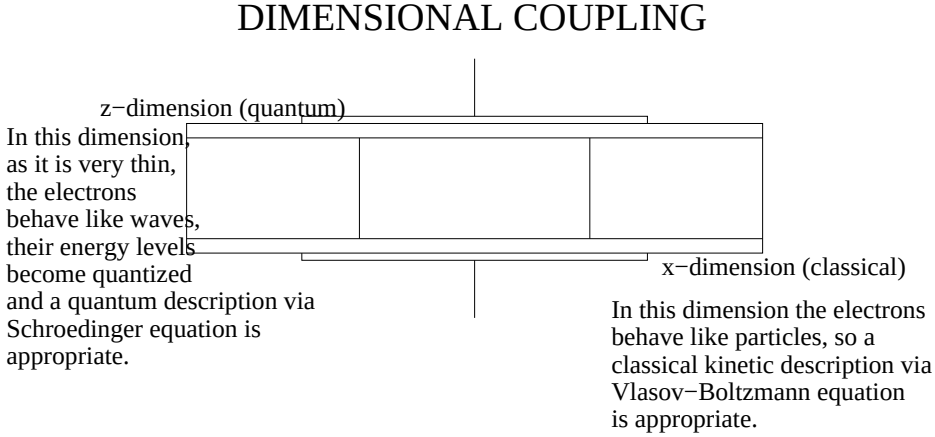


Figure 2: The coupling strategy is dimensional. We shall describe the electrons in the x-dimension as particles, and in the z-dimension as waves. The different descriptions are coupled in Poisson's equation through the total density number.

The coupling comes through the computation of the self-consistent electric field which depends on the total number of electrons in each of the energy bands. Also, the collision operator may couple the sub-bands, i.e.,

describe the electron jumps between energy levels or bands.

Then, the problem mainly divides into two blocks. On one hand, the solution of the Boltzmann Transport Equation (BTE) for each of the sub-bands p gives the evolution of the distribution $f_p(t,x,k)$ in phase space ($(x,k)=(x,k_1,k_2)\in[0,L]\times\mathbb{R}^2$) of the p-electron population

$$\frac{\partial f_p}{\partial t}+\frac{1}{\hbar}\nabla_k\epsilon_p^{kin}\cdot\nabla_x f_p-\frac{1}{\hbar}\nabla_x\epsilon_p^{pot}\cdot\nabla_k f_p=\mathcal{Q}_p[f_p] \tag{5.1}$$

plus boundary conditions

where $\hbar$ is the reduced Planck constant. On the other hand, the solution of the Schrödinger-Poisson (SP) system allows to compute the potential $V(x,z)$ the sub-band potential energies $\epsilon_p^{pot}(x)$, given the occupation numbers

$$\rho_p=\int_{\mathbb{R}^2} f_p dk$$

from the equations

$$-\frac{d}{dz}\left[\frac{1}{m_*(z)}\frac{d\chi_p}{dz}\right]-q\left(V+V_c\right)\chi_p=\epsilon_p^{pot}\chi_p \tag{5.2}$$

$$\{\chi_p\}_p\subseteq H_o^1,\qquad \langle\chi_p,\chi_q\rangle=\int\chi_p\chi_q dz=\delta_{p,q}$$

$$-\mathrm{div}_{x,z}\left[\varepsilon_R(x,z)\nabla_{x,z}V\right]=-\frac{q}{\varepsilon_0}\left[N-N_D\right] \tag{5.3}$$

plus boundary conditions,

Here, $m_*(z)$ is the effective electron mass, q the positive elementary charge and V_c the confining potential, i.e. the build-in potential drop near the SiO_2 layer which confines the carriers along the z-dimension, ε_0 is the vacuum dielectric permittivity, ε_R the possibly spatial-dependent relative dielectric permittivity, N the total density, which is a mixed quantum-classical state, sum of the densities of all the subbands

$$N(t,x,z) \;=\; \sum_p N_p(t,x,z)=\sum_p\int_{\mathbb{R}^2} f_p(t,x,k)dk|\chi_p(t,x,z)|^2,$$

and N_D is the doping profile which takes into account the injected impurities in the semiconductor lattice. The time t and the position x just act as parameters in this second block and were omitted. Also, we follow shortcuts H_o^1 for the Sobolev space $H_o^1(0,l_z)$ and the integral symbol for the integral on the interval $z\in(0,l_z)$.

The collisions are described in the low-density approximation

$$\mathcal{Q}_p[f_p]=\sum_{p'}\int_{\mathbb{R}^2}\alpha_{p,p'}\left[M_p(k)f_{p'}(k')-M_{p'}(k')f_p(k)\right]dk',$$

where $M(k)$ is the Maxwellian

$$M(k) = \frac{\hbar^2}{2\pi m_* k_B T_L} \exp\left(-\frac{\hbar^2|k|^2}{2m_* k_B T_L}\right).$$

For the scope of this work, we shall use a relaxation time operator

$$\mathcal{Q}_p[f_p] = \frac{1}{\tau}\left[M(k)\rho_p - f_p(k)\right],$$

where the relaxation time τ is chosen from the relation with the mobility μ

$$\tau = \frac{\mu m_*}{q}.$$

The band structure of the semiconductor crystal is described by the band energy function $\epsilon_p(t, x, k)$, which has contribution of a kinetic part

$$\epsilon_p^{kin}(k) = \frac{\hbar^2|k|^2}{2m_* k_B T_L}$$

and a potential part which results from the solution of a Schrödinger equation along the z-dimension (5.2). We simplify in the following the notation for the potential energy for the p^{th} band in the rest: $\epsilon_p^{pot} \equiv \epsilon_p$. Then, the velocities are

$$v_p(k) = \frac{1}{\hbar}\nabla_k \epsilon_p^{kin}(k) = \frac{\hbar k}{m_*} = v(k), \tag{5.4}$$

and they do not depend on the subband; this is strictly related to the parabolic band approximation. If we were using the Kane dispersion, then the velocities would be band-dependent and the problem would become more delicate to treat.

The total system (5.2)-(5.3) has to be completed by adding suitable initial and boundary conditions for this problem.

Initial condition.

As initial condition, we impose a thermodynamical equilibrium for the system when no drain-source voltage is applied, that is

$$f_p^{eq}(x, k) = M(k)\rho^{eq}(x).$$

In order to give the initial occupation numbers, we have to find the so-called Fermi levels ϵ_F^{eq}. If we have the same configuration at the source and the drain and no voltage is applied then the Fermi levels are constant in the device. Initial occupation numbers are assumed to be given by Boltzmann statistics

$$\rho_p^{eq} = \frac{m_* k_B T_L}{\hbar^2} e^{\frac{\epsilon_F^{eq} - \epsilon_p^{eq}}{k_B T_L}}$$

where the constraint of electrical neutrality $\int N^{eq}(x,z)dz = \int N_D(x,z)dz$ at contacts $x = 0$ and $x = L$ has to be imposed, which provides the final expression of the Fermi levels

$$\epsilon_F^{eq} = (k_B T_L) \log \left[\frac{\hbar^2}{m_* k_B T_L} \frac{\int_0^{l_z} N_D dz}{\sum_q e^{-\frac{\epsilon_q^{eq}}{k_B T_L}}} \right]. \tag{5.5}$$

In order to compute the energy levels ϵ_p^{eq} we still need to compute the set of solutions

$$\left(V^{eq}, \{\epsilon_p[V^{eq}], \chi_p[V^{eq}]\}_p \right)$$

of the Schrödinger-Poisson system (5.2)-(5.3) (at thermal equilibrium), where the density reads

$$N^{eq}(x,z) = \frac{\int N_D(0,z)dz}{\sum_q e^{\frac{\epsilon_q[V_b](0)}{k_B T_L}}} \sum_p e^{\frac{\epsilon_p[V^{eq}](x)}{k_B T_L}} |\chi_p[V^{eq}](x,z)|^2$$

and V_b is the solution of the Schrödinger equation (5.2) coupled with a 1D Poisson problem at the contact for computing the potential and the eigenproperties which respect the electrical neutrality

$$-\frac{d}{dz}\left[\varepsilon_R(0,z)\frac{dV_b}{dz}\right] = -\frac{q}{\varepsilon_0}\left[N[V_b] - N_D(0,z)\right]$$
Homogeneous Neumann at $z = 0$ and $z = l_z$

$$N[V_b] = \frac{\int N_D(0,z)dz}{\sum_q e^{-\frac{\epsilon_q[V_b]}{k_B T_L}}} \sum_p e^{-\frac{\epsilon_p[V_b]}{k_B T_L}} |\chi_p[V_b]|^2.$$

Boundary conditions for the BTE

We need to properly set up border conditions for the incoming particles at the contacts of the device since no flux boundary condition is imposed on the rest of the boundary. The proceeding is almost the same as for setting the initial condition, the only difference being that we look at the situation on either sides of the device. We force the system to fulfill electrical neutrality at contact by pushing the border carrier populations to stay close to f_p^{eq}, the equilibrium distribution: therefore we impose for *entering particles*

$$f_p^n(x \in \{0, L\}, k) = \frac{\rho_p^{eq}(x)}{\rho_p^n(x)} f_p^n(x,k)$$

so that we obtain

$$\int_{\mathbb{R}^2} f_p^n(x \in \{0, L\}, k)dk = \rho_p^{eq}(x),$$

while the *outgoing particles* are determined by the system itself, so we simply impose homogeneous Neumann boundary conditions.

The k_1-dimension is taken large enough in order to make $f_p^n(x, k_1, k_2)$ vanish at the borders, so that any boundary conditions would fit, because no population can be found there. Therefore we impose homogeneous Neumann conditions.

Refer to Figure 3 for an overview.

Boundary conditions for the Schrödinger-Poisson block

As for the diagonalization (5.2), we impose the wave function to be in $H_o^1(0, l_z)$.

We impose on the Poisson equation the typical Neumann boundary conditions on the whole boundary of the device except at the gates, in which Dirichlet boundary conditions are assumed, and at contacts, where the physically meaningful conditions are Dirichlet conditions

$$\begin{aligned} V(t,0) &= V_b(z) \\ V(t,L) &= V_b(z) + V_{DS}(t). \end{aligned}$$

Here $V_{DS}(t)$ is the applied drain-source potential at time t: the potential cannot be applied all of a sudden, because it would be physically meaningless and it would produce numerical strong instabilities.

Ideally we want to have both Dirichlet (for the applied drain-source drop) and Neumann (for the electrical neutrality) conditions at contacts, therefore we impose Robin conditions, which mix either constraints:

$$\frac{dV}{dn} + \alpha\,(V - V_b) = 0.$$

For $\alpha = 0$ we have homogeneous Neumann, while for $\alpha \to +\infty$ we recover Dirichlet. By a proper choice of this parameter we achieve acceptable potential drop and electrical neutrality. Conditions are resumed in Figure 3.

