Eric Sonnendrücker

Numerical Methods for the Vlasov-Maxwell equations

SPIN Springer's internal project number, if known

– Monograph –

January 29, 2015

Springer

Contents

1 Introduction 1
- 1.1 Plasmas 1
- 1.2 Controlled thermonuclear fusion 1
- 1.3 The ITER project 4

2 Some models used in plasma physics 5
- 2.1 The Vlasov-Maxwell equations 5
- 2.2 The N-body model 6
- 2.3 Kinetic models 7
  - 2.3.1 The Vlasov-Maxwell model 8
  - 2.3.2 The Boltzmann operator 9
  - 2.3.3 The non-linear Fokker-Planck-Landau operator 11
  - 2.3.4 The linear Fokker-Planck operator 11
  - 2.3.5 The BGK operator 12
- 2.4 Fluid models 12
- 2.5 Some application specific approximations 16
  - 2.5.1 The paraxial approximation 16
  - 2.5.2 The gyrokinetic approximation 20
  - 2.5.3 The guiding-center model 21
- 2.6 Expressions of the Maxwell equations 22
  - 2.6.1 The 3D Maxwell equations 22
  - 2.6.2 The 2D Maxwell equations 23
  - 2.6.3 The 1D Maxwell equations 23

3 Some theory on Vlasov systems 25
- 3.1 The linear Vlasov equation 25
- 3.2 The Vlasov-Poisson system 30
  - 3.2.1 The equations 30
  - 3.2.2 Conservation properties 30
- 3.3 Solution of the linearised Vlasov-Poisson equations 34
- 3.4 The Vlasov-Maxwell system 45

3.4.1 The equations . . . 45
3.4.2 Conservation properties . . . 45

4 Code verification . . . 49
4.1 Introduction . . . 49
4.2 Test of the field solver . . . 50
4.2.1 Test of the Poisson solver . . . 50
4.2.2 Test of the Maxwell solver . . . 50
4.3 Test of the Vlasov solver with a given advection field . . . 51
4.4 Comparison with analytical solution of the linearised Vlasov-Poisson system . . . 52
4.4.1 Computation of the zeros of an analytic function . . . 53
4.4.2 Landau damping . . . 54
4.4.3 The two stream instability . . . 57

5 Basic numerical tools . . . 63
5.1 Operator splitting . . . 63
5.2 Discrete Fourier Transform . . . 65
5.2.1 Definition . . . 65
5.2.2 Approximation of the coefficients of a Fourier series using the FFT . . . 67
5.3 Circulant matrices . . . 68
5.4 Interpolation . . . 70
5.4.1 Splines . . . 70
5.4.2 Using B-splines for spline interpolation . . . 73
5.4.3 Cubic spline interpolation . . . 74
5.4.4 Fast local spline interpolation . . . 76
5.4.5 Parallelization . . . 79
5.4.6 Lagrange interpolation . . . 79

6 Numerical methods for the Maxwell equation . . . 81
6.1 Spectral method for the Poisson equation . . . 81
6.2 Discretization of the 1D Maxwell equations . . . 82
6.2.1 Centered finite difference discretization . . . 83
6.2.2 Mixed Finite Element formulation . . . 84
6.2.3 B-spline Finite Elements . . . 88
6.3 High-order time schemes . . . 91
6.3.1 Schemes based on a Taylor expansion . . . 91
6.3.2 Leap-frog schemes . . . 91
6.3.3 Conservation of a discrete energy for the leap-frog scheme . . . 93
6.3.4 Stability of time schemes . . . 95
6.3.5 Explicit computation of the stability condition when the matrices are circulant . . . 99
6.4 Discretization of the 2D Maxwell equations . . . 99

6.4.1 The Yee scheme 99
6.4.2 Variationnal formulations for the 2D Maxwell equations 102
6.4.3 Discretization using conforming finite elements 104
6.5 Discretization of the 3D Maxwell equations 109
6.6 The Discontinuous Galerkin (DG) method 109
6.6.1 Principle of the method 110
6.6.2 Matrix formulation of the discrete problem 112
6.6.3 Computation of the fluxes 116
6.6.4 Semi-discrete energy conservation 118
6.6.5 The time scheme 119

7 Semi-Lagrangian approximation of the Vlasov equation 121
7.1 The classical semi-Lagrangian method 122
7.2 Conservation properties of the semi-Lagrangian method 126
7.2.1 Conservation of mass 126
7.2.2 Conservation of total momentum 127
7.3 Numerical integration of the characteristics 129
7.3.1 Origin of the characteristics for the non split 1D Vlasov-Poisson equations 129
7.3.2 The general case 130
7.3.3 Case of 1D characteristics with linear interpolation 131
7.4 Importance of conservativity 133
7.5 The conservative semi-Lagrangian method 134
7.5.1 Stabilization. 137
7.5.2 Equivalence of conservative and classical semi-Lagrangian methods 139
7.6 The forward semi-Lagrangian method 139
7.6.1 The general algorithm 140
7.6.2 An explicit computation of the characteristics 142

8 Particle approximation of the Vlasov equation 145
8.1 Introduction 145
8.2 The PIC method 146
8.2.1 Consequence 147
8.2.2 Choice of the initial condition. 147
8.2.3 Particle-Mesh coupling. 148
8.2.4 Conservation properties at the semi-discrete level. 149
8.2.5 Time scheme for the particles. 150
8.2.6 Time loop. 151
8.3 Monte Carlo Simulation 151
8.3.1 Principle 151
8.3.2 Estimation of the error in a Monte Carlo simulation . . . 152
8.3.3 Error monitoring in PIC codes 155
8.3.4 Error on the probability density 156
8.3.5 Aliasing 159

8.4 Initialisation of given PDF . . . 161
8.4.1 Inversion of the CDF . . . 161
8.4.2 Acceptance-rejection method . . . 163
8.4.3 Composition method . . . 163
8.5 Variance reduction techniques . . . 164
8.5.1 Control variates . . . 164
8.5.2 Importance sampling . . . 165
8.5.3 Application to the PIC method. . . . 167
8.6 Coupling the Monte Carlo Vlasov solver with a grid based Poisson solver . . . 170
8.6.1 Finite Difference PIC methods . . . 170
8.6.2 Finite Element PIC methods . . . 172

9 Coupling the Vlasov and Maxwell equations . . . 175
9.1 Introduction . . . 175
9.2 Generalised Maxwell's equations . . . 177
9.3 Structure preserving discretizations . . . 178
9.3.1 Enforcing a discrete continuity equation . . . 179
9.3.2 Conforming mixed Finite Elements . . . 180
9.3.3 The finite difference Yee scheme . . . 182

A Complex analysis and Laplace transform . . . 185
A.1 Analytic functions . . . 185
A.2 Path integration . . . 186
A.3 Laplace transform . . . 187

B Background in probability theory . . . 191
B.1 Probability spaces . . . 191
B.2 Random variables . . . 193
B.3 Distribution function . . . 193
B.4 Expected value, variance. . . . 194
B.5 Conditional probabilities and independence . . . 196
B.6 Stochastic processes . . . 198
B.7 The Itô integral . . . 199
B.8 Stochastic differential equations . . . 199
B.9 The Kolmogorov forward and backward equations . . . 200

References . . . 203

Index . . . 209

1

Introduction

1.1 Plasmas

When a gas is brought to a very high temperature ($10^4\,K$ or more) electrons leave their orbit around the nuclei of the atom to which they are attached. This gives an overall neutral mixture of charged particles, ions and electrons, which is called plasma. Plasmas are considered beside solids, liquids and gases, as the fourth state of matter.

You can also get what is called a non-neutral plasma, or a beam of charged particles, by imposing a very high potential difference so as to extract either electrons or ions of a metal chosen well. Such a device is usually located in the injector of a particle accelerator.

The use of plasmas in everyday life have become common. These include, for example, neon tubes and plasma displays. There are also a number industrial applications: amplifiers in telecommunication satellites, plasma etching in microelectronics, production of X-rays.

We should also mention that while it is almost absent in the natural state on Earth, except the Northern Lights at the poles, the plasma is 99% of the mass of the visible universe. Including the stars are formed from plasma and the energy they release from the process of fusion of light nuclei such as protons. More information on plasmas and their applications can be found on the web site `http://www.plasmas.org`.

1.2 Controlled thermonuclear fusion

The evolution of energy needs and the depletion of fossil fuels make it essential to develop new energy sources. According to the well-known formula $E = mc^2$, we can produce energy by performing a transformation that removes the mass. There are two main types of nuclear reactions with this. The fission reaction of generating two lighter nuclei from the nucleus of a heavy atom and the fusion reaction that is created from two light atoms a heavier nucleus. Fission is

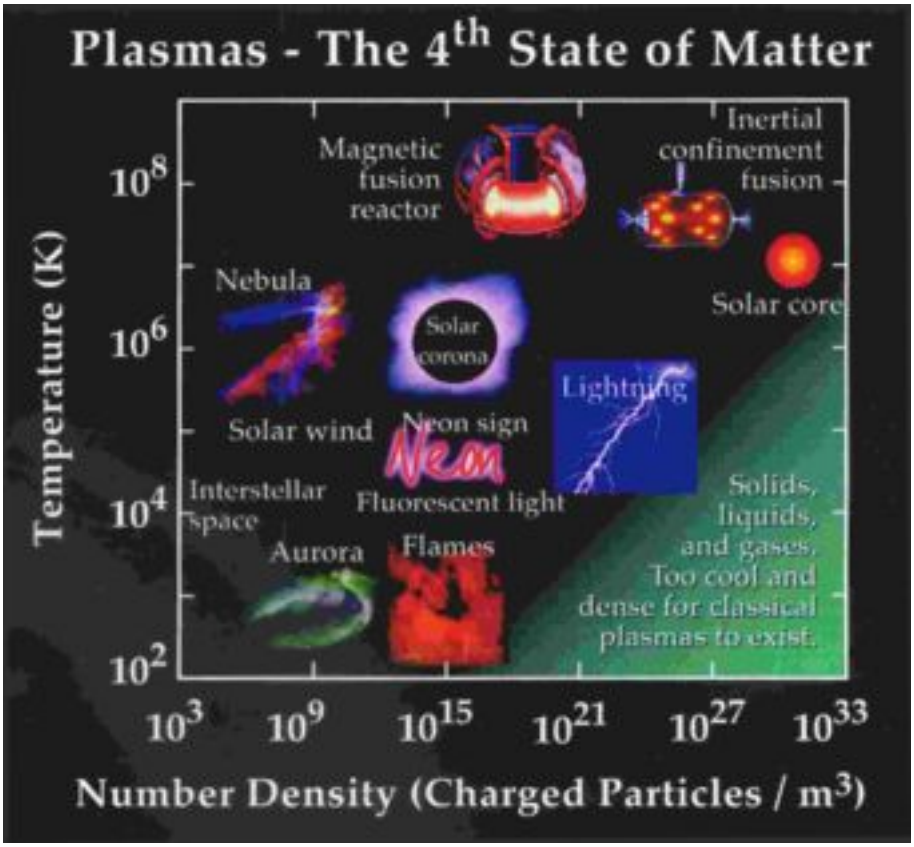

Fig. 1.1. Examples of plasmas at different densities and temperatures

used in existing nuclear power plants. Controlled fusion is still in the research stage.