Remark 5.1.1 (The three valleys question)**.** *There are six possible configuration for the particles in the Si. This phenomenon is strictly related with the quantum description of the Si atom. Consider the Block wave $E_p(k)$ and develop it in series*

$$E_p(k) = E_p(0) + E_p'(0)k + E_p''(0)\frac{k^2}{2} + ...$$

If we are developing near a minimum,

$$E_p(k) \approx E_p(0) + E_p''(0)\frac{k^2}{2}.$$

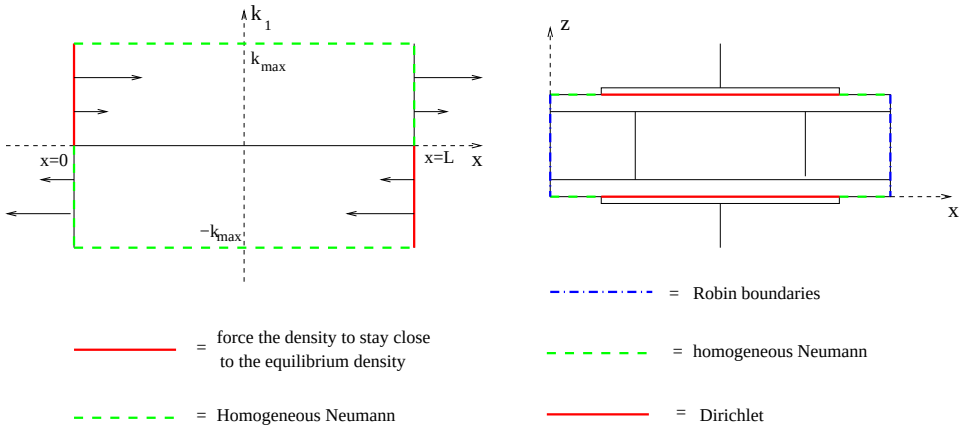


Figure 3: Boundary conditions for the Boltzmann Transport Equation and the Poisson-like equations.

The Hessian matrix $E_p''(0)$ (also called "effective mass" in physics literature) is symmetric, so in fact, three degrees of freedom are allowed. This is where the six configurations (three multiplied by two for symmetry arguments) come from. For each band, six configurations must be taken into account and treated separately.

$$\rho_p^n = \sum_{q=1}^{6} \int_{\mathbb{R}^2} f_{p,q}^n(x,k)dk.$$

This is a further improvement in the modelling of Si-based devices, but for the scope of this work we shall focus on the single-valley case.

5.2 Numerical schemes

Let us first reduce the complete system to dimensionless cartesian coordinates. We assume invariance along the y-dimension, therefore the device spans over the (x,z)-plane. We use the following adimensionalization:

adim.	parameter	value
$\tilde{x} = l^* x, \tilde{z} = l^* z$	$l^* = L_x$	$20 \times 10^{-9}\ m$
$\tilde{k} = k^* k$	$k^* = \frac{\sqrt{2m_* k_B T_L}}{\hbar}$	5.824664×10^{8}
$\tilde{t} = t^* t$	$t^* =$ "typical time"	$10^{-14}\ s$
$\tilde{V} = V^* V$	$V^* =$ typical Vbias	$1\ V$
$\tilde{\epsilon} = \epsilon^* \epsilon$	$\epsilon^* = \frac{\hbar^2 k^{*2}}{2m_*}$	4.141951×10^{-21}
$\tilde{\rho} = \rho^* \rho$	$\rho^* = k^{*2}$	3.392672×10^{17}
$\lvert\tilde{\chi}\rvert^2 = \chi^* \lvert\chi\rvert^2$	$\chi^* = \frac{1}{l^*}$	2.000000×10^{7}
$\tilde{N} = N^* N$	$N^* = \rho^* \chi^*$	6.785343×10^{24}

(5.6)

The BTE reduces to

$$\frac{\partial f_p}{\partial t} + 2C^V k_1 \frac{f_p}{\partial x} - C^V \frac{\epsilon_p^{pot}}{\partial x} \frac{\partial f_p}{\partial k_1} = \frac{1}{\tau} \left[M\rho_p - f_p \right], \tag{5.7}$$

where the dimensionless parameter is

$$C^V = \frac{t^* \epsilon^*}{\hbar l^* k^*},$$

τ is now the dimensionless relaxation time $\tau = \frac{\tilde{\tau}}{t^*}$ and the adimensionalized Maxwellian reads

$$M(k) = \frac{1}{\pi} e^{-|k|^2}.$$

The Schrödinger-Poisson block reads now

$$-C^{S,1} \frac{d}{dz} \left[\frac{1}{m_*(z)} \frac{d\chi_p}{dz} \right] - C^{S,2} \left[V + V_c \right] \chi_p = \epsilon_p \chi_p \tag{5.8}$$

$$-\mathrm{div}_{x,z} \left[\varepsilon_R(x,z) \nabla_{x,z} V \right] = -C^P \left[N - N_D \right], \tag{5.9}$$

where the dimensionless parameters are

$$C^{S,1} = \frac{\hbar^2}{\epsilon^* l^{*2} m_*}, \qquad C^{S,2} = \frac{qV^*}{\epsilon^*}, \qquad C^P = \frac{eN^* l^{*2}}{V^* \varepsilon_0}.$$

Numerical values for all the involved dimensionless parameters, derived from the physical constants, the problem data and the rescaling are given in the following table:

dimensionless constant	value
$C^V = \frac{\epsilon^* t^*}{\hbar k^* l^*}$	1.348615×10^{-2}
$\alpha = \frac{\hbar\omega}{k_B T_L}$	2.436946
$C^{S,1} = \frac{\hbar^2}{\epsilon^* l^{*2} m_*}$	2.358024×10^{-3}
$C^{S,2} = \frac{qV^*}{\epsilon^*}$	3.868169×10^{1}
$C^P = \frac{eN^* l^{*2}}{V^* \varepsilon_0}$	1.534770×10^{-5}

5.2.1 Discretization

The computational domain is discretized into a tensor product mesh, and a uniform mesh is taken in each direction:

$$\begin{aligned} x_i &= i\Delta x \\ (k_1)_l &= -\epsilon_{kin}^{-1}(\alpha \bar{N}) + l\Delta k_1 \\ (k_2)_m &= -\epsilon_{kin}^{-1}(\alpha \bar{N}) + m\Delta k_2, \end{aligned}$$

where α is the dimensionless energy $\alpha = \frac{\hbar\omega}{k_B T_L}$ and ω is the phonon frequency.

We essentially need two main building blocks for the numerical algorithm: one for transport along the x-dimension and one for the computation of the self-consistent potential and all the eigenproperties (band potential energies, Schrödinger eigenfunctions and the electron density itself) through the Schrödinger-Poisson equation, whose solver is the object of the next section.

5.2.2 Newton schemes for the SP block

The Schrödinger-Poisson block (5.8)-(5.9) cannot be decoupled: the potential is needed for the computation on the energy levels (Schrödinger eigenvalues) and Schrödinger eigenfunctions, and the density (which is an eigenproperty itself, requiring the energy levels for its computation) is needed for the self-consistent computation of the electric potential (through Poisson equation).

In [101] the author implemented a Gummel iteration to deal with this block; we have preferred the use of a Newton schemes, even if it requires the evaluation of some Gâteaux derivatives which are not straightforward. We try to minimize the following functional:

$$P[V] = -\mathrm{div}\left[\varepsilon_R \nabla V\right] + \frac{q}{\varepsilon_0}\left[N[V] - N_D\right] \tag{5.10}$$

whose minimum is, in this case, a zero. The Newton iteration reads

$$-dP(V^{old}, V^{new} - V^{old}) = -P[V^{old}], \tag{5.11}$$

where $dP(V,U)$ denotes the Gâteaux derivative of functional P, at point V in direction U.

We remind that the Gâteaux derivative has the following definition: given $F : X \longrightarrow Y$, being X and Y Banach spaces (in fact, only locally convex topological vector spaces is required), $U \subseteq X$ an open set, the Gâteaux derivative of F at point $u \in U$ in direction $\psi \in X$ is

$$dF(u,\psi) = \lim_{\varepsilon \to 0} \frac{F(u+\varepsilon\psi) - F(u)}{\varepsilon}.$$

In order to be able to compute the Gâteaux derivative of the functional (5.10), we have to know how to differentiate eigenvalues and eigenvectors in the generalized eigenvalue problem, whose details were given in Chapter 1, Section 1.8. We recall formulae (1.36) and (1.36):

$$\begin{aligned} d\epsilon_p(V,U) &= -q\langle U\chi_p[V], \chi_p[V]\rangle \\ d\chi_p(V,U) &= -q\sum_{p'} \frac{\langle U\chi_p[V], \chi_{p'}[V]\rangle}{\epsilon_p[V] - \epsilon_{p'}[V]}, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ just means integration. The Gâteaux derivative of (5.10) is therefore calculated:

$$dP(V,U) = -\mathrm{div}\,[\varepsilon_R \nabla U] + \frac{q}{\varepsilon_0} dN(V,U),$$

from which we can see that the central point is how the density is defined. All the different Schrödinger-Poisson problems come from the various definitions of the meaningful densities, whose derivative has, at any rate, form

$$dN(V,U) = \int \mathcal{A}[V](\zeta) U(\zeta) d\zeta,$$

where the kernel $\mathcal{A}[V]$ may have one or more contribution, each one needing to be positive definite.

Scheme (5.11) gives rise to a Poisson-like equation in which an extra term appears:

$$\begin{aligned} & -\mathrm{div}\,[\varepsilon_R \nabla V^{new}] + \int \mathcal{A}[V](\zeta) V^{new}(\zeta) d\zeta \\ = \quad & -\frac{q}{\varepsilon_0} \left[N[V^{old}] - N_D \right] + \int \mathcal{A}[V](\zeta) V^{old}(\zeta) d\zeta. \end{aligned}$$

Numerical solution. The divergence and gradient operators are approximated by alternate finite differences, in order to recover the classical three-points centered scheme for the Laplacian, and the integrals are computed through trapezoids approximation. Once the equation has been descretized, it is solved by means of a Lapack routine called DGESV, which has proven to be robust and fast. For more detail refer to Section 1.6 for the 1D case and Section 1.7 for the 2D case.

Three are the problems which we need to solve by Newton schemes (each one corresponding to a different definition of the density N and by consequence of the kernel $\mathcal{A}[V]$). For each problem we have to choose different boundary conditions.

Problem 1: boundary potential.

This is a 1D problem; we need to compute at $x = 0$ and $x = L$ the potential V_b and the density respecting the electrical neutrality condition $\int_{z=0}^{z=l_z} N(x,\zeta) d\zeta = \int_{z=0}^{z=l_z} N_D(x,\zeta) d\zeta$. The density is defined (written with physical dimensions)

$$N[V_b] = \frac{\int_{z=0}^{z=l_z} N_D(0,\zeta) d\zeta}{\sum_q e^{-\frac{\epsilon_q[V_b]}{k_B T_L}}} \sum_p e^{-\frac{\epsilon_p[V_b]}{k_B T_L}} |\chi_p[V_b]|^2,$$

and homogeneous Neumann boundary conditions are taken at $z = 0$ and $z = l_z$.

Problem 2: thermodynamical equilibrium.

Once we have computed the boundary potential, we must compute the thermodynamical equilibrium for the system when no drain-source voltage is applied:

$$N[V_{eq}] = \frac{\int_{z=0}^{z=l_z} N_D(0,\zeta)d\zeta}{\sum_q e^{-\frac{\epsilon_q[V_b]}{k_B T_L}}} \sum_p e^{-\frac{\epsilon_p[V_{eq}]}{k_B T_L}} |\chi_p[V_{eq}]|^2,$$

where we impose Dirichlet conditions at gates, Dirichlet ($V_{eq} = V_b$)) at contacts (but if homogeneous Neumann or Robin were imposed there would be no difference, as shown by numerical tests), and Homogeneous Neumann elsewhere.

Problem 3: potential.