The fusion reaction is the most accessible to fuse nuclei of deuterium and tritium, which are isotopes of hydrogen, for a helium atom and a neutron high energy will be used to produce the heat necessary to manufacture electricity (see Fig. 1.2).

The temperatures required for thermonuclear fusion exceed one hundred million degrees. At these temperatures the electrons are totally freed from their atoms so that one obtains a gas of electrons and ions which is a totally ionized plasma. To produce energy, it is necessary that the amplification factor Q which is the ratio of the power produced to the external power supplied is greater than one. Energy balance allows for the Lawson criterion that connects the amplification factor Q the product nTt_E where n is the plasma density, T its temperature and t_E energy confinement time in the plasma.

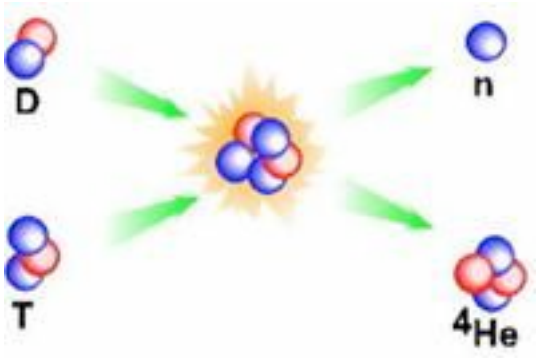

Fig. 1.2. The Deuterium-Tritium fusion reaction

Fusion is the basis of the energy of stars in which a confinement at a sufficient density is provided by their mass. The research on controlled fusion on Earth is considering two approaches. On the one hand inertial confinement fusion aims at achieving a very high density for a relatively short time by shooting on a capsule of deuterium and tritium beams with lasers. On the other hand magnetic confinement fusion consists in confining the plasma with a magnetic field at a lower density but for a longer time. The latter approach is pursued in the ITER project whose construction has just started at Cadarache in the south-eastern France. The plasma is confined in a toroidal-shaped chamber called a tokamak that for ITER is shown in Figure 1.3.

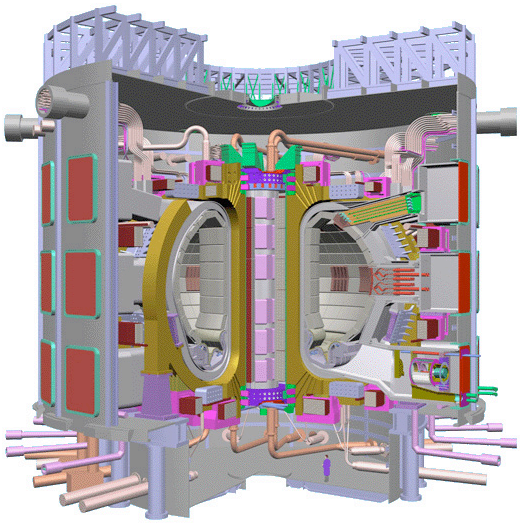

Fig. 1.3. Artist view of the ITER Tokamak

There are also experimental facilities (NIF in the USA and LMJ in France) are being built for experimental validation of the concept of inertial confinement fusion using lasers.

Note that an alternative option to lasers for inertial confinement using heavy ions beams is also pursued. See `http://hif.lbl.gov/tutorial/tutorial.html` for more details.

More information on fusion can be found on wikipedia `http://en.wikipedia.org/wiki/Inertial_confinement_fusion` for inertial fusion and `http://en.wikipedia.org/wiki/Magnetic_confinement_fusion` for magnetic fusion.

The current record fusion power produced for a deuterium-tritium reaction is equal to 16 megawatts, corresponding to an amplification factor $Q = 0.64$. It was obtained in the JET tokamak in England. It is well established that to obtain an amplification factor much greater than one, it is necessary to use a greater machine, hence the need for the construction of the ITER tokamak, which will contain a plasma volume five times larger than that of JET, to demonstrate the feasibility of a power plant based on magnetic fusion. The amplification factor provided in ITER should be greater than 10.

1.3 The ITER project

The ITER project is a partnership between the European Union, Japan, China, South Korea, Russia, the United States and India for which an international agreement was signed November 21, 2006 in Paris. It aims to demonstrate the scientific and technical feasibility of producing electricity from fusion energy for which there are significant resources of fuel and which has a low impact on the environment.

The construction of the ITER tokamak is under way in Cadarache in the south-eastern France and the operational phase is expected to begin in 2019 and last for two decades. The main objectives of ITER are firstly to achieve an amplification factor greater than 10 and so really allow the production of energy, secondly to implement and test the technologies needed for a fusion power plant and finally to test concepts for the production of Tritium from Lithium belt used to absorb the energy of neutrons.

If successful the next step called DEMO will be to build a fusion reactor fusion that will actually produce energy before moving on to commercial fusion power plants.

More information is available on the web site `http://www.iter.org`.

2

Some models used in plasma physics

2.1 The Vlasov-Maxwell equations

We consider in this lecture more specifically one of the models commonly used to describe the evolution of a plasma and which is called a kinetic model. It is based on the Vlasov equation which describes the evolution of charged particles in an electromagnetic field which can either be self-consistent, that is to say, generated by the particles themselves, or externally applied, or most often, both. It is written for non-relativistic particles

$$\frac{\partial f_s}{\partial t} + \mathbf{v} \cdot \frac{\partial f_s}{\partial \mathbf{x}} + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial f_s}{\partial \mathbf{v}} = 0,$$

where m is the mass particles, q their charge and $f \equiv f(\mathbf{x}, \mathbf{v}, t)$ represents the particle density in phase space at point $(\mathbf{x}, \mathbf{v})$ and at time t. It has the structure of a transport equation in phase space which includes the three dimensions of physical space and the three dimensions of velocity space (or momentum in the relativistic case). The self-consistent electromagnetic field can be calculated by coupling with Maxwell's equation with sources that are the charge densities and current calculated from the particles:

$$\begin{aligned} -\frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} &= \mu_0 \mathbf{J}, \\ \frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} &= 0, \\ \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}, \\ \nabla \cdot \mathbf{B} &= 0, \end{aligned}$$

with

$$\rho(\mathbf{x}, t) = q \int f(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v}, \quad \mathbf{J}(\mathbf{x}, t) = q \int f(\mathbf{x}, \mathbf{v}, t)\mathbf{v}\, d\mathbf{v}.$$

Plasmas, in particular fusion plasmas are extremely complex objects, involving nonlinear interactions and a large variety of time and space scales.

They are subject to many instabilities and turbulence phenomena that make their confinement challenging. The road to fusion as an energy source therefore requires a very fine understanding of plasmas using appropriate models and numerical simulations based on these models.

The numerical solution of the three-dimensional Vlasov-Maxwell system is a major challenge if only because of the huge size of the system due to the fact that the Vlasov equation is posed in the 6D phase space and the non linear coupling between Vlasov and Maxwell. The seven variables to consider are the three variables giving the position in physical space and the three variable velocity over time. For the model to be used in practice, it will be necessary to use reduced models that can be precise enough with respect to certain characteristics of the studied system: symmetry, small parameters, etc.. Furthermore the specific properties of the Vlasov equation will require the use of numerical methods specifically designed for this kind of equations.

2.2 The *N*-body model

At the microscopic level, a plasma or a particle beam is composed of a number of particles that evolve following the laws of classical or relativistic dynamics. So each particle obeys Newton's law

$$\frac{d\gamma m\mathbf{v}}{dt} = \sum F_{ext},$$

where m is the mass of the particle, $\mathbf{v}$ its velocity $\gamma = (1 - \frac{|\mathbf{v}|^2}{c^2})^{-\frac{1}{2}}$ is the Lorentz factor (c being the speed of light). The right hand side F_{ext} is composed of all the forces applied to the particle, which in our case reduce to the Lorentz force induced by the external and self-consistent electromagnetic fields. Other forces as the weight of the particles are in general negligible. Whence we have

$$\frac{d\gamma_i m\mathbf{v}_i}{dt} = \sum_j q(\mathbf{E}_j + \mathbf{v}_i \times \mathbf{B}_j).$$

On the other hand the velocity of a particle $\mathbf{v}_i$ is linked to its position $\mathbf{x}_i$ by

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{v}_i.$$

Thus, if the initial positions and velocities of the particles are known as well as the external fields, the evolution of the particles is completely determined by the equations

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{v}_i, \tag{2.1}$$

$$\frac{d\gamma_i m\mathbf{v}_i}{dt} = \sum_j q(\mathbf{E}_j + \mathbf{v} \times \mathbf{B}_j), \tag{2.2}$$

where the sum contains the electric and magnetic field generated by each of the other particles as well as the external fields.

Remark 1 *This system is Hamiltonian, which can be seen easily in the non relativistic case without magnetic field. In this case the electric field derives from a scalar potential:* $\mathbf{E} = -\nabla\phi$. *The hamiltonian then reads*

$$H = \frac{v_i^2}{2} + \frac{q}{m}\phi.$$

And so

$$\frac{d\mathbf{x}_i}{dt} = \frac{\partial H}{\partial \mathbf{v}_i} = \mathbf{v}_i,$$
$$\frac{d\mathbf{v}_i}{dt} = -\frac{\partial H}{\partial \mathbf{x}_i} = -\frac{q}{m}\nabla\phi = \frac{q}{m}\mathbf{E}$$

The motion of the particles is also hamiltonian in the general case, but one needs to transform to specific coordinates which are called canonical coordinates to exhibit the hamiltonian structure.

In general a plasma consists of a large number of particles, 10^{10} and more. The microscopic model describing the interactions of particles with each other is not used in a simulation because it would be far too expensive. We must therefore find approximate models which, while remaining accurate enough can reach a reasonable computational cost. There is actually a hierarchy of models describing the evolution of a plasma. The base model of the hierarchy and the most accurate model is the N-body model we have described, then there are intermediate models called kinetic and which are based on a statistical description of the particle distribution in phase space and finally the macroscopic or fluid models that identify each species of particles of a plasma with a fluid characterized by its density, its velocity and energy. Fluid models are becoming a good approximation when the particles are close to thermodynamic equilibrium, to which they return in long time do to the effects of collisions and for which the distribution of particle velocities is a Gaussian.

2.3 Kinetic models

In a *kinetic* model, each particle species s in the plasma is characterized by a distribution function $f_s(\mathbf{x}, \mathbf{v}, t)$ which corresponds to a statistical mean of the repartition of particles in phase space for a large number of realisations of the considered physical system. The product $f_s\, d\mathbf{x}\, d\mathbf{v}$ is the average number of particles of species s, whose position and velocity are in the box of volume $d\mathbf{x}\, d\mathbf{v}$ centred at $(\mathbf{x}, \mathbf{v})$.

The distribution function contains much more information than a fluid description as it includes information on the distributions of particle velocities at each position. A kinetic description of a plasma is essential when the

distribution function is far away from the Maxwell-Boltzmann distribution (also called Maxwellian) that corresponds to the thermodynamic equilibrium of plasma. Otherwise a fluid description is sufficient.