While making the code progress in time, we need to be able to update the potential during all the transition states. Two versions are possible: an explicit computation and a semi-implicit scheme, which shows less numerical stability and much longer convergence times (even if it should push the eigenproperties towards the equilibrium).

$$N[V] = \begin{cases} \sum_p \rho_p^{n+1} |\chi_p[V]|^2 & \text{explicit} \\ \sum_p \rho_p^{n+1} e^{\frac{\epsilon_p^n}{k_B T_L}} |\chi_p[V]|^2 e^{-\frac{\epsilon_p[V]}{k_B T_L}} & \text{semi-implicit} \end{cases}.$$

As for the boundary conditions, at contacts we would like to have both the electrical neutrality (i.e. homogeneous Neumann conditions) and the correct potential drop which we impose (i.e. Dirichlet conditions); we accomodate it by the use of Robin conditions

$$\frac{dV}{dn} + \alpha\,(V - V_b) = 0,$$

so that we can mix them: we do not achieve the desired potential drop but the electrical neutrality is satisfied enough. By imposing Dirichlet conditions, we loose too much of the electrical neutrality. To give an idea of how these schemes work, we give the Gâteaux derivative of the potential relative to this problem (in the explicit case):

$$\begin{aligned} dN(V,U) &= \sum_p \rho_p 2d\chi_p(V,U)\chi_p[V] \\ &= \int \mathcal{A}[V](z,\zeta)U(\zeta)d\zeta \end{aligned}$$

once we have defined the kernel $\mathcal{A}[V]$ as

$$\begin{aligned}\mathcal{A}[V](z,\zeta) &= -2q\sum_{p,p'}\frac{\rho_p}{\epsilon_p[V]-\epsilon_{p'}[V]}\chi_p[V](\zeta)\chi_{p'}[V](\zeta)\chi_p[V](z)\chi_{p'}[V](z)d\zeta \\ &= q\sum_{p,p'}\frac{\rho_{p'}-\rho_p}{\epsilon_p[V]-\epsilon_{p'}[V]}\chi_p[V](\zeta)\chi_{p'}[V](\zeta)\chi_p[V](z)\chi_{p'}[V](z)d\zeta.\end{aligned}$$

The positive definiteness can be easily checked noticing that

$$\frac{\rho_{p'}-\rho_p}{\epsilon_p[V]-\epsilon_{p'}[V]}\geq 0,$$

because the energies are taken in increasing order, which makes the occupation factors be decreasing (physically it is obvious that the lowest energy levels are the most occupied).

5.2.3 Numerical schemes for the direction of transport

We have several possibilities as for the integration of the BTE (5.7): first of all we have the choice of integrating the original pdf $f_p(t,x,k)$ or defining a slotboom variable $g_p(t,x,k)$ by

$$f_p(t,x,k)=g_p(t,x,k)e^{-\epsilon_p(t,x)-|k|^2},$$

and integrating this last one instead. As for the time discretization, it can be achieved by a third order TVD (Total Variation Diminishing) Runge-Kutta scheme [28], or by splitting techniques [35]. The advantage of integration the slotboom variable is that if we are in an equilibrium state, then $g_p(t,x,k)$ is constant and the equilibrium is numerically well preserved. The BTE in the slotboom variable has one additional term:

$$\frac{\partial g_p}{\partial t}+2C^V k_1\frac{g_p}{\partial x}-C^V\frac{\epsilon_p^{pot}}{\partial x}\frac{\partial g_p}{\partial k_1}-\frac{\partial\epsilon_p^{pot}}{\partial t}g_p=\frac{1}{\tau}\left[\frac{1}{\pi}\int_{\mathbb{R}^2}g_p(k')e^{-|k'|^2}-g_p(k)\right].$$

We resume now the available schemes for the solution of the transport.

WENO second order time splitting schemes in the original variable

The approximation denoted by $f^n_{p,i,l,m}$ to the point values of the solution $f_p(t^n,x_i,(k_1)_l,(k_2)_m)$ is made evolve in time through the second order time splitting scheme subdividing the BTE (5.7):

- **Step 1.** Solve

$$\frac{\partial f_p^n}{\partial t}+2C^V k_1\frac{f_p^n}{\partial x}-C^V\left(\frac{\partial\epsilon_p^{pot}}{\partial x}\right)^n\frac{\partial f_p^n}{\partial k_1}=0$$

  for a $\frac{\Delta t}{2}$-time step; call $f_p^{n+1/3}$ the solution.

- **Step 2.** Solve

$$\frac{\partial f_p^{n+1/3}}{\partial t} = \frac{1}{\tau}\left[M\rho_p^{n+1/3} - f_p^{n+1/3}\right]$$

for a Δt-time step; call $f_p^{n+2/3}(x,k)$ the solution.

- **Step 3.** Solve

$$\frac{\partial f_p^{n+2/3}}{\partial t} + 2C^V k_1 \frac{f_p^{n+2/3}}{\partial x} - C^V\left(\frac{\partial \epsilon_p^{pot}}{\partial x}\right)^{n+2/3}\frac{\partial f_p^{n+2/3}}{\partial k_1} = 0$$

for a $\frac{\Delta t}{2}$-time step; its solution is the desired $f^{n+1}(x,k)$ pdf.

The same procedure is recursively adopted inside the transport part to subdivide this block into the solution of one-dimensional advection problems. Refer to Section 1.2 for details about splitting schemes.

Numerical Scheme: 1D Advection Step.- Each transport block is solved by the Flux Balance Method [46, 26]: when solving the x-transport, k_1 and k_2 act as parameters, as well as x and k_2 when solving the k_1-transport. This method is based on the semi-lagrangian approach of following the characteristics backwards; the improvement is that we force the mass conservation, unlike the direct method, which gives no guarantee about this point. The solution of the x-transport gives

$$\begin{aligned} f^{**}_{p,i,l,m} &= f^{*}_{p,i,l,m} + \frac{1}{\Delta x}\left\{\left[F(x_{i-1/2}) - F(x_{i-1/2} - 2C^V k_1 \Delta t)\right]\right. \\ &- \left.\left[F(x_{i+1/2}) - F(x_{i+1/2} - 2C^V k_1 \Delta t)\right]\right\} \\ F(x) &= \int_0^x f_p^*\left[\xi, (k_1)_l, (k_2)_m\right] d\xi \end{aligned}$$

and, as for the solution of the k_1-transport,

$$\begin{aligned} f^{**}_{p,i,l,m} &= f^{*}_{p,i,l,m} \\ &+ \frac{1}{\Delta k_1}\left\{\left[F((k_1)_{j-1/2}) - F\left((k_1)_{j-1/2} + C^V\frac{\partial \epsilon_p^{pot}}{\partial x}(x_i)\Delta t\right)\right]\right. \\ &- \left.\left[F((k_1)_{i+1/2}) - F\left((k_1)_{i+1/2} + C^V\frac{\partial \epsilon_p^{pot}}{\partial x}(x_i)\Delta t\right)\right]\right\} \\ F(k_1) &= \int_0^{k_1} f_p^*\left[x_i, \xi, (k_2)_m\right] d\xi. \end{aligned}$$

More details about the FBM method can be found in [46, 26]. In order to compute the fluxes, for instance,

$$F(x_{i+1/2}) - F(x_{i+1/2} - 2C^V k_1)$$

we reconstruct the values $F(x_{i+1/2} - 2C^V k_1)$, given the known values of the primitive at the grid points $F(x_{i+1/2})$, by the fifth order Pointwise WENO-6,4 interpolation summarized in next subsection.

Solution via Finite Differences in the original variable.

The approximations to the point values of the solution $f^n_{p,i,l,m}$ are obtained with a dimension-by-dimension (not dimension splitting, in this case) approximation to the spatial derivatives using fifth order WENO method in [64]. The BTE (5.7) is rewritten (in conservation form)

$$\frac{\partial f_p}{\partial t}(t) = L(f,t) \approx -\frac{\partial}{\partial x}(a_1(k) f_p) + \frac{\partial}{\partial k_1}(a_2(t,x) f_{k1}) + \mathcal{Q}_p[f_p] \qquad (5.12)$$

with

$$a_1(k) = 2C^V k_1, \qquad a_2(t,x) = \frac{\partial \epsilon^{pot}_p}{\partial x}(t,x).$$

When approximating $\frac{\partial}{\partial x}(a_1(k) f_p)$, for instance, the other variables (k_1, k_2) are fixed and the approximation is performed along the x-line:

$$\frac{\partial}{\partial x}\left(a_1((k_1)_l, (k_2)_m) f^n_{p,i,l,m}\right) \approx \frac{1}{\Delta x}(\hat{h}_{i+1/2} - \hat{h}_{i-1/2}),$$

where the numerical flux $\hat{h}_{i+1/2}$ is obtained with the fifth order, once computed the "wind direction", i.e. the sign of the coefficient $a_1((k_1)_l, (k_2)_m)$: as this sign is independent of i, when l and m are fixed, the wind direction is fixed. Suppose that $a_1((k_1)_l, (k_2)_m) > 0$ (otherwise the procedure would just be a mirror symmetry with respect to $i + 1/2$ when computing $\hat{h}_{i+1/2}$. We denote

$$\begin{aligned} &h_i = a_1((k_1)_l, (k_2)_m) f^n_p(x_i, (k_1)_l, (k_2)_m), \\ &i = -N_{gh}, ..., N_x + N_{gh} - 1, \qquad (n,p,l,m) \text{ fixed}, \end{aligned}$$

where N_{gh} is the number of ghost points needed for imposing the boundary conditions (with fifth order WENO for flux reconstruction, $N_{gh} = 3$). The numerical flux is

$$\hat{h}_{i+1/2} = \omega_0 \hat{h}^{(0)}_{i+1/2} + \omega_1 \hat{h}^{(1)}_{i+1/2} + \omega_2 \hat{h}^{(2)}_{i+1/2},$$

where $\hat{h}^{(k)}_{i+1/2}$ are the third order fluxes on three different stencils given by

$$\begin{aligned} \hat{h}^{(0)}_{i+1/2} &= \frac{1}{3}h_{i-2} - \frac{7}{6}h_{i-1} + \frac{11}{6}{}_i \\ \hat{h}^{(1)}_{i+1/2} &= -\frac{1}{6}h_{i-1} + \frac{5}{6}h_i + \frac{1}{3}{}_{i+1} \\ \hat{h}^{(2)}_{i+1/2} &= \frac{1}{3}h_i + \frac{5}{6}h_{i+1} - \frac{1}{6}{}_{i+2}, \end{aligned}$$

the non-linear weights are given by

$$\omega_k = \frac{\tilde{\omega}_k}{\sum_{k'} \tilde{\omega}_{k'}}, \qquad \tilde{\omega}_k = \frac{\gamma_k}{(\varepsilon + \beta_k)^2},$$

the linear weights are

$$\gamma_0 = \frac{1}{10}, \qquad \gamma_1 = \frac{3}{5}, \qquad \gamma_2 = \frac{3}{10},$$

the smoothness indicators are

$$\begin{aligned}
\beta_0 &= \frac{13}{12}\left(h_{i-2} - 2h_{i-1} + h_i\right)^2 + \frac{1}{4}\left(h_{i-2} - 4h_{i-1} + 3h_i\right)^2 \\
\beta_1 &= \frac{13}{12}\left(h_{i-1} - 2h_i + h_{i+1}\right)^2 + \frac{1}{4}\left(h_{i-1} - h_{i+1}\right)^2 \\
\beta_2 &= \frac{13}{12}\left(h_i - 2h_{i+1} + h_{i+2}\right)^2 + \frac{1}{4}\left(3h_i - 4h_{i+1} + h_{i+2}\right)^2,
\end{aligned}$$

and $\varepsilon = 10^{-6}$ is a numerical parameters which avoids the denominator to become zero.

The approximation to $\frac{\partial(a_2 f_p)}{\partial k_1}$ is performed in the same fashion. Again, once (p, n, i, m) are fixed, the wind direction is fixed.

For the time discretization a third order TVD (Total variation Diminishing) RUnge-Kutta method [95]:

$$\begin{aligned}
f_p^{(1)} &= f^n + \Delta t L(f^n, t^n) \\
f_p^{(2)} &= \frac{3}{4} f^n + \frac{1}{4} f_p^{(1)} + \frac{1}{4}\Delta t L(t^n + \Delta t, f^{(1)} + \Delta t) \\
f_p^{n+1} &= \frac{1}{3} f^n + \frac{2}{3} f_p^{(2)} + \frac{2}{3}\Delta t L\left(t^n + \frac{\Delta t}{2}, f^{(2)} + \Delta t\right),
\end{aligned}$$

where $L(t, f)$ is defined at (5.12). The time stepping is restricted by the CFL condition due to the explicit character of the time evolution solver (see [94] for a discussion):

$$\Delta t \leq CFL\left(\frac{\Delta x}{\max|a_1|} + \frac{\Delta k_1}{\max|a_2|}\right).$$

Integration in the slotboom variable. If we integration the BTE in the slotboom variable we can use the previously defined schemes; as for the Finite Differences Runge-Kutta scheme, the adaptation is straighforward. The use of the time splitting scheme is a bit more tricky due to the presence of the extra term $\frac{\partial \epsilon_p^{pot}}{\partial x}$. That is why a first order time splitting is more suitable: it reduces to solving for a Δt-time step every single part of the kinetic equation.

Boundary conditions

We have implemented the following boundary conditions:

- at $x = 0$ and $x = L$ we use the following inflow/outflow condition:

$$f^n_{p,-i,l,m} = \begin{cases} f^n_{p,0,l,m} & k_1 < 0 \\ \frac{\rho_p^{eq}(0)}{\rho_p^n(0)} f^n_{p,0,l,m} & k_1 \geq 0 \end{cases}$$

  and

$$f^n_{p,N_x-1+i,l,m} = \begin{cases} f^n_{p,N_x-1,l,m} & k_1 > 0 \\ \frac{\rho_p^{eq}(L)}{\rho_p^n(L)} f^n_{p,N_x-1,l,m} & k_1 \leq 0 \end{cases}$$

  in order to have the ghost points we need for the PWENO interpolation and to preserve the correct values of the distribution function at the drain and the source of the MOSFET;

- at $k_1 = -\varepsilon_{kin}^{-1}(\alpha\bar{N})$ and $k_1 = \varepsilon_{kin}^{-1}(\alpha\bar{N})$ a Neumann type boundary condition in used:

$$f^n_{p,i,-l,m} = f^n_{p,i,0,m} \qquad \text{and} \qquad f^n_{i,N_{k_1}-1+l,m} = f^n_{i,N_{k_1}-1,m}.$$

  WENO schemes are the object of investigation in section 1.1.

5.3 Numerical experiments

We take into account the single-valley model, i.e. electrons are only allowed to have one configuration, without distinguishing between longitudinal effective mass and transverse effective mass. In [101] the author used a Gummel iteration instead of a Newton iteration in order to solve the Schrödinger-Poisson problems and a stationary-state Drift-Diffusion model for the transport part. Our scheme is made for transient-state computation at a microscopic level, so it cannot compete with a macroscopic steady-state code in terms of times of calculation. We have chosen a very restrictive CFL conditions for the Runge-Kutta discretization, and the potential is applied very softly, due to the preconditional character of Newton schemes used for computing the potential: for the sake of convergence they cannot be initialized too far from the equilibrium, so we cannot impose a shock as potential. In our tests, the potential $V(x, z)$ increases linearly from $0V$ to the desired drop in $1ps$ time.

In the tests which we show here, the CFL condition is set 0.01. It might be improved, but extensive tests on the stability of the scheme with less restrictive conditions have not been performed yet. Anyway, we expect we

cannot make it much larger, because the time stepping should stay below $0.0001ps$ at any rate, even if the CFL condition does not constraint it that much: the self-consistent electrostatic field needs be computed very often, not to introduce instabilities (oscillations of the band-energies and potential energy, in this case) due to growing overcorrections to the electrostatic field computed at the previous stage.

In our test, we have used as effective Silicon mass

$$m_* = 0.5 m_{Si}.$$

Other tests with other effective masses need to be performed, but for the purpose of this work, i.e. setting the basis for a deeply detailed description of the nanoMOSFET in Figure 1, it is already meaningful to observe that the model is properly simulated in this test case.

Our tests have been performed with a structure of six subbands. The more subband we introduce, the more detailed the description is. In [101] the author used twelve subbands; anyway, over the third band the carrier polulation is very low, almost uninfluent.

5.3.1 Border potential

In Figure 4 we show the border potential computed via the scheme in Section 5.2.2. Results coincide with [101], where they have been obtained via Gummel iteration. As a remark, Newton is initialized by a constant null potential and its convergence is very fast, usually achieved in about four steps.

5.3.2 Thermodynamical equilibrium

In Figures 5 and 6, we show the results of thermodynamical equilibrium resulting from the solution of the SP-block discussed in problem 2 above in section 5.2.2. In [101] the author obtained equivalent results by means of a Gummel iteration. In our tests, Newton converges very fast, usually no more than ten steps, depending on the meshing and on the number of subbands taken into account.

5.3.3 Long-time behavior

We have used a $64 \times 32 \times 16 \times 16$-grid in the (x, z, k_1, k_2)-space with a Runge-Kutta-3 solver for the original variable f_p. The CFL condition is set 0.01, a very restrictive value used in order to avoid oscillations due to Poisson overcorrections for it has not been solved frequently enough. We take 1 ps time to apply the potential, because at any step Newton iterations cannot be initialized too far from the equilibrium. As for the potential at contacts, we can use either Robin or Dirichlet conditions; in fact, no difference is osberved: this last one works properly too, the electrical neutrality already

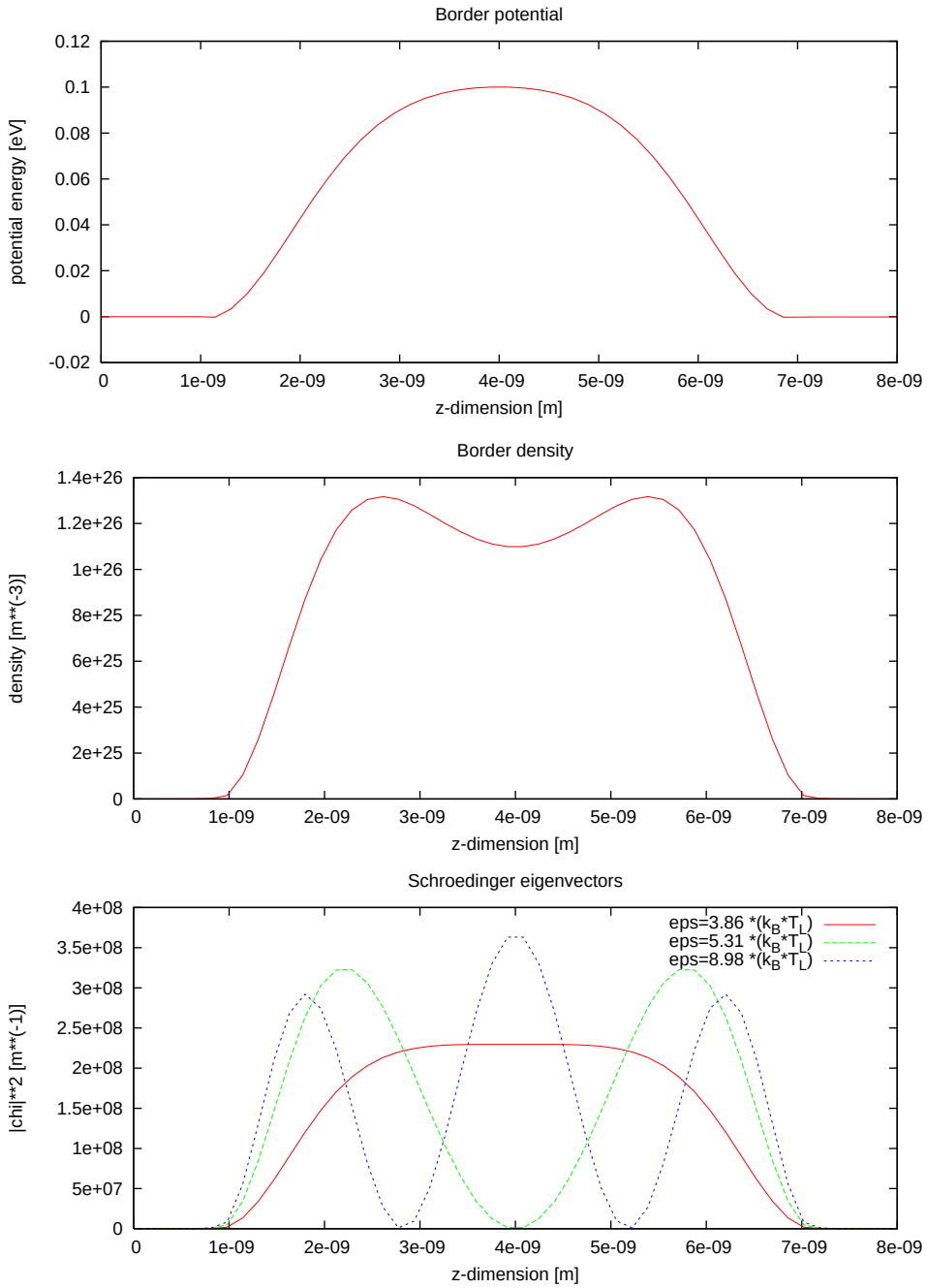


Figure 4: Border potential. The grid is 64 points along the z-dimension. Top: the potential energy at contacts. Center: the free electron density at contacts. Bottom: the Schrödinger eigenfunctions corresponding to the three first energy levels.