2.3.1 The Vlasov-Maxwell model

In the limit where the collective effects are dominant on binary collisions between particles, the kinetic equation that is derived, by methods of statistical physics from the N-body model is the *Vlasov* equation which reads

$$\frac{\partial f_s}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f_s + \frac{q_s}{m_s}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{v}} f_s = 0, \tag{2.3}$$

in the non relativistic case. In the relativistic case it becomes

$$\frac{\partial f_s}{\partial t} + \mathbf{v}(\mathbf{p}) \cdot \nabla_{\mathbf{x}} f_s + q_s(\mathbf{E} + \mathbf{v}(\mathbf{p}) \times \mathbf{B}) \cdot \nabla_{\mathbf{p}} f_s = 0. \tag{2.4}$$

We denote by $\nabla_{\mathbf{x}} f_s$, $\nabla_{\mathbf{v}} f_s$ and $\nabla_{\mathbf{p}} f_s$, the respective gradients of f_s with respect to the three position, velocity and momentum variables. The constants q_s and m_s denote the charge and mass of the particle species. The velocity is linked to the momentum by the relation $\mathbf{v}(\mathbf{p}) = \frac{\mathbf{p}}{m_s \gamma_s}$, where γ is the Lorentz factor which can be expressed from the momentum by $\gamma_s = \sqrt{1 + |\mathbf{p}|^2/(m_s^2 c^2)}$.

This equation expresses that the distribution function f is conserved along the trajectories of the particles which are determined by the mean electric field. We denote by $f_{s,0}(\mathbf{x}, \mathbf{v})$ the initial value of the distribution function. The Vlasov equation, when it takes into account the self-consistent electromagnetic field generated by the particles, is coupled to the Maxwell equations which enable to computed this self-consistent electromagnetic field from the particle distribution:

$$\begin{aligned} -\frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} &= \mu_0 \mathbf{J}, \\ \frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} &= 0, \\ \nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon_0}, \\ \nabla \cdot \mathbf{B} &= 0. \end{aligned}$$

The source terms for Maxwell's equation, the charge density $\rho(\mathbf{x}, t)$ and the current density $\mathbf{J}(\mathbf{x}, t)$ can be expressed from the distribution functions of the different species of particles $f_s(\mathbf{x}, \mathbf{v}, t)$ using the relations

$$\begin{aligned} \rho(\mathbf{x}, t) &= \sum_s q_s \int f_s(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v}, \\ \mathbf{J}(\mathbf{x}, t) &= \sum_s q_s \int f_s(\mathbf{x}, \mathbf{v}, t)\mathbf{v}\, d\mathbf{v}. \end{aligned}$$

Note that in the relativistic case the distribution function becomes a function of position and momentum (instead of velocity): $f_s \equiv f_s(\mathbf{x}, \mathbf{p}, t)$ and charge and current densities verify

$$\rho(\mathbf{x}, t) = \sum_s q_s \int f_s(\mathbf{x}, \mathbf{p}, t)\, d\mathbf{p}, \qquad \mathbf{J}(\mathbf{x}, t) = \sum_s q_s \int f_s(\mathbf{x}, \mathbf{p}, t)\mathbf{v}(\mathbf{p})\, d\mathbf{p}.$$

The macroscopic quantities, associated to each particle species are defined as follows:

- The particle density, in physical space, for species s, is defined by

$$n_s(\mathbf{x}, t) = \int f_s(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v},$$

- The mean velocity $\mathbf{u}_s(\mathbf{x}, t)$ verifies

$$n_s(\mathbf{x}, t)\mathbf{u}_s(\mathbf{x}, t) = \int f_s(\mathbf{x}, \mathbf{v}, t)\mathbf{v}\, d\mathbf{v},$$

- The kinetic energy is defined by

$$n_s(\mathbf{x}, t)\mathcal{E}_s(\mathbf{x}, t) = \frac{m}{2} \int f_s(\mathbf{x}, \mathbf{v}, t)|\mathbf{v}|^2\, d\mathbf{v},$$

- The temperature $T_s(\mathbf{x}, t)$ is related to the kinetic energy, mean velocity and density by

$$T_s(\mathbf{x}, t) = \mathcal{E}_s(\mathbf{x}, t) - u_s^2(\mathbf{x}, t).$$

2.3.2 The Boltzmann operator

When binary collisions between particles are dominant with respect to mean field effects, the distribution function satisfies the Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} = \sum_s \mathcal{Q}(f, f_s),$$

where $\mathcal{Q}$ is the non linear Boltzmann operator. This operator is sometimes replaced by simpler models. A sum on the collisions with all the species of particles represented by f_s, including the particles of the same species, is considered. In many cases not all the collisions might be considered. In some intermediate cases, the collision operator appears on the right-hand side of the full Vlasov equation.

The Boltzmann collision operator for two species of particles (that might be identical, in which case $f_s = f$) writes

$$\mathcal{Q}(f, f_s)(\mathbf{v}) = \frac{1}{m} \int_{\mathbb{R}^3} \int_{S^2} B(|\mathbf{v} - \mathbf{v}_1|, \theta) \left[f(\mathbf{v}')f_s(\mathbf{v}_1') - f(\mathbf{v})f_s(\mathbf{v}_1) \right] d\mathbf{v}_1\, d\mathbf{n},$$

where $\mathbf{v}'$ and $\mathbf{v}_1'$ are the velocities after collision of the particles with velocity $\mathbf{v}$ and $\mathbf{v}_1$ before collision. The deflection angle θ is the angle between $\mathbf{v}-\mathbf{v}_1$ and $\mathbf{v}'-\mathbf{v}_1'$. The post-collision velocities are expressed by

$$\mathbf{v}' = \mathbf{v} - \frac{2\mu}{m}[(\mathbf{v}-\mathbf{v}_1)\cdot\mathbf{n}]\mathbf{n}, \quad \mathbf{v}_1' = \mathbf{v}_1 + \frac{2\mu}{m_s}[(\mathbf{v}-\mathbf{v}_1)\cdot\mathbf{n}]\mathbf{n},$$

with $\mu = \frac{mm_s}{m+m_s}$ and $\mathbf{n}$ a unit vector on the sphere S^2. These expressions are obtained by writing that momentum and kinetic energy are conserved during a collision. The collision kernel B is given. Its precise form depends on the properties of the gas.

Let us briefly sketch out some basic properties of the Boltzmann collision operator, see the book of Cercignani [31] for details.

Proposition 1 *For any continuous function φ, we have*

$$\int_{\mathbb{R}^3} \mathcal{Q}(f,f)\varphi(\mathbf{v})\,d\mathbf{v} = \frac{1}{4m}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{S^2}[\varphi(\mathbf{v})+\varphi(\mathbf{v}_1)-\varphi(\mathbf{v}')-\varphi(\mathbf{v}_1')]\,B(|\mathbf{v}-\mathbf{v}_1|,\theta) \\ [f(\mathbf{v}')f(\mathbf{v}_1') - f(\mathbf{v})f(\mathbf{v}_1)]\,d\mathbf{v}\,d\mathbf{v}_1\,d\mathbf{n}.$$

From which it follows that

$$\int_{\mathbb{R}^3} \mathcal{Q}(f,f)\varphi(\mathbf{v})\,d\mathbf{v} = 0$$

if and only if $\varphi(\mathbf{v})$ is a linear combination of 1, v_x, v_y, v_z and $|v|^2$.

The first relation is obtained by writing four equal expressions for $\int_{\mathbb{R}^3} \mathcal{Q}(f,f)\varphi(\mathbf{v})\,d\mathbf{v}$ obtained by changes of variables conserving $|\mathbf{v}-\mathbf{v}_1|$ and θ so that $B(|\mathbf{v}-\mathbf{v}_1|,\theta)$ is not modified and then expression the integral as the average of the four expressions.

Proposition 2 (Boltzmann inequality) *For $f>0$ we have*

$$\int_{\mathbb{R}^3} \mathcal{Q}(f,f)\ln f\,d\mathbf{v} \le 0.$$

Proposition 3 (H theorem) *For $f(t,\mathbf{x},\mathbf{v}) > 0$ a solution of the Vlasov-Boltzmann equation, we define*

$$H(t) = \int_{\mathbb{T}^3}\int_{\mathbb{R}^3} f\ln f\,d\mathbf{x}\,d\mathbf{v} \le 0.$$

Then

$$\frac{dH}{dt} \le 0,$$

and the inequality is strict if f is not of the form $f(\mathbf{v}) = \exp(a+\mathbf{b}\cdot\mathbf{v}+cv^2)$ (i.e. a Maxwellian).

A consequence of the H theorem, is that the solution of the Vlasov-Boltzmann equation relaxes when time goes to infinity to a minimum of H which is a Maxwellian. The precise definition of a Maxwellian with a given density, mean velocity and temperature will be given in section 2.4.

2.3.3 The non-linear Fokker-Planck-Landau operator

The Fokker-Planck-Landau operator, which is the most commonly used in plasma physics, is the limit of the Boltzmann operator for grazing collisions [75]. It reads for collisions with particles of the same species

$$Q_L(f,f)(v) = \nabla_v \cdot \int A(\mathbf{v}-\mathbf{v}_*)[f(\mathbf{v}_*)\nabla_v f(\mathbf{v}) - f(\mathbf{v})\nabla_{v_*} f(\mathbf{v}_*)]\,\mathrm{d}\mathbf{v}_*] \quad (2.5)$$

where, for the case of Coulomb collisions, Λ being a positive constant,

$$A(\mathbf{v}-\mathbf{v}_*) = \frac{\Lambda}{|\mathbf{v}-\mathbf{v}_*|}\left(\mathbb{I}_3 - \frac{(\mathbf{v}-\mathbf{v}_*)\otimes(\mathbf{v}-\mathbf{v}_*)}{|\mathbf{v}-\mathbf{v}_*|^2}\right).$$

As for the Boltzmann operator, the first three velocity moments of the Landau operator vanish, which implies conservation of mass, momentum and kinetic energy. Moreover the H-theorem is satisfied, so that the equilibrium states are also the Maxwellians.