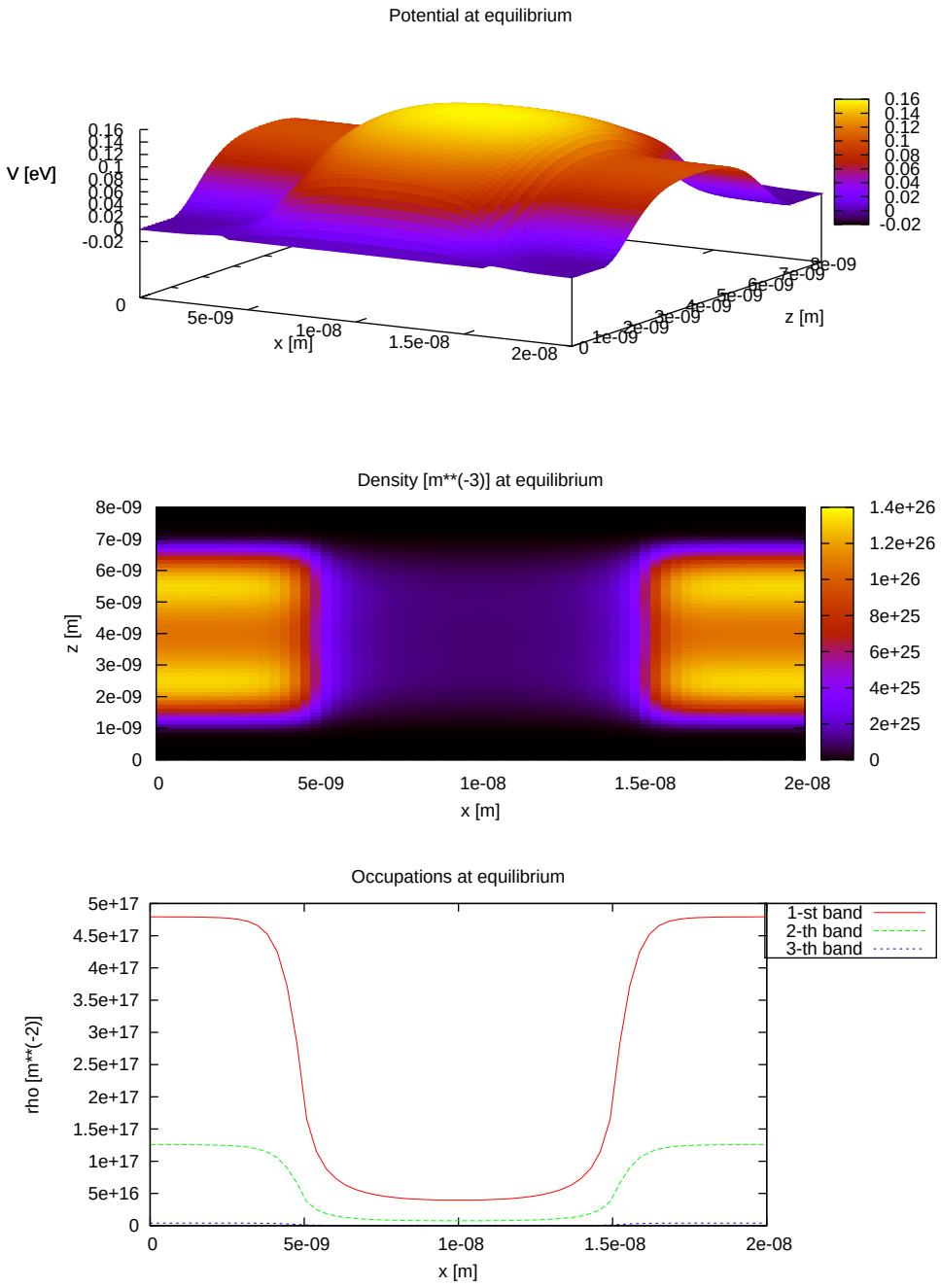


Figure 5: Thermodynamical equilibrium. The grid is 64×64 in the (x, z)-dimensions. Top: the potential energy. Center: the free electron density. Bottom: the occupation factor of the first three energy bands.

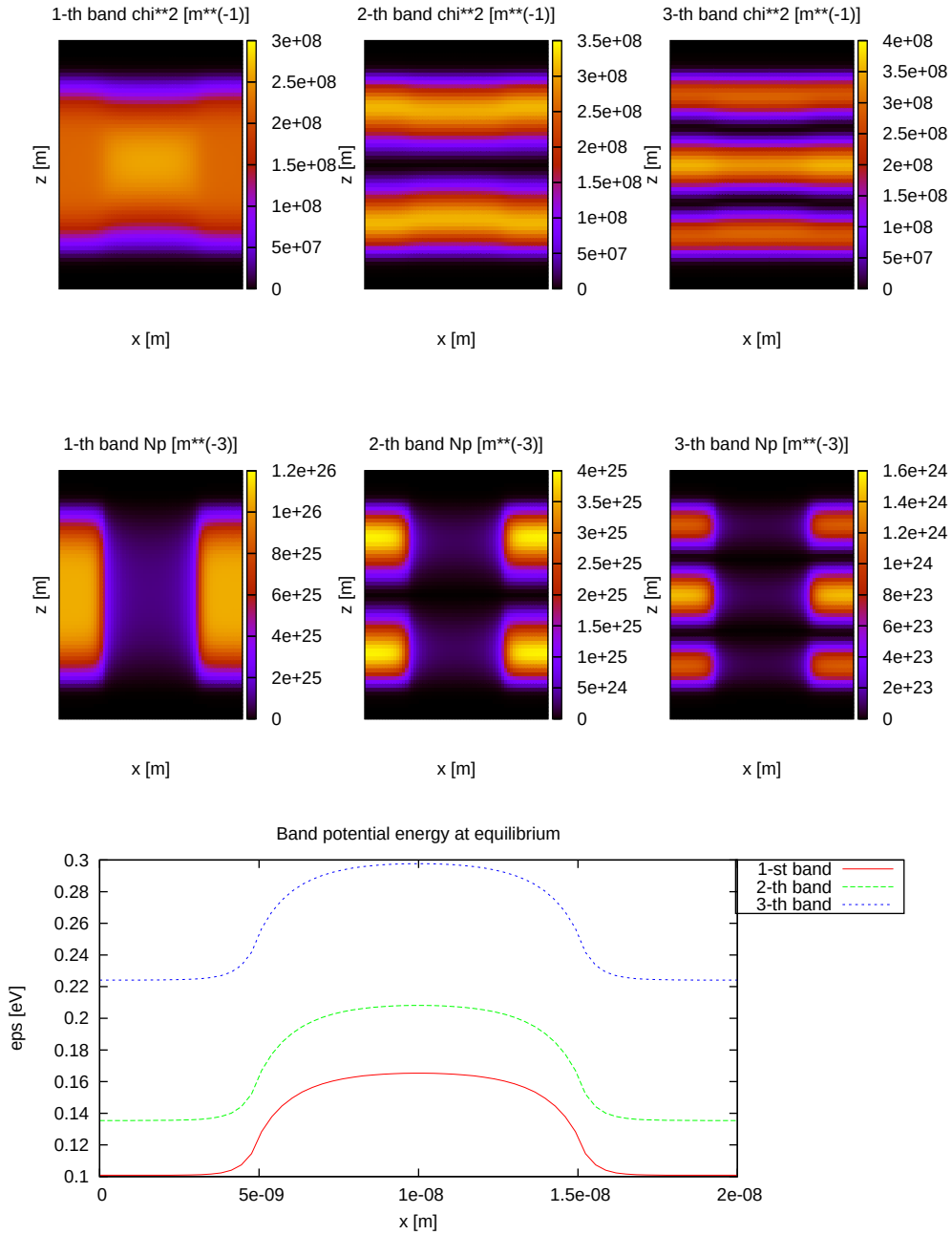


Figure 6: Thermodynamical equilibrium. The grid is 64×64 in the (x, z)-dimensions. Top: the Schrödinger eigenfunctions for the first three bands. Center: the band densities of the free electrons. Bottom: the band potential energies of the first three bands.

being imposed in the transport equation for the entering particles.
We present results concerning the long-time behavior: we expect the macroscopic magnitudes to stabilize after the potential has been completely applied. In Figure 9 and 10 we plot the long-time behavior for a potential drop of $V_{DS} = 0.2V$, and in Figure 7 and 8 the long-time behavior for a potential drop of $V_{DS} = 0.5V$.

Results are satisfactory and close the reference results in [101]. Anyway, this should be the beginning for some improvements needed for a more detailed description of the MOSFET devices: first of all a complete scattering operator, not just a relaxation time operator; then the three valleys case, to distinguish between transversal and longitudinal effective masses; also achieving a 3D geometry would be very useful.

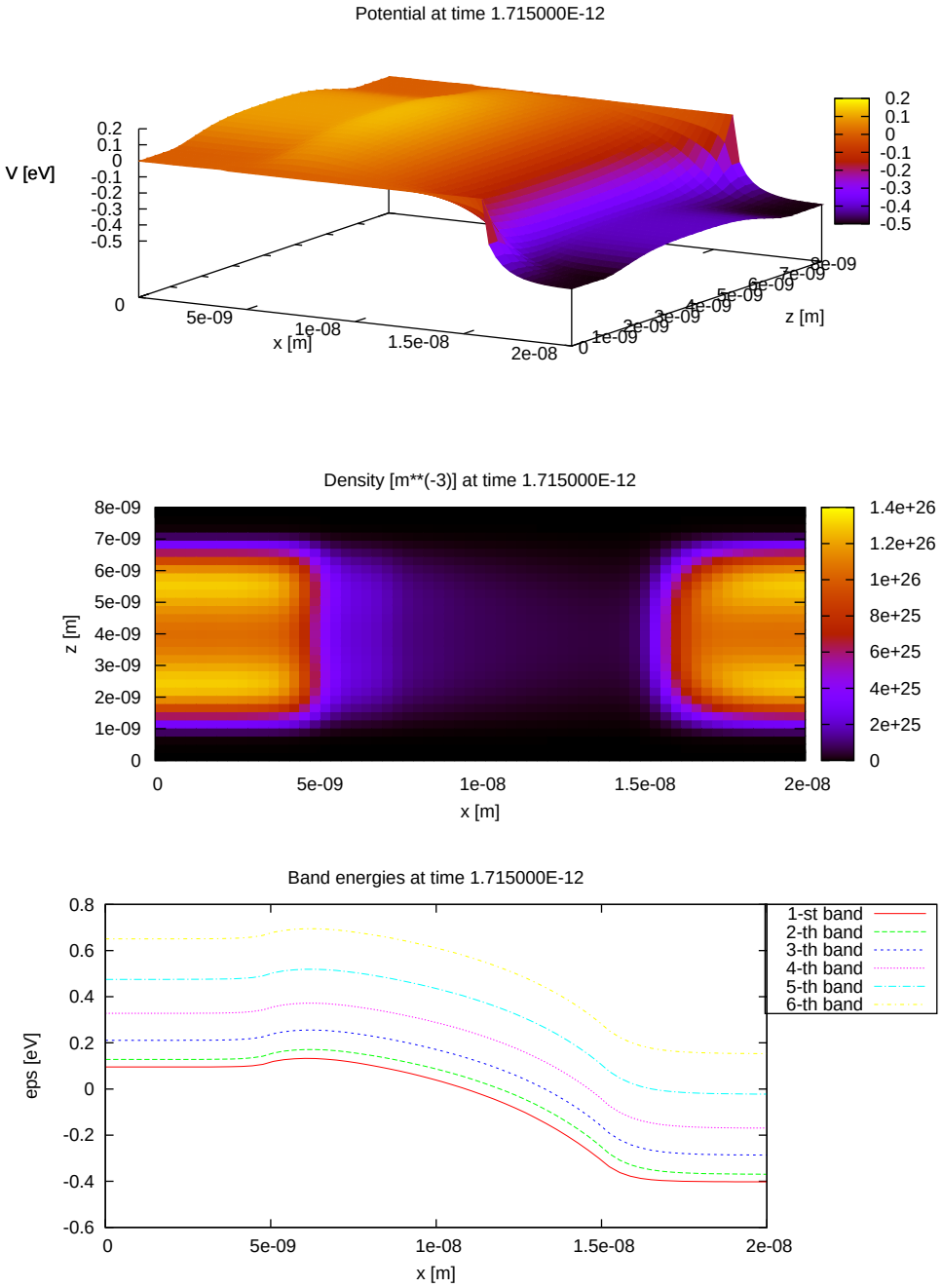


Figure 7: The single-valley case (the effective mass is set $m_* = 0.5$) when the drain-source potential is $V_{DS} = 0.5V$, the mesh is $64 \times 32 \times 16 \times 16$ in the (x, z, k_1, k_2)-space, Poisson boundary conditions are Robin (with $\alpha = 5$), the solver for the BTE is Runge-Kutta-3 (FDWENO-5,3 for interpolation) with $CFL = 0.01$, the potential grows from 0 to V_{DS} in $1ps$-time. Top: the potential energy. Center: the total free electrons density. Bottom: the band potential energies.

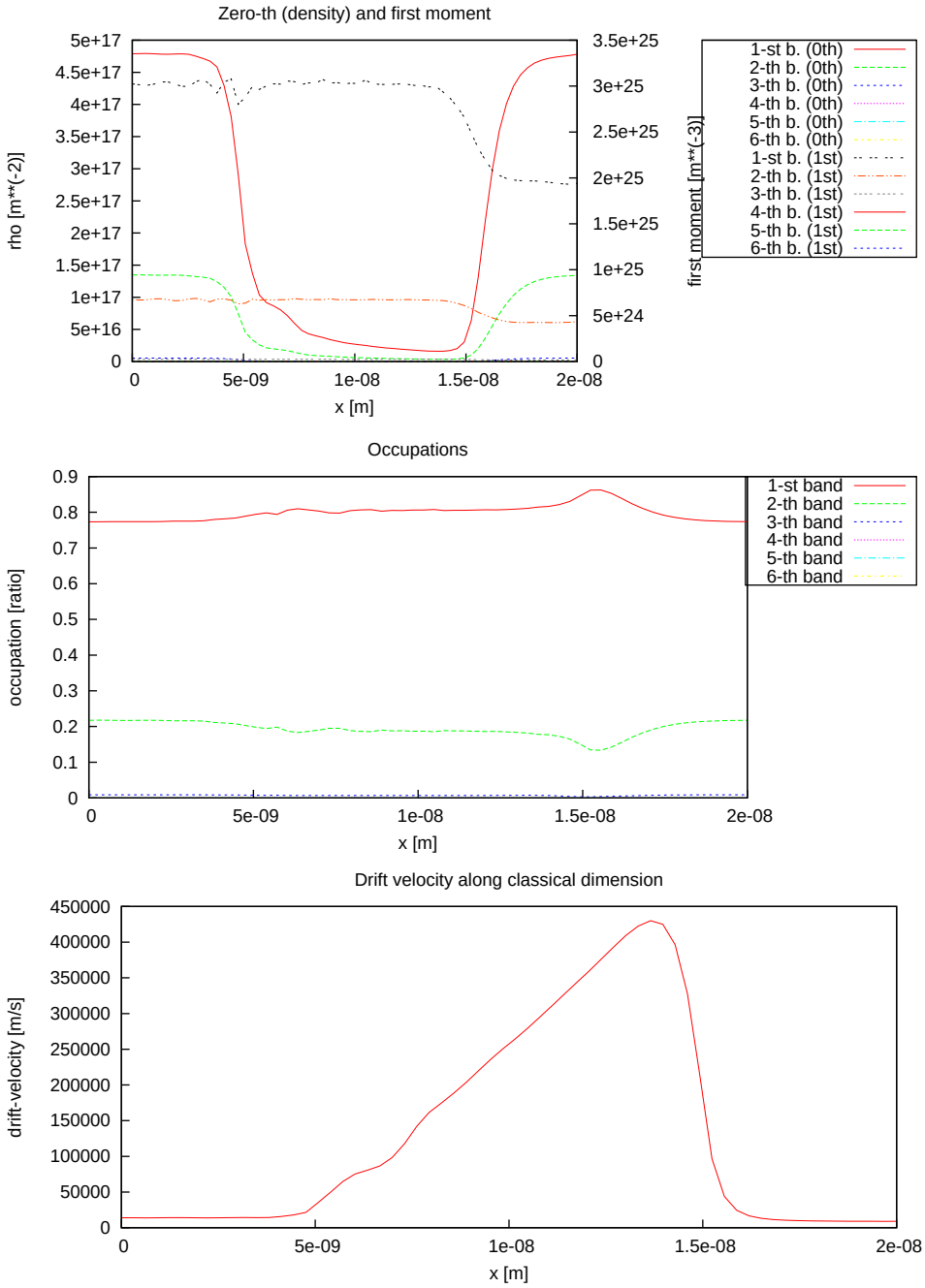


Figure 8: The single-valley case (the effective mass is set $m_* = 0.5$) when the drain-source potential is $V_{DS} = 0.5V$, the mesh is $64 \times 32 \times 16 \times 16$ in the (x, z, k_1, k_2)-space, Poisson boundary conditions are Robin (with $\alpha = 5$), the solver for the BTE is Runge-Kutta-3 (FDWENO-5,3 for interpolation) with $CFL = 0.01$, the potential grows from 0 to V_{DS} in $1ps$-time. Top: the band density and band first moment along the classical dimension. Center: The occupation factors. Bottom: The drift velocity along the classical dimension.

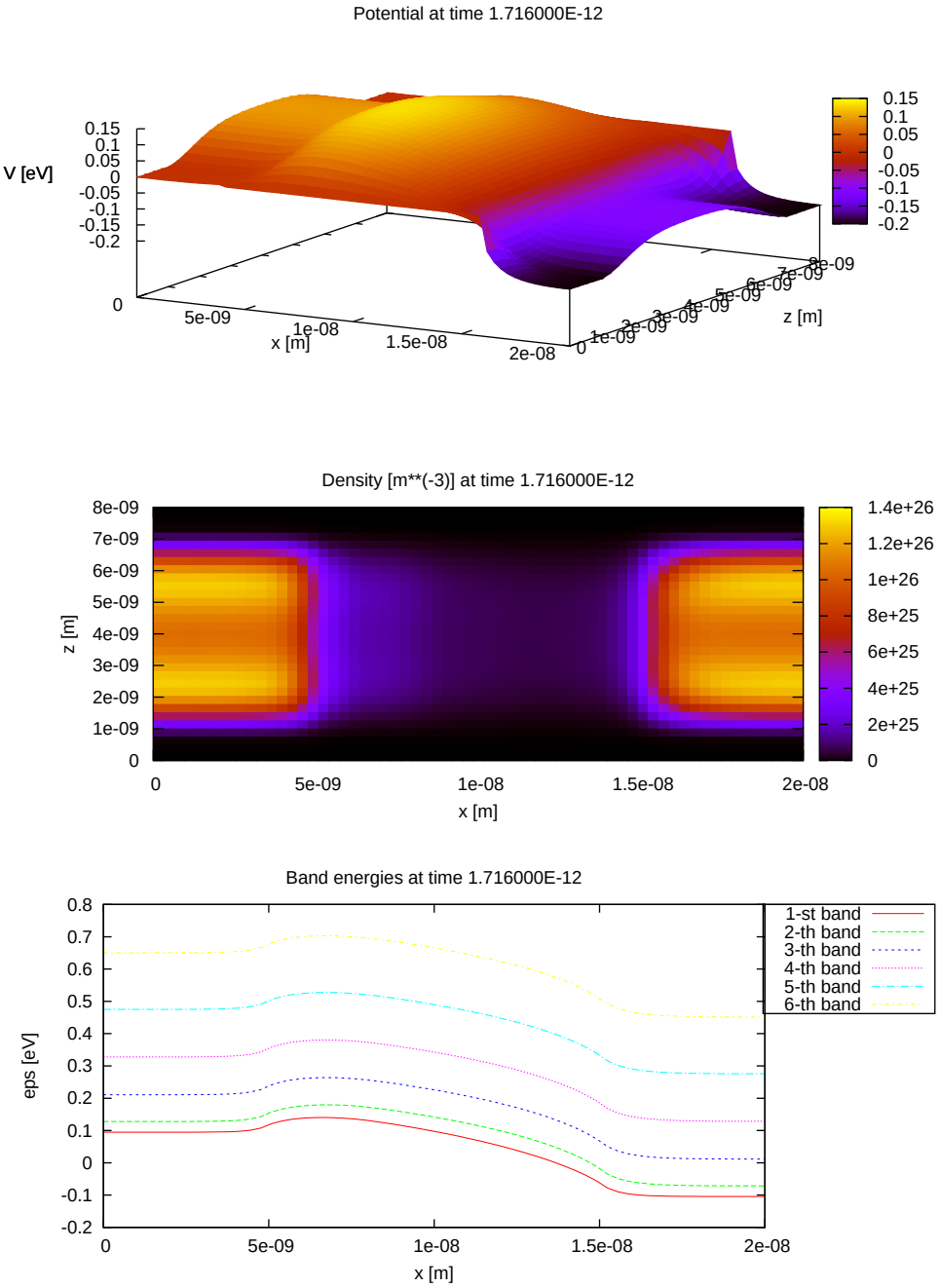


Figure 9: The single-valley case (the effective mass is set $m_* = 0.5$) when the drain-source potential is $V_{DS} = 0.2V$, the mesh is $64 \times 32 \times 16 \times 16$ in the (x, z, k_1, k_2)-space, Poisson boundary conditions are Robin (with $\alpha = 5$), the solver for the BTE is Runge-Kutta-3 (FDWENO-5,3 for interpolation) with $CFL = 0.01$, the potential grows from 0 to V_{DS} in $1ps$-time. Top: the potential energy. Center: the total free electrons density. Bottom: the band potential energies.

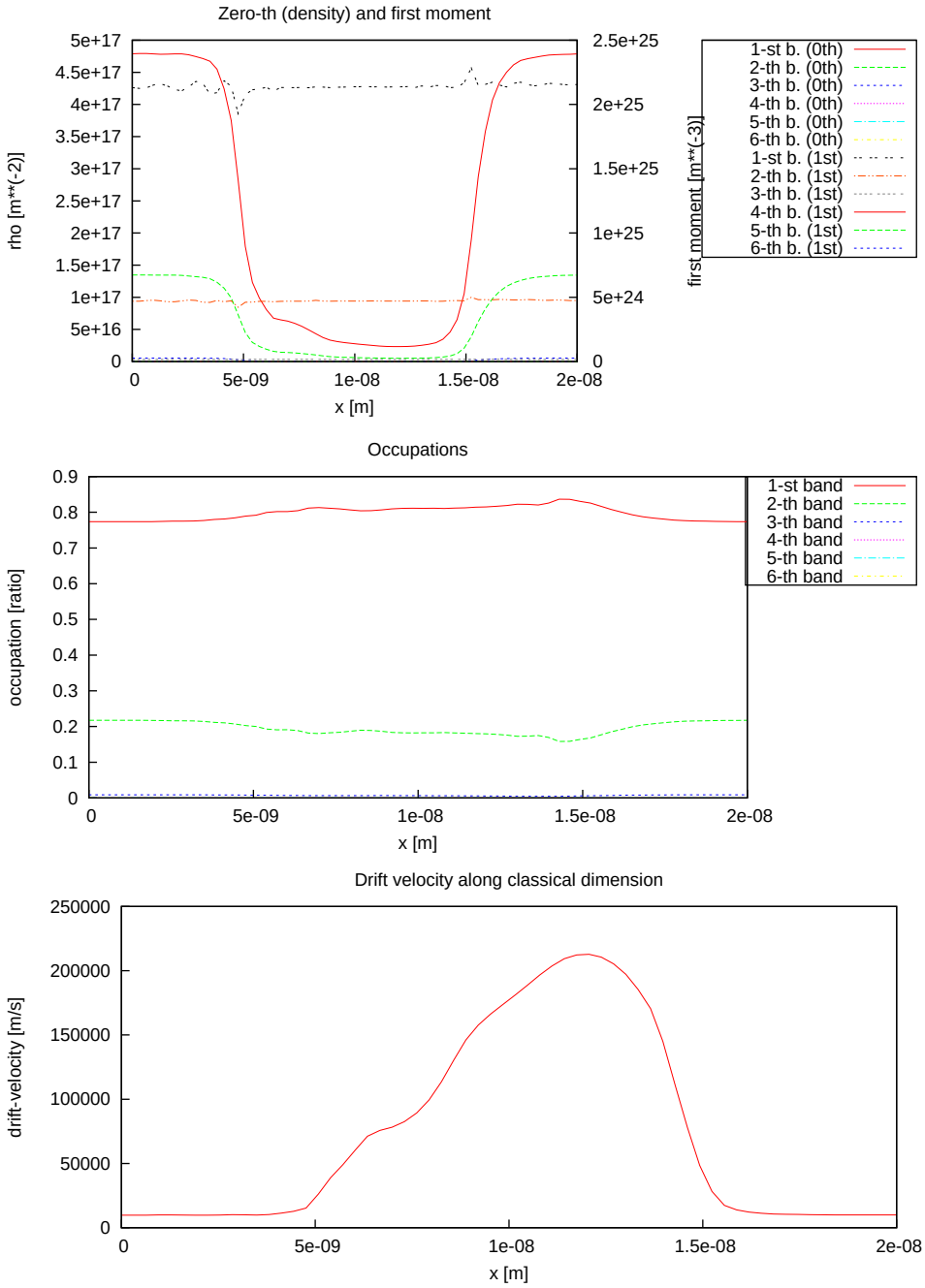


Figure 10: The single-valley case (the effective mass is set $m_* = 0.5$) when the drain-source potential is $V_{DS} = 0.2V$, the mesh is $64 \times 32 \times 16 \times 16$ in the (x, z, k_1, k_2)-space, Poisson boundary conditions are Robin (with $\alpha = 5$), the solver for the BTE is Runge-Kutta-3 (FDWENO-5,3 for interpolation) with $CFL = 0.01$, the potential grows from 0 to V_{DS} in $1ps$-time. Top: the band density and band first moment along the classical dimension. Center: The occupation factors. Bottom: The drift velocity along the classical dimension.