Sometimes it is more convenient to express the Landau collision operator using the Rosenbluth potentials [92], which reads

$$Q_{L,R}(f,f_*)(v) = \Lambda\nabla_v \cdot [\nabla_v \cdot (f\nabla_v \otimes \nabla_v \mathbf{G}(f_*)) - 4f\nabla_v \cdot \mathbf{H}(f_*)]$$

where the Rosenbluth potentials are defined by

$$\mathbf{G}(f_*)(\mathbf{v}) = \int |\mathbf{v}-\mathbf{v}_*| f_*(\mathbf{v}_*)\,\mathrm{d}\mathbf{v}_*, \qquad \mathbf{H}(f_*)(\mathbf{v}) = \int \frac{1}{|\mathbf{v}-\mathbf{v}_*|} f_*(\mathbf{v}_*)\,\mathrm{d}\mathbf{v}_*. \quad (2.6)$$

2.3.4 The linear Fokker-Planck operator

When the distribution function is close enough to a Maxwellian, the Fokker-Planck-Landau operator can be linearised around a Maxwellian, the collision operator takes a much simpler for as can be seen from the Rosenbluth potential for by taking f_* in the expression of the potentials $\mathbf{G}$ and $\mathbf{H}$ to be a given Maxwellian. Then we get the linear Fokker-Planck operator, which takes the form

$$Q_{FP}(f)(v) = \nu\nabla_v \cdot (\mu f\mathbf{v} + \frac{D^2}{2}\nabla_v f). \quad (2.7)$$

This operator is also known as the Lenard-Bernstein operator in the plasma physics community [78]. For given constants ν, μ and D, the Lenard-Bernstein operator conserves mass, but not momentum and energy and its equilibrium function is a Maxwellian of the form $\alpha e^{-\frac{\mu v^2}{D^2}}$, where the constant α is determined by the total mass (or number of particles).

The linear Fokker-Planck operator can be made to conserve also total momentum and kinetic energy by using the mean velocity $\mathbf{u}$ and the temperature T associated to f. Then the operator reads

$$Q_{FPC}(f)(v) = \nu\nabla_v \cdot (f\frac{\mathbf{v}-\mathbf{u}}{T} + \nabla_v f).$$

2.3.5 The BGK operator

A simplified collision operator that has been build to conserve mass, momentum and kinetic energy and have the Maxwellian as equilibrium states, as the Boltzmann and Fokker-Planck-Landau operators, has been derived by Bhatnagar, Gross and Krook [16]. In the mathematics community this is known as the BGK operator and in the physics community it is called the Krook operator. It simply reads

$$Q_K(f)(v) = \nu(f_M[f] - f),$$

where $f_M[f]$ is the Maxwellian, which has the same mass, mean velocity and temperature as f. It reads

$$f_M(t, \mathbf{x}, \mathbf{v}) = \frac{n(t, \mathbf{x})}{(2\pi T(t, \mathbf{x})/m)^{\frac{3}{2}}} e^{-\frac{|\mathbf{v}-\mathbf{u}(\mathbf{x},t)|^2}{2T(\mathbf{x},t)/m}}.$$

2.4 Fluid models

Due to collisions, the particles relax in long time to a Maxwellian, which is a thermodynamical equilibrium. When this state is approximately attained particles can be described by a fluid like model.

This fluid model can be derived from the Vlasov equations. The fluid model will still be coupled to Maxwell's equation for the determination of the self-consistent electromagnetic field.

We start from the Vlasov-Boltzmann equation:

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial f}{\partial \mathbf{v}} = \mathcal{Q}(f, f). \tag{2.8}$$

Remark 2 *The Boltzmann collision operator $Q(f, f)$ on the right hand side is necessary to provide the relaxation to thermodynamic equilibrium. However it will have no direct influence on our derivation, as we will consider only the first three velocity moments which vanish for the Boltzmann operator.*

The macroscopic quantities on which the fluid equations will be established are defined using the first three velocity moments of the distribution function $f(\mathbf{x}, \mathbf{v}, t)$

- The particle density is defined by

$$n(\mathbf{x}, t) = \int f(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v},$$

- The mean velocity $\mathbf{u}(\mathbf{x}, t)$ verifies

$$n(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t) = \int f(\mathbf{x}, \mathbf{v}, t)\mathbf{v}\, d\mathbf{v},$$

- The pressure tensor $\mathbb{P}(\mathbf{x},t)$ is defined by

$$\mathbb{P}(\mathbf{x},t) = m\int f(\mathbf{x},\mathbf{v},t)(\mathbf{v}-\mathbf{u}(\mathbf{x},t))\otimes(\mathbf{v}-\mathbf{u}(\mathbf{x},t))\,d\mathbf{v}.$$

- The scalar pressure is one third of the trace of the pressure tensor

$$p(\mathbf{x},t) = \frac{m}{3}\int f(\mathbf{x},\mathbf{v},t)|\mathbf{v}-\mathbf{u}(\mathbf{x},t)|^2\,d\mathbf{v},$$

- The temperature $T(\mathbf{x},t)$ is related to the pressure and the density by

$$T(\mathbf{x},t) = \frac{p(\mathbf{x},t)}{n(\mathbf{x},t)}.$$

- The energy flux is a vector defined by

$$\mathbf{Q}(\mathbf{x},t) = \frac{m}{2}\int f(\mathbf{x},\mathbf{v},t)|\mathbf{v}|^2\mathbf{v}(\mathbf{x},t))\,d\mathbf{v}.$$

where we denote by $|\mathbf{v}| = \sqrt{\mathbf{v}\cdot\mathbf{v}}$ and for two vectors $\mathbf{a} = (a_1,a_2,a_3)^T$ and $\mathbf{b} = (b_1,b_2,b_3)^T$, their tensor product $\mathbf{a}\otimes\mathbf{b}$ is the 3×3 matrix whose components are $(a_ib_j)_{1\le i,j\le 3}$.

We obtain equations relating these macroscopic quantities by taking the first velocity moments of the Vlasov equation. In the actual computations we shall make use that f vanishes at infinity and that the plasma is periodic in space. This takes care of all boundary condition problems.

Let us first notice that as $\mathbf{v}$ is a variable independent of $\mathbf{x}$, we have $\mathbf{v}\cdot\nabla_x f = \nabla_x\cdot(f\mathbf{v})$. Moreover, as $\mathbf{E}(\mathbf{x},t)$ does not depend on $\mathbf{v}$ and that the i^{th} component of

$$\mathbf{v}\times\mathbf{B}(\mathbf{x},t) = \begin{pmatrix} v_2B_3(\mathbf{x},t) - v_3B_2(\mathbf{x},t)\\ v_3B_1(\mathbf{x},t) - v_1B_3(\mathbf{x},t)\\ v_1B_2(\mathbf{x},t) - v_2B_1(\mathbf{x},t)\end{pmatrix}$$

is independent of v_i, we also have

$$(\mathbf{E}(\mathbf{x},t)+\mathbf{v}\times\mathbf{B}(\mathbf{x},t))\cdot\nabla_v f = \nabla_v\cdot(f(\mathbf{E}(\mathbf{x},t)+\mathbf{v}\times\mathbf{B}(\mathbf{x},t))).$$

Integrating the Vlasov equation (2.8) with respect to velocity $\mathbf{v}$ we obtain

$$\frac{\partial}{\partial t}\int f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} + \nabla_x\cdot\int f(\mathbf{x},\mathbf{v},t)\mathbf{v}\,d\mathbf{v} + 0 = 0.$$

Whence, as $n(\mathbf{x},t)\mathbf{u}(\mathbf{x},t) = \int f(\mathbf{x},\mathbf{v},t)\mathbf{v}\,d\mathbf{v}$, we get

$$\frac{\partial n}{\partial t} + \nabla_x\cdot(n\mathbf{u}) = 0. \tag{2.9}$$

Multiplying the Vlasov by $m\mathbf{v}$ and integrating with respect to $\mathbf{v}$, we get

$$m\frac{\partial}{\partial t}\int f(\mathbf{x},\mathbf{v},t)\mathbf{v}\,d\mathbf{v} + m\nabla_x \cdot \int (\mathbf{v}\otimes\mathbf{v}) f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} \\ - q(\mathbf{E}(\mathbf{x},t)\int f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} + \int f(\mathbf{x},\mathbf{v},t)\mathbf{v}\,d\mathbf{v}\times\mathbf{B}(\mathbf{x},t) = 0.$$

Moreover,

$$\int \mathbf{v}\otimes\mathbf{v} f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} = \int (\mathbf{v}-\mathbf{u})\otimes(\mathbf{v}-\mathbf{u}) f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} + n\mathbf{u}\otimes\mathbf{u}.$$

Whence

$$m\frac{\partial}{\partial t}(n\mathbf{u}) + m\nabla\cdot(n\mathbf{u}\otimes\mathbf{u}) + \nabla\cdot\mathbb{P} = qn(\mathbf{E}+\mathbf{u}\times\mathbf{B}). \tag{2.10}$$

Finally multiplying the Vlasov equation by $\frac{1}{2}m|\mathbf{v}|^2 = \frac{1}{2}m\mathbf{v}\cdot\mathbf{v}$ and integrating with respect to $\mathbf{v}$, we obtain

$$\frac{1}{2}m\frac{\partial}{\partial t}\int f(\mathbf{x},\mathbf{v},t)|\mathbf{v}|^2\,d\mathbf{v} + \frac{1}{2}m\nabla_x\cdot\int(|\mathbf{v}|^2\mathbf{v}) f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} \\ + \frac{1}{2}q\int |\mathbf{v}|^2\nabla_v\cdot[(\mathbf{E}(\mathbf{x},t)+\mathbf{v}\times\mathbf{B}(\mathbf{x},t)) f(\mathbf{x},\mathbf{v},t)]\,d\mathbf{v} = 0.$$

An integration by parts then yields

$$\int |\mathbf{v}|^2\nabla_v\cdot(\mathbf{E}(\mathbf{x},t)+\mathbf{v}\times\mathbf{B}(\mathbf{x},t)) f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v} \\ = -2\int \mathbf{v}\cdot[(\mathbf{E}(\mathbf{x},t)+\mathbf{v}\times\mathbf{B}(\mathbf{x},t)) f(\mathbf{x},\mathbf{v},t)]\,d\mathbf{v}.$$

Then, developing $\int f|\mathbf{v}-\mathbf{u}|^2\,d\mathbf{v}$ we get

$$\int f|\mathbf{v}-\mathbf{u}|^2\,d\mathbf{v} = \int f|\mathbf{v}|^2\,d\mathbf{v} - 2\mathbf{u}\cdot\int \mathbf{v} f\,d\mathbf{v} + |\mathbf{u}|^2\int f\,d\mathbf{v} = \int f|\mathbf{v}|^2\,d\mathbf{v} - n|\mathbf{u}|^2,$$

whence

$$\frac{\partial}{\partial t}(\frac{3}{2}p + \frac{1}{2}mn|\mathbf{u}|^2) + \nabla\cdot\,\mathbf{Q} = \mathbf{E}\cdot(qn\mathbf{u}). \tag{2.11}$$

We could continue to calculate moments of f, but we see that each new expression reveals a moment of higher order. So we need additional information to have as many unknowns as equations to solve these equations. This additional information is called a *closure relation.*

In our case, we will use as a closure relation the physical property that at thermodynamic equilibrium the distribution function approaches a Maxwellian distribution function that we will note $f_M(\mathbf{x},\mathbf{v},t)$ and that can be expressed as a function of the macroscopic quantities $n(\mathbf{x},t)$, $\mathbf{u}(\mathbf{x},t)$ and $T(\mathbf{x},t)$ which are the density, mean velocity and temperature of the charged fluid:

$$f_M(\mathbf{x},\mathbf{v},t) = \frac{n(\mathbf{x},t)}{(2\pi T(\mathbf{x},t)/m)^{3/2}} e^{-\frac{|\mathbf{v}-\mathbf{u}(\mathbf{x},t)|^2}{2T(\mathbf{x},t)/m}}.$$