Appendix A

Appendix

A.1 Definition of Gâteaux-derivative

Let X and Y be two locally convex topological spaces (take for instance Banach spaces), and let $U \subset X$ be an open set, and

$$F : X \longrightarrow Y$$

The Gâteaux (directional) derivative of F at point $u \in U$ in the direction ψ is

$$dF(u, \psi) = \lim_{t \to 0} \frac{F(u + t\psi) - F(u)}{t}.$$

Remind that the Gâteaux derivative is unique (if it exists), that it is homogeneous, i.e.

$$\begin{aligned} u \in U, \qquad & dF(u, \cdot) : X \longrightarrow Y \\ & dF(u, \alpha\psi) = \alpha dF(u, \psi) \end{aligned}$$

but it is not always linear.

A.2 The Gauss-Siegel and the Successive OverRelaxation methods for solving the linear system

Suppose we have to solve the linear system

$$AX = b,$$

with

$$A \in \mathcal{M}(\mathbb{R}, N \times N), X \in \mathbb{R}^N, b \in \mathbb{R}^N.$$

The Gauss-Siegel iteration starts from an initial guess, say

$$X^{(0)} = \left(X_0^{(0)}, X_1^{(0)}, ..., X_{N-1}^{(0)} \right).$$

Then we update the i-th entry from the i-th row of the linear system A, starting from the first row:

$$\begin{aligned} X_0^{(1)} &= \frac{b_0 - \sum_{j\neq 0} A_{0,j} X_j^{(0)}}{A_{0,0}} \\ X_1^{(1)} &= \frac{b_1 - A_{1,0} X_0^{(1)} - \sum_{j\geq 20} A_{1,j} X_j^{(0)}}{A_{1,1}} \\ &\vdots \\ X_i^{(1)} &= \frac{b_i - \sum_{j<i} A_{i,j} X_j^{(1)} - \sum_{j>i} A_{i,j} X_j^{(0)}}{A_{i,i}} \\ &\vdots \end{aligned}$$

So, the iteration is: from guess

$$X^{(k)} = \left(X_0^{(k)}, X_1^{(k)}, ..., X_{N-1}^{(k)} \right)$$

perform

$$\text{for } i = 0, ..., N-1 \ , \qquad X_i^{(k+1)} = \frac{b_i - \sum_{j<i} A_{i,j} X_j^{(k+1)} - \sum_{j>i} A_{i,j} X_j^{(k)}}{A_{i,i}}$$

until convergece is fulfilled, i.e.

$$\| X^{(k+1)} - X^{(k)} \| < \lambda_{tolerance},$$

for some tolerance (e.g. $\lambda_{tolerance} = 10^{-6}$) and some appropriate norm (use L^2 or L^∞ for instance).

The SOR method is a generalization of the Gauss-Siegel iteration. Given a parameter $0 < \omega < 2$, the update transforms into

$$X_i^{(k+1)} = (1-\omega) X_i^{(k)} + \omega \frac{b_i - \sum_{j<i} A_{i,j} X_j^{(k+1)} - \sum_{j>i} A_{i,j} X_j^{(k)}}{A_{i,i}}.$$

For $\omega = 1$ we recover the Gauss-Siegel iteration.

Bibliography

[1] Cálculo numérico I. *UNED*, 1989.

[2] N. Ben Abdallah, P. Degond, and S. Génieys. An Energy-Transport model derived from the Boltzmann equation of semiconductors. *J. Stat. Phys.*, 7(37):3306–3333, 1996.

[3] N. Ben Abdallah, F. Méhats, and N. Vauchelet. Analysis of a Drift-Diffusion-Schrödinger-Poisson model. *C.R. Acad. Sci. Paris, Ser I*, (335):1007–1012, 2002.

[4] N. Ben Abdallah, F. Méhats, and N. Vauchelet. A note on the long time behavior for the Drift-Diffusion-Poisson system. *C.R. Acad. Sci. Paris*, 10(339):683–688, 2004.

[5] N. Ben Abdallah, F. Méhats, and N. Vauchelet. Diffusive transport of partially quantized particles: existence, uniqueness and long-time behaviour. *Proc. Edinb. Math. Soc.*, 2(49):513–549, 2006.

[6] R. Abgral and C. Basdevant. A semi-lagrangian numerical scheme for two-dimensional turbulence. *C. R. Acad. Sci. Paris Sér. I Math.*, (305):315–318, 1987.

[7] D. Aregba-Driollet and R. Natalini. Discrete kinetic schemes for multidimensional systems of conservation. *SIAM J. Numer. Anal.*, 6(37):1973–2004 (electronic), 2000.

[8] C. Auer, A. Majorana, and F. Schürrer. Numerical schemes for solving the non-stationary Boltzmann-Poisson system for two-dimensional semiconductor devices. *ESIAM: Proceedings*, (15):75–86, 2005.

[9] C. Bardos, F. Golse, and C.D. Levermore. Fluid dynamiac limits of kinetic equations, II: Convergence proofs for the Boltzmann equation. *Comm. Pure Appl. Math.*, 5(46):667–753, 1993.

[10] C. Bardos, F. Golse, B. Perthame, and R. Sentis. The nonaccretive radiative transfer equations: existence of solutions and Rosseland approximation. *J. Funct. Anal.*, 2(77):434–460, 1988.

[11] C. Bardos, R. Santos, and R. Sentis. Diffusion approximation and computation of the critical size. *Trans. Amer. Math. Soc.*, 2(284):617–649, 1984.

[12] J.R. Bates. Semi-lagrangian advective schemes and their use in meteorological modeling. Large-scale computations in fluid mechanics, part 1 (La Jolla, Calif. 1983). *Lectures in Appl. Math., Amer. Math. Soc.*, 1(22):1–29, 1985.

[13] N. Besse and E. Sonnendrücker. Semi-lagrangian schemes for the Vlasov equation on a unstructured mesh of phase space. *J. Comp. Phys.*, (191):341–376, 2003.

[14] F. Bouchut. Global weak solutions of the Vlasov-Poisson system for small electron mass. *Comm. Partial Diff. Equations*, (16):1337–1365, 1991.

[15] F. Bouchut, F. Golse, and M. Pulvirenti. Kinetic equations and asymptotic theory. *Series in Appl. Math. bf 4, Gauthiers-Villars*, 2000.

[16] J.P. Bourgade. On Sperical Harmonics Expansion type models for electron-phonon collisions. *Math. Methods Appl. Sci.*, (26):247–271, 2003.

[17] Y. Brenier. Systèmes hyperboliques de lois de conservation. *Cours de DEA 92–93, Publ. Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie*, 1992.

[18] C. Buet and S. Cordier. Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models. *C. R. Math. Acad. Sci. Paris*, 12(338):951–956, 2004.

[19] C. Buet and B. Despres. Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics. *JQSRT*, 3-4(85):385–418, 2004.

[20] C. Buet and B. Despres. Asymptotic preserving and positive schemes for radiation hydrodynamics. *J. Comput. Phys.*, 2(215):717–740, 2006.

[21] M.J. Cáceres, J.A. Carrillo, I.M. Gamba, A. Majorana, and C.-W. Shu. Deterministic kinetic solvers for charged particle transport in semiconductor devices. *Cercignani, C., Gabetta, E. (eds.) Transport Phenomena and Kinetic Theory: Applications to Gases, Semiconductors, Photons and Biological Systems, Series: Modelling and Simulation in Science, Engineering and Technology, Birkhäuser*, 2007.

[22] M.J. Cáceres, J.A. Carrillo, and A. Majorana. Deterministic simulation of the Boltzmann-Poisson system in GaAs-based semiconductors. *SIAM J. Sci. Comp.*, (27):1981–2009, 2006.

[23] J. A. Carrillo, T. Goudon, P. Lafitte, and F. Vecil. Numerical schemes of diffusion asymptotics and moment closures for kinetic equations. *to appear in Journal of Scientific Computing*, 2007.

[24] J. A. Carrillo, Th. Goudon, and P. Lafitte. Simulations of two-phase flows involving a dispersed phase: Bubbling and flowing regimes. *preprint*.

[25] J. A. Carrillo, A. Majorana, and F. Vecil. A semi-lagrangian deterministic solver for the semiconductor Boltzmann-Poisson system. *Communications in Computational Physics*, (2):1027–1054, 2007.

[26] J. A. Carrillo and F. Vecil. Non oscillatory interpolation methods applied to Vlasov-based models. *SIAM Journal of Scientific Computing*, (29):1179–1206, 2007.

[27] J.A. Carrillo, M.J. Cáceres, and T. Goudon. Equilibration rate for the linear inhomogeneous relaxation-time Boltzmann equation for charged particles. *Comunications in Partial Differential Equations*, (28):969–989, 2003.

[28] 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. *J. Comput. Phys.*, (184):498–525.

[29] J.A. Carrillo, I.M. Gamba, A. Majorana, and C.-W. Shu. 2D semiconductor device simulations by WENO-Boltzmann schemes: efficiency, boundary conditions and comparison to Monte Carlo methods. *J. Comput. Phys.*, (214):55–80, 2006.

[30] J.A. Carrillo, I.M. Gamba, and C.-W. Shu. Computational macroscopic approximations to the one-dimensional relaxation-time kinetic system for semiconductors. *Phys. D*, (146):289–306, 2000.

[31] J. Cartier and A. Munnier. Geometric Eddington factor for radiative transfer problems. *IRMA Lect. Math. Theor. Phys.*, 7:271–293, 2005.

[32] C. Cercignani, I. Gamba, J. Jerome, and C.-W. Shu. Device benchmark comparisons via kinetic, hydrodynamic, and high-field models. *Comput. Methods Appl. Mech. Engrg.*, (181):381–392, 2000.

[33] C. Cercignani, I. Gamba, J. Jerome, and C.-W. Shu. Device benchmark comparisons via kinetic, hydrodynamic, and high-field models. *Comput. Methods Appl. Mech. Engrg.*, (181):381–392, 2000.

[34] P.H. Chavanis, J. Sommeria, and R. Robert. Statistical mechanics of two-dimensional vortices and collisionless stellar systems. *The Astrophysical J.*, (471):385–399, 1996.

[35] C.Z. Cheng and G. Knorr. The integration of the Vlasov equation in configuration space. *J. Comput. Phys.*, (22):330–351, 1976.

[36] J.-F. Coulombel, F. Golse, and Th. Goudon. Diffusion approximation and entropy-based moment closure for kinetic equations. *Asymptot. Anal.*, 1-2(45):1–39, 2005.

[37] J.-F. Coulombel and Th. Goudon. Entropy-based moment closure for kinetic equations: Riemann problem and invariant regions. *J. Hyperbolic Diff. Eq.*, (3):649–671, 2006.

[38] N. Crouseilles and F. Filbet. Numerical approximation of collisional plasmas by high order methods. *J. Comp. Phys.*, (201):546–572, 2004.

[39] P. Degond. An infinite system of diffusion equations arising in transport theory: the coupled Spherical Harmonic Expansion model. *Math. Models Methods Appl. Sci.*, (11):903–932, 2001.

[40] J. Dolbeault, P.A. Markowich, D. Ölz, and C. Schmeiser. Nonlinear diffusion as limit of kinetic equations with relaxation collision kernels. *preprint*, 2006.

[41] B. Dubroca. Etude de régimes microscopiques, macroscopiques et transitionnels basés sur des équations cinétiques : modélisation et approximation numérique. *Habilitation à diriger les recherches. Université Bordeaux 1*, 2000.

[42] F. Filbet. Thèse doctorale: Constribution à l'analyse et a la simulation numérique de l'equation de Vlasov. *Université Henri Poincaré*, 2001.

[43] F. Filbet, C. Mouhot, and L. Pareschi. Solving the Boltzmann equation in $n \log n$. *SIAM J. Scientific Computing*, (28):1029–1053, 2006.

[44] F. Filbet and G. Russo. Accurate numerical methods for the Boltzmann equation, in modeling and computational methods for kinetic equations. *Model. Simul. Sci. Eng. Technol., Birkhäuser Boston, Boston*, pages 117–145, 2004.

[45] F. Filbet and E. Sonnendrücker. Comparison of eulerian Vlasov solvers. *Comput. Phys. Commun.*, (150):247–266, 2003.

[46] F. Filbet, E. Sonnendrücker, and P. Bertrand. Conservative numerical schemes for the Vlasov equation. *J. Comput. Phys.*, (172):166–187, 2001.

[47] J. Fort. Information-theoretical approach to radiative transfer. *Physica A*, 3-4(243):275–303, 1997.

[48] M. Galler. Multigroup equations for the description of the particle transport in semiconductors. *Series on Advances in Mathematics for Applied Sciences 70, World Scientific Publishing*, 2005.

[49] M. Galler and A. Majorana. Deterministic and stochastic simulations of electron transport in semiconductors. *Bull. Inst. Math. Acad. Sin. (N.S.)*, 2(2):349–365, 2007.

[50] M. Galler and F. Schürrer. A deterministic solver for the transport of the AlGaN/GaN 2D electron gas including hot-phonon and degeneracy effects. *J. Comput. Phys.*, (210):519–534, 2005.

[51] M. Galler and F. Schürrer. A direct MultiGroup-WENO solver for the 2D non-stationary Boltzmann-Poisson system for GaAs devices: GaAs-MESFET. *J. Comput. Phys.*, (212):778–797, 2006.

[52] Y. Giga and T. Miyakawa. A kinetic construction of global solutions of first order quasilinear equations. *Duke Math. J.*, 2(50):505–515, 1983.

[53] P. Godillon-Lafitte and Th. Goudon. A coupled model for radiative transfer: Doppler effects, equilibrium and non equilibrium diffusion asympotics. *SIAM MMS*, (4):1245–1279, 2005.

[54] F. Golse, S. Jin, and C. D. Levermore. A domain decomposition analysis for a two-scale linear transport problem. *M2AN. Mathematical Modelling and Numerical Analysis*, 6(37):869–892, 2003.

[55] F. Golse and L. Saint-Raymond. Hydrodynamics limit for the Boltzmann equation. *Lectures Porto Ercole*, 2004.

[56] L. Gosse and G. Toscani. An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *C. R. Math. Acad. Sci. Paris*, 4(334):337–342, 2002.

[57] L. Gosse and G. Toscani. Asymptotic-preserving & well-balanced schemes for radiative transfer and the Rosseland approximation. *Numer. Math.*, 2(98):223–250, 2004.

[58] Th. Goudon and P. Lafitte. Splitting schemes for the simulation of non equilibrium radiative flows. *preprint*, 2006.

[59] Th. Goudon and A. Mellet. On fluid limit for the semiconductors Boltzmann equation. *J. Differential Equations*, 1(189):17–45, 2003.

[60] Th. Goudon and F. Poupaud. Approximation by homogenization and diffusion of kinetic equations. *Comm. Partial Differential Equations*, 3-4(26):537–569, 2001.

[61] M. Gutnic, M. Haefele, I. Paun, and E. Sonnendrücker. Vlasov simulations on an adaptative phase-space grid. *Comput. Phys. Commun.*, (164):214–219, 2004.

[62] F. Hérau. Hypocoercitivity and exponential time decay for the linear inhomogeneous relaxation Boltzmann equation. *Asymptotic Anal.*, 3-4(46):349–359, 2005.

[63] F. Huot, A. Ghizzo, P. Bertrand, E. Sonnendrücker, and O. Coulaud. Instability of the time-splitting scheme for the one-dimensional and relativistic Vlasov-Maxwell system. *J. Comput. Phys.*, (185):512–531, 2003.

[64] G.S. Jiang and C.-W. Shu. Efficient implementation of Weighted ENO schemes. *J. Comp. Phys.*, (126):202–228, 1996.

[65] S. Jin, L. Pareschi, and G. Toscani. Diffusive relaxation schemes for discrete-velocity kinetic equations. *SIAM J. Numer. Anal.*, (35):2405–2439, 1998.

[66] S. Jin, L. Pareschi, and G. Toscani. Uniformly accurate diffusive relaxation schemes for mustiscale transport equations. *SIAM J. Numer. Anal.*, (38):913–936, 2000.

[67] S. Jin and Z.-P. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm. Pure Appl. Math.*, (48):235–276, 1995.

[68] D. Kincaid and W. Cheney. Análisis numérico. *Addison-Wesley Iberoamericana*, 1994.

[69] A. Klar. An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit. *SIAM J. Numer. Anal.*, 3(35):1073–1094, 1998.

[70] A. Klar. An asymptotic preserving numerical scheme for kinetic equations in the low Mach number limit. *SIAM J. Numer. Anal.*, 5(36):1507–1527, 1999.

[71] A. Klar and A. Unterreiter. Uniform stability of a finite difference scheme for transport equations in diffusive regime. *SIAM J. Numer. Anal.*, 3(40):891–913, 2002.

[72] S. Labrunie, J.A. Carrillo, and P. Bertrand. Numerical study on hydrodynamic and quasi-neutral approximations for collisionless two-species plasmas. *J. Comput. Phys.*, (200):267–298, 2004.

[73] C. D. Levermore. A Chapman–Enskog approach to flux limited diffusion theory. *Technical Report, Lawrence Livermore Laboratory*, UCID-18229, 1979.

[74] C.D. Levermore. Moment closure hierarchies for kinetic theories. *J. Statist. Phys.*, 5-6(83):1021–1065, 1996.

[75] C.D. Levermore. Entropy-based moment closures for kinetic equations. *Proceedings of the International Conference on Latest Developments and Fundamental Advances in Radiative Transfer (Los Angeles, CA, 1996)*, 26:591–606, 1997.

[76] C.D. Levermore and W.J. Morokoff. The Gaussian moment closure for gas dynamics. *SIAM J. Appl. Math.*, 1(59):72–96, 1999.

[77] C.D. Levermore and G.C. Pomraning. A flux-limited diffusion theory. *Astrophys. J*, (248):321–334, 1981.

[78] P.-L. Lions, B. Perthame, and P. Souganidis. Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. *Comm. Pure Appl. Math.*, 6(49):599–638, 1996.

[79] P.-L. Lions, B. Perthame, and E. Tadmor. Kinetic formulation of the isentropic gas dynamics and p-systems. *Comm. Math. Phys.*, 2(163):415–431, 1994.

[80] P.-L. Lions and G. Toscani. Diffuse limit for finite velocity Boltzmann kinetic models. *Rev. Mat. Ib.*, (13):473–513, 1997.

[81] F. Liotta, V. Romani, and G. Russo. Central scheme for balance law of relaxation type. *SIAM J. Numer. Anal.*, (38):1337–1356, 2000.

[82] A. Majorana, O. Muscato, and C. Milazzo. Comparison of Monte Carlo and Boltzmann equation simulations for 1-D silicon devices. *COMPEL*, (23):410–425, 2004.

[83] A. Majorana and R.M. Pidatella. A finite difference scheme solving the Boltzmann-Poisson system for semiconductor devices. *J. Comput. Phys.*, (174):649–668, 2001.

[84] P.A. Markowich, C. Ringhofer, and C. Schmeiser. Semiconductor equations. *Springer-Verlag, New-York*, 1990.

[85] G. Naldi, L. Pareschi, and G. Toscani. Relaxation schemes for partial differential equations and applications to degenerate diffusion problems. *Surveys Math. Indust.*, 4(10):315–343, 2002.

[86] R. Natalini. A discrete kinetic approximation of entropy solutions to multidimensional scalar conservation laws. *J. Differential Equations*, 2(148):292–317, 1998.

[87] G.L. Olson, L.H. Auer, and M.L. Hall. Diffusion, P1, and other approximation of radiation transport. *JQSRT*, (64):619–634, 2000.

[88] B. Perthame. Kinetic formulation of conservation laws. *Oxford lecture Series in Mathematics and its Applications, Oxford University Press, Oxford*, 21, 2002.

[89] B. Perthame and E. Tadmor. A kinetic equation with kinetic entropy functions for scalar conservation laws. *Comm. Math. Phys.*, 3(136):501–517, 1991.

[90] L. Reggiani. Hot-electron transport in semiconductors. *Topics in Applied Physics, Springer Verlag, Berlin*, (58), 1985.

[91] C. Ringhofer. Space-time discretization of series expansion methods for the Boltzmann Transport Equation. *SIAM J. Numer. Anal.*, (38):442–465, 2000.

[92] C. Ringhofer. A mixed spectral - difference method for the steady state Boltzmann - Poisson system. *SIAM J. Numer. Anal.*, (41):64–89, 2003.

[93] K. Sebastian and C.-W. Shu. Multidomain WENO finite difference method with interpolation at subdomain interfaces. *J. Sci. Comput.*, (19):405–438, 2003.

[94] C.-W. Shu. Essentially Non-Oscillatory and Weighted Essentially Non-Oscillatory schemes for hyperbolic conservation laws, Advanced numerical approximation of nonlinear hyperbolic equations (cetraro, 1997). *Lecture Notes in Math.*, 1697:325–432, 1998.

[95] C.-W. Shu and S. Osher. Efficient implementation of Essentially Non-Oscillatory shock capturing schemes. *J. Comput. Phys.*, (77):439–471, 1988.

[96] C.-W. Shu and S. Osher. Efficient implementation of Essentially Non Oscillatory shock capturing schemes. *J. Comput. Phys.*, (77):439–471, 1988.

[97] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo. The semi-lagrangian method for the numerical resolution of the Vlasov equation. *J. Comput. Phys.*, (149):201–220, 1999.

[98] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, (5):506–517, 1968.

[99] B. Su and G.L. Olson. Analytical benchmark for non-equilibrium radiative transfer in anisotropically scattering medium. *Annals of Nuclear Energy*, 13(24):1035–1055, 1997.

[100] K. Tomizawa. Numerical simulation of submicron semiconductor devices. *Artech House, Boston*, 1983.

[101] N. Vauchelet. Ph.D. Thesis. *Université Paul Sabatier*, 2006.

[102] X. Yang, F. Golse, Z. Huang, and S. Jin. Numerical study of a domain decomposition method for a two-scale linear transport equation. *Networks and Heterogenous Media*, 1(1):143–166, 2006.

[103] T. Zhou, Y. Guo, and C.-W. Shu. Numerical study on Landau damping. *Physica D*, (157):322–333, 2001.