We also introduce a classical quantity in plasma physics which is the thermal velocity of the particle species considered

$$v_{th} = \sqrt{\frac{T}{m}}.$$

It is easy to verify that the first three moments of the distribution function f_M are consistent with the definition of the macroscopic quantities n, $\mathbf{u}$ and T defined for an arbitrary distribution function. We have indeed performing each time the change of variable $\mathbf{w} = \frac{\mathbf{v}-\mathbf{u}}{v_{th}}$

$$\int f_M(\mathbf{x},\mathbf{v},t)\, d\mathbf{v} = n(\mathbf{x},t),$$
$$\int f_M(\mathbf{x},\mathbf{v},t)\mathbf{v}\, d\mathbf{v} = n(\mathbf{x},t)\mathbf{u}(\mathbf{x},t),$$
$$\int f_M(\mathbf{x},\mathbf{v},t)|\mathbf{v}-\mathbf{u}|^2\, d\mathbf{v} = 3n(\mathbf{x},t)T(\mathbf{x},t)/m.$$

On the other hand, replacing f by f_M in the definitions of the pressure tensor $\mathbb{P}$ and the energy flux $\mathbf{Q}$, we can express these terms also in function of n, $\mathbf{u}$ and T which enables us to obtain a closed system in these three unknowns as opposed to the case of an arbitrary distribution function f. Indeed, we first notice that, denoting by w_i the i^{th} component of $\mathbf{w}$,

$$\int w_i w_j e^{-\frac{|\mathbf{w}|^2}{2}}\, d\mathbf{w} = \begin{cases} 0 \text{ if } i \neq j, \\ \int e^{-\frac{|\mathbf{w}|^2}{2}}\, d\mathbf{w} \text{ if } i = j. \end{cases}$$

It follows that the pressure tensor associated to the Maxwellian is

$$\mathbb{P} = m\frac{n}{(2\pi T/m)^{3/2}} \int e^{-\frac{|\mathbf{v}-\mathbf{u}|^2}{2T/m}} (\mathbf{v}-\mathbf{u}) \otimes (\mathbf{v}-\mathbf{u})\, d\mathbf{v},$$

and so, thanks to our previous computation, the off diagonal terms of $\mathbb{P}$ vanish, and by the change of variable $\mathbf{w} = \frac{\mathbf{v}-\mathbf{u}}{v_{th}}$, we get for the diagonal terms

$$\mathbb{P}_{ii} = m\frac{n}{(2\pi)^{3/2}}\frac{T}{m} \int e^{-\frac{\mathbf{w}^2}{2}} w_i^2\, d\mathbf{v} = nT.$$

It follows that $\mathbb{P} = nT\mathbb{I} = p\mathbb{I}$ where $\mathbb{I}$ is the 3×3 identity matrix. It now remains to compute in the same way $\mathbf{Q}$ as a function of n, $\mathbf{u}$ and T for the Maxwellian with the same change of variables, which yields

$$\begin{aligned}\mathbf{Q} &= \frac{m}{2}\frac{n}{(2\pi T/m)^{3/2}}\int e^{-\frac{|\mathbf{v}-\mathbf{u}|^2}{2T/m}}|\mathbf{v}|^2\mathbf{v}(\mathbf{x},t))\,d\mathbf{v},\\ &= \frac{m}{2}\frac{n}{(2\pi)^{3/2}}\int e^{-\frac{\mathbf{w}^2}{2}}(v_{th}\mathbf{w}+\mathbf{u})^2(v_{th}\mathbf{w}+\mathbf{u})\,d\mathbf{w},\\ &= \frac{m}{2}\frac{n}{(2\pi)^{3/2}}\int e^{-\frac{\mathbf{w}^2}{2}}(v_{th}^2\mathbf{w}^2\mathbf{u}+2v_{th}^2\mathbf{u}\cdot\mathbf{w}\,\mathbf{w}+|\mathbf{u}|^2\,\mathbf{u})\,d\mathbf{w},\\ &= \frac{m}{2}n(3\frac{T}{m}\mathbf{u}+2\frac{T}{m}\mathbf{u}+|\mathbf{u}|^2\mathbf{u}),\end{aligned}$$

as the odd moments in $\mathbf{w}$ vanish. We finally get

$$\mathbf{Q} = \frac{5}{2}nT\mathbf{u}+\frac{m}{2}n|\mathbf{u}|^2\mathbf{u} = \frac{5}{2}p\mathbf{u}+\frac{m}{2}n|\mathbf{u}|^2\mathbf{u}.$$

Then, plugging the expressions of $\mathbb{P}$ and of $\mathbf{Q}$ in (2.9)-(2.10)-(2.11) we obtain the fluid equations for one species of particles of a plasma:

$$\frac{\partial n}{\partial t}+\nabla_x\cdot(n\mathbf{u}) = 0 \tag{2.12}$$

$$m\frac{\partial}{\partial t}(n\mathbf{u})+m\nabla\cdot(n\mathbf{u}\otimes\mathbf{u})+\nabla p = qn(\mathbf{E}+\mathbf{u}\times\mathbf{B}) \tag{2.13}$$

$$\frac{\partial}{\partial t}(\frac{3}{2}p+\frac{1}{2}mn|\mathbf{u}|^2)+\nabla\cdot\,(\frac{5}{2}p\mathbf{u}+\frac{m}{2}n|\mathbf{u}|^2\mathbf{u}) = \mathbf{E}\cdot(qn\mathbf{u}), \tag{2.14}$$

which corresponds in three dimensions to a system of 5 scalar equation with 5 scalar unknowns which are the density n, the three components of the mean velocity $\mathbf{u}$ and the scalar pressure p. These equations need of course to be coupled to Maxwell's equations for the computation of the self-consistent electromagnetic field with, in the case of only one particle species $\rho = q\,n$ and $\mathbf{J} = q\,n\mathbf{u}$. Let us also point out that an approximation often used in plasma physics is that of a cold plasma, for which $T = 0$ and thus $p = 0$. Only the first two equations are needed in this case.

2.5 Some application specific approximations

2.5.1 The paraxial approximation

The paraxial model is often used in Accelerator Physics for analysing the propagation of beams possessing an optical axis, which is assumed to be a straight line. For a physicist's derivation of this model one can refer to the recent book by Davidson and Qin [45]. A mathematically rigorous derivation of this model as an approximation of the steady-state Vlasov-Maxwell equations was proposed by P. Degond and P.-A. Raviart [49].

Let us recall the derivation of the model as presented in [61]. Starting from the steady-state relativistic Vlasov equation, which reads

$$\mathbf{v} \cdot \nabla_{\mathbf{x}} \tilde{f} + q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{p}} \tilde{f} = 0, \tag{2.15}$$

where q is the charge and m is the mass of one particle, c the velocity of light in free space, $\mathbf{v} = \mathbf{p}/(\gamma m)$, $\beta = |\mathbf{v}|/c$, $\gamma = (1-\beta^2)^{-\frac{1}{2}}$, $\tilde{f}(\mathbf{x}, \mathbf{p})$ represents the distribution function of one species of particles (ions, electrons), depending on position $\mathbf{x} \in \mathbb{R}^3$ and momentum $\mathbf{p} \in \mathbb{R}^3$. From the distribution function $\tilde{f}$, we compute the charge and current densities

$$\rho(\mathbf{x}) = q \int_{\mathbb{R}^3} \tilde{f}(\mathbf{x}, \mathbf{p}) d\mathbf{p}, \quad \mathbf{J}(\mathbf{x}) = q \int_{\mathbb{R}^3} \mathbf{v}\, \tilde{f}(\mathbf{x}, \mathbf{p}) d\mathbf{p} \tag{2.16}$$

and the electromagnetic fields $\mathbf{E}$ (electric field) and $\mathbf{B}$ (magnetic field) are given by the steady-state Maxwell equations

$$\nabla \times \mathbf{B} = -\mu_0 \mathbf{J}, \tag{2.17}$$

$$\nabla \times \mathbf{E} = 0, \tag{2.18}$$

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}, \tag{2.19}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{2.20}$$

In the case of particle beams, we can derive a simplified model based on the following assumption which are often satisfied in physical problems involving particle beams.

- The beam is steady-state: All partial derivatives with respect to time vanish.
- The beam is sufficiently long so that longitudinal self-consistent forces can be neglected.
- The beam is propagating at constant velocity v_b along the propagation axis z.
- Electromagnetic self-forces are included.
- $p_x, p_y << p_b$ where $p_b = \gamma m v_b$ is the beam momentum. It follows in particular that

$$\beta \approx \beta_b = (v_b/c)^2, \qquad \gamma \approx \gamma_b = (1-\beta_b^2)^{-1/2}.$$

- The beam is thin: the transverse dimensions of the beam are small compared to the characteristic longitudinal dimension.

The paraxial model of approximation of Vlasov-Maxwell's equations is obtained by retaining only the first terms in the asymptotic expansion of the distribution function and the electromagnetic fields with respect to $\eta = l/L$, where l denotes the transverse characteristic length and L is the longitudinal characteristic length [48].

$$\eta = l/L \ll 1.$$

Moreover, for simplicity we will neglect the variation with respect to the longitudinal mean velocity v_b.

Let us now introduce

$$f = f(z, \mathbf{x}, \mathbf{v}),\ \mathbf{x} = (x, y),\ \mathbf{v} = (v_x, v_y),\ \Phi = \Phi(\mathbf{x}, z),\ \mathbf{B} = (B_x(\mathbf{x}, z), B_y(\mathbf{x}, z))$$

where the new distribution function f is linked to the solution $\tilde{f}$ of the original Vlasov equation (2.15) by $\tilde{f}(x, y, z, p_x, p_y, p_z) = f(z, \mathbf{x}, \mathbf{v})\delta(p_z - p_b)$. Then making the assumptions above f is a solution (in the sense of distributions) of

$$v_b \frac{\partial f}{\partial z} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f + \frac{q}{\gamma_b m} \mathbf{F} \cdot \nabla_{\mathbf{v}} f = 0, \tag{2.21}$$

$$\mathbf{E} = -\nabla_{\mathbf{x}} \Phi, \quad -\Delta_{\mathbf{x}} \Phi = q\, n/\epsilon_0, \quad n = \int_{\mathbb{R}^2} f d\mathbf{v}, \tag{2.22}$$

$$\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \mu_0\, q\, v_b\, n, \tag{2.23}$$

$$\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} = -\frac{dB_z}{dz}. \tag{2.24}$$

and the transverse Lorentz force $\mathbf{F} = q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B})_\perp = (F_x, F_y)$ is given by

$$F_x = -\frac{\partial \Phi}{\partial x} - v_b\, B_y + v_y\, B_z, \quad F_y = -\frac{\partial \Phi}{\partial y} + v_b\, B_x - v_x\, B_z.$$

Since Maxwell's equations are linear, we can split the transverse fields into their external and self-consistent parts :

$$\mathbf{E} = \mathbf{E}^e + \mathbf{E}^s, \quad \mathbf{B} = \mathbf{B}^e + \mathbf{B}^s.$$

which are respectively of the forms

$$\mathbf{E}^e = -\nabla_{\mathbf{x}} \Phi^e, \quad \mathbf{B}^e = -\nabla_{\mathbf{x}} \chi^e, \tag{2.25}$$

and

$$\mathbf{E}^s = -\nabla_{\mathbf{x}} \Phi^s, \quad \mathbf{B}^s = \mathbf{curl}_{\mathbf{x}} \psi^s = (\partial_y \psi^s, -\partial_x \psi^s). \tag{2.26}$$

On the one hand, from the equations (2.23)–(2.24), we check that the functions ϕ^e and χ^e satisfy

$$-\Delta_{\mathbf{x}} \Phi^e = 0,\ -\Delta_{\mathbf{x}} \chi^e = -\frac{dB_z}{dz}, \tag{2.27}$$

where B_z is an external magnetic field.

On the other hand, from the equations (2.22), (2.23), (2.24) the self-consistent forces satisfy

$$-\Delta_{\mathbf{x}} \Phi^s = q\, n/\epsilon_0, \qquad -\Delta_{\mathbf{x}} \psi^s = \mu_0\, v_b\, q\, n = q \frac{v_b}{\epsilon_0 c^2}\, n. \tag{2.28}$$

Then, we have

$$\psi^s = \frac{v_b}{c^2}\Phi^s$$

and the self-consistent force field is given by

$$F_x^s = \frac{q}{\gamma_b m}(-\frac{\partial \Phi^s}{\partial x} - v_b\, B_y^s) = -\frac{q}{\gamma_b m}(1-\beta_b^2)\frac{\partial \Phi^s}{\partial x} = -\frac{q}{\gamma_b^3 m}\frac{\partial \Phi^s}{\partial x}, \quad (2.29)$$

$$F_y^s = \frac{q}{\gamma_b m}(-\frac{\partial \Phi^s}{\partial y} + v_b\, B_x^s) = -\frac{q}{\gamma_b m}(1-\beta_b^2)\frac{\partial \Phi^s}{\partial y} = -\frac{q}{\gamma_b^3 m}\frac{\partial \Phi^s}{\partial y}, \quad (2.30)$$

where $\beta = v_b/c$.

For the external forces, we consider the three different types of external focusing forces which are mostly used in accelerators for modeling purposes:

1. Uniform focusing by a uniform electric field of the form

$$\mathbf{E}(\mathbf{x}) = -\frac{\gamma_b m}{q}\omega_0^2\,(x\,\mathbf{e}_x + y\,\mathbf{e}_y).$$

2. Periodic focusing by a magnetic field of the form

$$\mathbf{B}(\mathbf{x}) = B(z)\,\mathbf{e}_z - \frac{1}{2}B'(z)\,(x\,\mathbf{e}_x + y\,\mathbf{e}_y),$$

   where the longitudinal component of the magnetic field $B(z)$ is given and satisfies the periodicity condition $B(z+S) = B(z)$.
3. Alternating gradient focusing:
   - either by a magnetic field of the form

$$\mathbf{B}(\mathbf{x}) = B'(z)\,(y\,\mathbf{e}_x + x\,\mathbf{e}_y),$$

     which corresponds to a potential

$$\psi^e(\mathbf{x}) = -\frac{1}{2}B'(z)(x^2 - y^2),$$

   - or by an electric field of the form

$$\mathbf{E}(\mathbf{x}) = E'(z)\,(x\,\mathbf{e}_x - y\,\mathbf{e}_y),$$

     which corresponds to a potential

$$\Phi^e(\mathbf{x}) = -\frac{1}{2}E'(z)(x^2 - y^2).$$

The Vlasov equation (2.21) we consider, can be interpreted, dividing all the terms by the strictly positive velocity v_b, as a transverse Vlasov equation where z plays the role of time. We then get the paraxial model

$$\frac{\partial f}{\partial z} + \frac{\mathbf{v}}{v_b}\cdot\nabla_{\mathbf{x}} f + \frac{q}{\gamma_b m v_b}\left(-\frac{1}{\gamma_b^2}\nabla\Phi^s + \mathbf{E}^e + (\mathbf{v}, v_b)^T \times \mathbf{B}^e\right)\cdot\nabla_{\mathbf{v}} f = 0, \quad (2.31)$$

coupled with the Poisson equation

$$-\Delta_{\mathbf{x}}\Phi^s = \frac{q}{\epsilon_0}\int_{\mathbb{R}^2} f(z,\mathbf{x},\mathbf{v})\,d\mathbf{v}. \tag{2.32}$$

The characteristic curves associated to this Vlasov equation are given by

$$x' = \frac{v_x}{v_b}, \tag{2.33}$$

$$y' = \frac{v_y}{v_b}, \tag{2.34}$$

$$v_x' = -\frac{q}{\gamma_b^3 m v_b}\frac{\partial \Phi^s}{\partial x} + \frac{q}{\gamma_b m v_b}\left(E_x^e + v_y\,B_z^e - v_b\,B_y^e\right) \tag{2.35}$$

$$v_y' = -\frac{q}{\gamma_b^3 m v_b}\frac{\partial \Phi^s}{\partial y} + \frac{q}{\gamma_b m v_b}\left(E_y^e - v_x\,B_z^e + v_b\,B_x^e\right), \tag{2.36}$$

where the notation $'$ corresponds to the derivative with respect the the longitudinal variable z.

The paraxial model is much simpler than the full Vlasov-Maxwell model. On the one hand, one replaces the stationary Vlasov equation by the paraxial Vlasov equation (2.21) where the longitudinal co-ordinates z plays the role of a time variable and which can be solved numerically by a marching procedure. On the other hand the stationary Maxwell's equations are replaced by the two dimensional Poisson equations (2.28) where z only acts like a parameter.

2.5.2 The gyrokinetic approximation

In the large external magnetic field characteristic of tokamak plasmas, the Vlasov-Poisson equations are reduced to the so-called gyrokinetic approximation, where one on the velocity components is removed as well as the stiffness in time induce by the fast rotation of particles around these field lines.

We consider a given equilibrium magnetic field $\mathbf{B} = B_0\mathbf{e}_\varphi + (\nabla\psi)\times\mathbf{e}_\varphi$, where φ is the toroidal angle in a tokamak and $\mathbf{e}_\varphi$ the unit vector in the toroidal direction. We suppose that ψ does not depend on φ. Its norm is denoted by B and $\mathbf{b} = \mathbf{B}/B$ the unit vector in the direction of the magnetic field. Denoting by

$$\mathbf{B}^* = \mathbf{B} + v_\parallel \nabla\times\mathbf{b}, \quad B_\parallel^* = \mathbf{b}\cdot\mathbf{B}^* = B + v_\parallel \nabla\times\mathbf{b}\cdot\mathbf{b},$$

the gyrokinetic Vlasov equation in cartesian coordinates used in classical simulations is recalled in the review paper by Garbet et. al. [63] or Grandgirard and Sarazin [65]. In the electrostatic case it reduces to

$$\frac{\partial f}{\partial t} + \frac{d\mathbf{X}}{dt}\cdot\nabla_x f + \frac{dV_\parallel}{dt}\frac{\partial f}{\partial v_\parallel} = 0,$$

with

$$B_\parallel^* \frac{d\mathbf{X}}{dt} = V_\parallel \mathbf{B}^* + \mathbf{b} \times (\mu \nabla B + \nabla J(\phi)), \tag{2.37}$$

$$B_\parallel^* \frac{dV_\parallel}{dt} = -\mathbf{B}^* \cdot (\mu \nabla B + \nabla J(\phi)), \tag{2.38}$$

where $J(\phi)$ is the gyro-average operator applied to the electrostatic potential ϕ.

The gyrokinetic Vlasov equation is coupled with the gyrokinetic Poisson equation, which also relies on a quasi-neutrality approximation:

$$-\nabla_\perp \cdot (\frac{n_0(r)}{B\omega_c} \nabla_\perp \phi) + \frac{e\, n_0(r)}{T_e(r)} (\phi - \lambda \langle \phi \rangle) = \int J(f)\, \mathrm{d}v_\parallel d\mu - n_0,$$

where $\langle \phi \rangle$ is the average of ϕ on a magnetic flux surface.

The gyrokinetic approximation satisfies the following conservation properties:

- Conservation of mass: We have the relations

$$\begin{aligned} \nabla \cdot (B^* \frac{d\mathbf{X}}{dt}) &= \nabla \cdot (\mathbf{b} \times \nabla J(\phi)) + \frac{1}{q} \nabla \cdot (\mathbf{b} \times \mu \nabla B) \\ &= \nabla J(\phi) \cdot \nabla \times \mathbf{b} + \frac{\mu}{q} \nabla B \cdot \nabla \times \mathbf{b} \end{aligned}$$

  On the other hand

$$\frac{\partial}{\partial v_\parallel} (B^* \frac{dV_\parallel}{dt}) = -\nabla \times \mathbf{b} \cdot (\frac{\mu}{q} \nabla B + \nabla J(\phi)).$$

  Hence the phase space divergence vanishes, from which conservativity follows.
- Conservation of energy

$$\frac{d}{dt} \left(\int m(\mu B + v_\parallel^2) f(t, \mathbf{x}, v_\parallel, \mu) B^* \, \mathrm{d}\mathbf{x}\, \mathrm{d}v_\parallel d\mu + \int \phi \tilde{J}(n)\, \mathrm{d}\mathbf{x} \right) = 0.$$

2.5.3 The guiding-center model

When considering only the restriction of the gyrokinetic model to the perpendicular direction, *i.e.* in the poloidal plane, for a constant and uniform magnetic field we get the guiding-center approximation. Denoting the density by $f = f(t, x, y)$, which is no longer dependent on velocity the model reads

$$\frac{\partial f}{\partial t} + E^\perp(x, y) \cdot \nabla f = 0, \tag{2.39}$$

coupled self-consistently to Poisson's equation for the electric field which derives from a potential $\Phi = \Phi(x, y)$ that satisfies the Poisson equation

$$-\Delta\Phi(t,x,y) = f(t,x,y), \quad E(t,x,y) = -\nabla\Phi(t,x,y). \tag{2.40}$$

In equation (2.39), the advection term $E^\perp = (E_y, -E_x)$ depends on (x, y) and the time-splitting that we will see in Chapter 5 cannot be simply applied like in the Vlasov-Poisson case. Hence, this simple model appears to be interesting in order to test numerical methods.

The guiding-center model (2.39)-(2.40) also presents conserved quantities as the total number of particles and L^2 norm of f (energy) and E (enstrophy)

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(t,x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{\mathrm{d}}{\mathrm{d}t}\int f^2(t,x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{\mathrm{d}}{\mathrm{d}t}\int E^2(t,x)\,\mathrm{d}x = 0. \tag{2.41}$$

2.6 Expressions of the Maxwell equations

2.6.1 The 3D Maxwell equations

The general expression for the Maxwell equations reads

$$-\frac{\partial \mathbf{D}}{\partial t} + \nabla \times \mathbf{H} = \mathbf{J}, \tag{2.42}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0, \tag{2.43}$$

$$\nabla \cdot \mathbf{D} = \rho, \tag{2.44}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{2.45}$$

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{2.46}$$

$$\mathbf{B} = \mu \mathbf{H}. \tag{2.47}$$

Initial and boundary conditions are needed in addition to fully determine the solution.

The last two relations are called the constitutive laws and permittivity ϵ and the permeability μ depend on the material. They can be discontinuous if several materials are considered. In vacuum $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$ are constants and they verify $\varepsilon_0\mu_0 c^2 = 0$ where c is the speed of light. Then $\mathbf{D}$ and $\mathbf{H}$ are generally eliminated of the system.

Note that taking the divergence of (2.42) yields

$$\frac{\partial \nabla \cdot \mathbf{D}}{\partial t} = -\nabla \cdot \mathbf{J} = \frac{\partial \rho}{\partial t}$$

using the continuity equation $\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$. Hence if (2.44) is satisfied at time $t = 0$ it will be satisfied at all times. In the same way if $\nabla \cdot \mathbf{B} = 0$ at the initial time it will remain so for all times.

2.6.2 The 2D Maxwell equations

We consider the Maxwell equations in vacuum on a two dimensional domain Ω on which the fields are independent of the z variable. Then the electromagnetic field obeys to two sets of decoupled equations, the first of which involving the (E_x, E_y, B_z) components (TE mode) and the second involving the (E_z, B_x, B_y) components (TM mode). We present only the first system, the other can be dealt with in a similar manner. This system reads

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\, B = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{2.48}$$

$$\frac{\partial B}{\partial t} + \operatorname{curl} \mathbf{E} = 0, \tag{2.49}$$

$$\operatorname{div} \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{2.50}$$

where $\mathbf{E} = (E_x, E_y)^T$, $B = B_z$, $\mathbf{curl}\, B_z = (\partial_y B_z, -\partial_x B_z)^T$, $\operatorname{curl} \mathbf{E} = \partial_x E_y - \partial_y E_x$, and $\operatorname{div} \mathbf{E} = \partial_x E_x + \partial_y E_y$.

2.6.3 The 1D Maxwell equations

The system can be further decoupled in 1D, assuming that the fields only depend on x. Then (2.48) becomes

$$\frac{\partial E_x}{\partial t} = -\frac{1}{\varepsilon_0} J_x, \tag{2.51}$$

$$\frac{\partial E_y}{\partial t} + c^2 \frac{\partial B_z}{\partial x} = -\frac{1}{\varepsilon_0} J_y, \tag{2.52}$$

$$\frac{\partial B_z}{\partial t} + \frac{\partial E_y}{\partial x} = 0, \tag{2.53}$$

$$\frac{\partial E_x}{\partial x} = \frac{\rho}{\varepsilon_0}. \tag{2.54}$$

Note that here we decouple completely the propagative part of the electric field which is in 1D only E_y from its "static" part E_x. Components E_y and B_z are coupled by equations (2.52) and (2.53) and E_x is given either by the first component of the Ampère equation (2.51) or by Gauss's law (2.54), which are equivalent provided the initial condition satisfies Gauss's law and the 1D continuity equation $\frac{\partial \rho}{\partial t} + \frac{\partial J_x}{\partial x} = 0$, which are compatibility conditions.

3

Some theory on Vlasov systems

3.1 The linear Vlasov equation

The Vlasov equation is a linear scalar hyperbolic partial differential equation when $\mathbf{E}$ and $\mathbf{B}$ are assumed to be known independently of f. Setting all constants to one the Vlasov can then be written

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_x f + (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_v f = 0, \tag{3.1}$$

where $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$ are given fields. Setting

$$\mathbf{A}(\mathbf{x}, \mathbf{v}, t) = \begin{pmatrix} \mathbf{v} \\ \mathbf{E} + \mathbf{v} \times \mathbf{B} \end{pmatrix},$$

equation (3.1) becomes

$$\frac{\partial f}{\partial t} + \mathbf{A} \cdot \nabla_{(x,v)} f = 0. \tag{3.2}$$

Hence it is a linear advection equation in phase space. Moreover

$$\begin{aligned} \nabla_{(x,v)} \cdot \mathbf{A} &= \nabla_x \cdot \mathbf{v} + \nabla_v \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \\ &= 0 + \frac{\partial}{\partial v_1}(E_1 + v_2 B_3 - v_3 B_2) + \frac{\partial}{\partial v_2}(E_2 + v_3 B_1 - v_1 B_3) \\ &\qquad + \frac{\partial}{\partial v_2}(E_3 + v_1 B_2 - v_2 B_1) \\ &= 0. \end{aligned}$$

The Vlasov equation can be written in a conservative form

$$\frac{\partial f}{\partial t} + \nabla_{(x,v)} \cdot (\mathbf{A} f) = 0, \tag{3.3}$$

as $\nabla_{(x,v)} \cdot (\mathbf{A} f) = \mathbf{A} \cdot \nabla_{(x,v)} f + f \nabla_{(x,v)} \cdot \mathbf{A}$.

Remark 3 *These properties do not rely on the fact that* $\mathbf{E}$ *and* $\mathbf{B}$ *are given independently of* f *and are also valid in the non linear case.*

The Vlasov equation can thus be written as a classical advection equation

$$\frac{\partial f}{\partial t} + \mathbf{A} \cdot \nabla f = 0, \tag{3.4}$$

with $f : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$ and $\mathbf{A} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$.

Consider now for $s \in \mathbb{R}^+$ given, the differential system

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}(\mathbf{X}, t), \tag{3.5}$$

$$\mathbf{X}(s) = \mathbf{x}, \tag{3.6}$$

which is naturally associated to the advection equation (3.4).

Definition 1 *The solutions of the system* (3.5) *are called characteristics of the linear advection equation* (3.4)*. We denote by* $\mathbf{X}(t; s, \mathbf{x})$ *the solution of* (3.5) - (3.6).

Let us recall the classical theorem of the theory of ordinary differential equations (ODE) which gives existence and uniqueness of the solution of (3.5)-(3.6). The proof can be found in [3] for example.

Theorem 1 *Assume that* $\mathbf{A} \in C^{k-1}(\mathbb{R}^d \times [0,T])$, $\nabla \mathbf{A} \in C^{k-1}(\mathbb{R}^d \times [0,T])$ *for* $k \geq 1$ *and that*

$$|\mathbf{A}(\mathbf{x}, t)| \leq \kappa(1 + |\mathbf{x}|) \quad \forall t \in [0,T] \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

Then for all $s \in [0,T]$ *and* $\mathbf{x} \in \mathbb{R}^d$*, there exists a unique solution* $\mathbf{X} \in C^k([0,T]_t \times [0,T]_s \times \mathbb{R}^d_x)$ *of* (3.5) - (3.6)*.*

Proposition 4 *Under the hypotheses of the previous theorem we have the following properties:*

(i) $\forall t_1, t_2, t_3 \in [0,T]$ *and* $\forall \mathbf{x} \in \mathbb{R}^d$

$$\mathbf{X}(t_3; t_2, \mathbf{X}(t_2; t_1, \mathbf{x})) = \mathbf{X}(t_3; t_1, \mathbf{x}).$$

(ii)$\forall (t,s) \in [0,T]^2$*, the application* $\mathbf{x} \mapsto \mathbf{X}(t; s, \mathbf{x})$ *is a* C^1*- diffeomorphism of* $\mathbb{R}^d$ *of inverse* $\mathbf{y} \mapsto \mathbf{X}(s; t, \mathbf{y})$*.*

(iii)The jacobian $J(t; s, 1) = \det(\nabla_x \mathbf{X}(t; s, \mathbf{x}))$ *verifies*

$$\frac{\partial J}{\partial t} = (\nabla \cdot \mathbf{A})(t; \mathbf{X}(t; s, \mathbf{x}))J,$$

and $J > 0$*. In particular if* $\nabla \cdot \mathbf{A} = 0$*,* $J(t; s, 1) = J(s; s, 1) = \det \mathbb{I}_d = 1$*, where* $\mathbb{I}_d$ *is the identity matrix of order* d*.*

Proof. (i) The points $\mathbf{x} = \mathbf{X}(t_1; t_1, \mathbf{x})$, $\mathbf{X}(t_2; t_1, \mathbf{x})$, $\mathbf{X}(t_3; t_1, \mathbf{x})$ are on the same characteristic curve. This curve is characterized by the initial condition $\mathbf{X}(t_1) = \mathbf{x}$. So, taking any of these points as initial condition at the corresponding time, we get the same solution of (3.5)-(3.6). We have in particular $\mathbf{X}(t_3; t_2, \mathbf{X}(t_2; t_1, \mathbf{x})) = \mathbf{X}(t_3; t_1, \mathbf{x})$.

(ii) Taking $t_1 = t_3$ in the equality (i) we have

$$\mathbf{X}(t_3; t_2, \mathbf{X}(t_2; t_3, \mathbf{x})) = \mathbf{X}(t_3; t_3, \mathbf{x}) = \mathbf{x}.$$

Hence $\mathbf{X}(t_3; t_2, .)$ is the inverse of $\mathbf{X}(t_2; t_3, .)$ (we denote by $g(.)$ the function $x \mapsto g(x)$) and both applications are of class C^1 because of the previous theorem.

(iii) Let

$$J(t; s, 1) = \det(\nabla_x \mathbf{X}(t; s, \mathbf{x})) = \det((\frac{\partial \mathbf{X}_i(t; s, \mathbf{x})}{\partial x_j}))_{1 \le i,j \le d}.$$

But $\mathbf{X}$ verifies $\frac{d\mathbf{X}}{dt} = \mathbf{A}(\mathbf{X}(t), t)$. So we get in particular taking the ith line of this equality $\frac{dX_i}{dt} = A_i(\mathbf{X}(t), t)$. And taking the gradient we get

$$\frac{d}{dt}\nabla X_i = \sum_{k=1}^{d} \frac{\partial A_i}{\partial x_k} \nabla X_k.$$

For a $d \times d$ matrix M the determinant of M is a d-linear alternated form taking as arguments the columns of M. So, denoting by $(., \ldots, .)$ this alternated d-linear form, we can write $\det M = (M_1, \ldots, M_d)$ where M_j is the jth column of M. Using this notation in our case, we get

$$\begin{aligned}
\frac{\partial J}{\partial t}(t; s, 1) &= \frac{\partial}{\partial t} \det(\nabla_x \mathbf{X}(t; s, \mathbf{x})) \\
&= (\frac{\partial \nabla X_1}{\partial t}, \nabla X_2, \ldots, \nabla X_d) + \cdots + (\nabla X_1, \nabla X_2, \ldots, \frac{\partial \nabla X_d}{\partial t}) \\
&= (\sum_{k=1}^{d} \frac{\partial A_1}{\partial x_k} \nabla X_k, \nabla X_2, \ldots, \nabla X_d) + \ldots \\
&\qquad + (\nabla X_1, \nabla X_2, \ldots, \sum_{k=1}^{d} \frac{\partial A_d}{\partial x_k} \nabla X_k) \\
&= \frac{\partial A_1}{\partial x_1} J + \cdots + \frac{\partial A_d}{\partial x_d} J,
\end{aligned}$$

as $(., \ldots, .)$ is alternated and d-linear. Thus we have $\frac{\partial J}{\partial t}(t; s, 1) = (\nabla \cdot \mathbf{A})J$. On the other hand $\nabla_x \mathbf{X}(s; s, \mathbf{x}) = \nabla_x \mathbf{x} = \mathbb{I}_d$ and so $J(s; s, 1) = \det \mathbb{I}_d = 1$. J is a solution of the differential equation

$$\frac{dJ}{dt} = (\nabla \cdot \mathbf{A})\, J, \quad J(s) = 1,$$

which admits as the unique solution $J(t) = e^{\int_s^t \nabla \cdot \mathbf{A}\, dt} > 0$ and in particular, if $\nabla \cdot \mathbf{A} = 0$, we have $J(t; s, 1) = 1$ for all t.

After having highlighted the properties of the characteristics, we can now express the solution of the linear advection equation (3.4) using the characteristics.

Theorem 2 *Let $f_0 \in C^1(\mathbb{R}^d)$ and $\mathbf{A}$ a vector field verifying the hypotheses of the previous theorem. Then there exists a unique solution of the linear advection equation* (3.4) *associated to the initial condition $f(\mathbf{x},0) = f_0(\mathbf{x})$. It is given by*

$$f(\mathbf{x},t) = f_0(\mathbf{X}(0;t,\mathbf{x})), \tag{3.7}$$

where $\mathbf{X}$ represent the characteristics associated to $\mathbf{A}$.

Proof. The function f given by (3.7) is C^1 as f_0 and $\mathbf{X}$ are, and $\mathbf{X}$ is defined uniquely. Let's verify that f is a solution of (3.4) and that it verifies the initial condition. We first have using formula (3.7)

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

Then

$$\frac{\partial f}{\partial t}(\mathbf{x},t) = \frac{\partial X}{\partial s}(0;t,\mathbf{x}) \cdot \nabla f_0(0;t,\mathbf{x}),$$

and

$$\begin{aligned}\nabla_x f(\mathbf{x},t) &= \nabla_x (f_0(\mathbf{X}(0;t,\mathbf{x})) \\ &= \sum_{k=1}^{d} \frac{\partial f_0}{\partial x_k} \nabla_x X_k(0;t,\mathbf{x})), \\ &= \nabla_x \mathbf{X}(0;t,\mathbf{x})^T \nabla_x f_0(\mathbf{X}(0;t,\mathbf{x})),\end{aligned}$$

in the sense of a matrix vector product with the jacobian matrix

$$\nabla_x \mathbf{X}(0;t,\mathbf{x}) = ((\frac{\partial X_k}{\partial x_l}(0;t,\mathbf{x})))_{1\le k,l\le d}.$$

We then get

$$\begin{aligned}(\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla_x f)(\mathbf{x},t) = \frac{\partial X}{\partial s}(0;t,\mathbf{x}) \cdot \nabla f_0(0;t,\mathbf{x}) \\ + \mathbf{A}(\mathbf{x},t) \cdot \left(\nabla_x \mathbf{X}(0;t,\mathbf{x})^T \nabla_x f_0(\mathbf{X}(0;t,\mathbf{x}))\right).\end{aligned} \tag{3.8}$$

Because of the properties of the characteristics we also have that

$$\mathbf{X}(t;s,\mathbf{X}(s;r,\mathbf{x})) = \mathbf{X}(t;r,\mathbf{x})$$

and taking the derivative with respect to s, we get

$$\frac{\partial \mathbf{X}}{\partial s}(t;s,\mathbf{X}(s;r,\mathbf{x})) + \nabla_x \mathbf{X}(t;s,\mathbf{X}(s;r,\mathbf{x})) \frac{\partial \mathbf{X}}{\partial t}(s;r,\mathbf{x}) = 0.$$

But by definition of the characteristics $\frac{\partial \mathbf{X}}{\partial t}(s;r,\mathbf{x}) = \mathbf{A}(\mathbf{X}(s;r,\mathbf{x}),s)$ and as this equation is verified for all values of t,r,s and so in particular for $r=s$. It becomes in this case

$$\frac{\partial \mathbf{X}}{\partial s}(t;s,\mathbf{x}) + \nabla_x \mathbf{X}(t;s,\mathbf{x})\mathbf{A}(\mathbf{x},s) = 0.$$

Plugging this expression into (3.8) we obtain

$$\begin{aligned}(\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla_x f)(\mathbf{x},t) = -\nabla_x \mathbf{X}(0;t,\mathbf{x})\mathbf{A}(\mathbf{x},t))\cdot\nabla f_0(\mathbf{X}(0;t,\mathbf{x})) \\ + \mathbf{A}(\mathbf{x},t)\cdot\left(\nabla_x \mathbf{X}(0;t,\mathbf{x})^T \nabla_x f_0(\mathbf{X}(0;t,\mathbf{x}))\right).\end{aligned}$$

But for a matrix $M \in \mathcal{M}_d(\mathbb{R})$ and two vectors $\mathbf{u},\mathbf{v} \in \mathbb{R}^d$, on a $(M\mathbf{u})\cdot\mathbf{v} = \mathbf{u}^T M^T \mathbf{v} = \mathbf{u}\cdot(M^T\mathbf{v})$. Whence we get

$$\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla_x f = 0,$$

which means that f defined by (3.7) is solution of (3.4).

The problem being linear, if f_1 and f_2 are two solutions we have

$$\frac{\partial}{\partial t}(f_1 - f_2) + \mathbf{A}\cdot\nabla_x(f_1 - f_2) = 0,$$

and using the characteristics $\frac{d}{dt}(f_1 - f_2)(\mathbf{X}(t),t) = 0$. So if f_1 and f_2 verify the same initial condition, they are identical, which gives the uniqueness of the solution which is thus the function given by formula (3.7).

Examples

1. The free streaming equation
$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = 0.$$
The characteristics are solution of
$$\frac{dX}{dt} = V,\ \frac{dV}{dt} = 0.$$
This we have $V(t;s,x,v) = v$ and $X(t;s,x,v) = x + (t-s)v$ which gives us the solution
$$f(x,v,t) = f_0(x - vt, v).$$
2. Uniform focusing in a particle accelerator ($1D$ model). We then have $E(x,t) = -x$ and the Vlasov writes
$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - x\frac{\partial f}{\partial v} = 0.$$
$$\frac{dX}{dt} = V,\ \frac{dV}{dt} = -X.$$
Whence get $X(t;s,x,v) = x\cos(t-s) + v\sin(t-s)$ and $V(t;s,x,v) = -x\sin(t-s) + v\cos(t-s)$ form which we compute the solution
$$f(x,v,t) = f_0(x\cos t - v\sin t, x\sin t + v\cos t).$$

3.2 The Vlasov-Poisson system

3.2.1 The equations

The Poisson equation is obtained from the Maxwell equations, when the electric and magnetic fields are not, or only very little, time dependent. We then get the stationary Maxwell equations

$$\nabla \times \mathbf{B} = \mathbf{J}, \tag{3.9}$$

$$\nabla \times \mathbf{E} = 0, \tag{3.10}$$

$$\nabla \cdot \mathbf{E} = \rho, \tag{3.11}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{3.12}$$

In this case the electric and magnetic fields are decoupled, and in many cases, because $\mathbf{B}$ itself is small, or because its contribution in the Lorentz force $\mathbf{v} \times \mathbf{B}$ is small, we shall only need the electric field, which is given by equations (3.10) and (3.11). Equation (3.10) implies that $\mathbf{E}$ derives from a scalar potential $\mathbf{E} = -\nabla\phi$, and then (3.11) implies the Poisson equation

$$-\Delta\phi = \rho,$$

along with adequate boundary conditions.

We consider the dimensionless Vlasov-Poisson equation for one species with a neutralizing background

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_x f - \mathbf{E} \cdot \nabla_v f = 0, \tag{3.13}$$

$$-\Delta\phi = 1 - \rho, \quad \mathbf{E} = -\nabla\phi, \tag{3.14}$$

with

$$\rho(\mathbf{x}, t) = \int f(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v}.$$

The domain on which the system is posed is considered periodic in $\mathbf{x}$ and the whole space $\mathbb{R}^3$ in velocity.

We first notice that the Vlasov equation (3.13) can also be written in conservative form

$$\frac{\partial f}{\partial t} + \nabla_{\mathbf{x},\mathbf{v}} \cdot (\mathbf{F} f) = 0, \tag{3.15}$$

with $\mathbf{F} = (\mathbf{v}, -\mathbf{E})^T$ such that $\nabla_{\mathbf{x},\mathbf{v}} \cdot \mathbf{F} = 0$.

3.2.2 Conservation properties

The Vlasov-Poisson system has a number of conservation properties that need special attention when developing numerical methods. In principle it is beneficial to retain the exact invariants in numerical methods and when it is not possible to keep them all as is the case here, they can be use to monitor the validity of the simulation by checking that they are approximately conserved with good accuracy.

Proposition 5 *The Vlasov-Poisson system verifies the following conservation properties:*

- *Maximum principle*

$$0 \leq f(\mathbf{x}, \mathbf{v}, t) \leq \max_{(\mathbf{x},\mathbf{v})}(f_0(\mathbf{x}, \mathbf{v})). \tag{3.16}$$

- *Conservation of L^p, norms for p integer, $1 \leq p \leq \infty$*

$$\frac{d}{dt}\left(\int (f(\mathbf{x}, \mathbf{v}, t))^p \, d\mathbf{x}\, d\mathbf{v}\right) = 0 \tag{3.17}$$

- *Conservation of volume. For any volume V of phase space*

$$\int_V f(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{x}\, d\mathbf{v} = \int_{F^{-1}(V)} f_0(\mathbf{y}, \mathbf{u})\, d\mathbf{y}\, d\mathbf{u}. \tag{3.18}$$

- *Conservation of momentum*

$$\frac{d}{dt}\int \mathbf{v} f \, d\mathbf{x} d\mathbf{v} = \frac{d}{dt}\int \mathbf{J}\, d\mathbf{x} = 0. \tag{3.19}$$

- *Conservation of energy*

$$\frac{d}{dt}\left[\frac{1}{2}\int v^2 f \, d\mathbf{x} d\mathbf{v} + \frac{1}{2}\int E^2 \, d\mathbf{x}\right] = 0. \tag{3.20}$$

Proof. The system defining the associated characteristics writes

$$\frac{d\mathbf{X}}{dt} = \mathbf{V}(t), \tag{3.21}$$

$$\frac{d\mathbf{V}}{dt} = -\mathbf{E}(\mathbf{X}(t), t). \tag{3.22}$$

We denote by $(\mathbf{X}(t; \mathbf{x}, \mathbf{v}, s), \mathbf{V}(t; \mathbf{x}, \mathbf{v}, s))$, or more concisely $(\mathbf{X}(t), \mathbf{V}(t))$ when the dependency with respect to the initial conditions is not explicitly needed, the unique solution at time t of this system which takes the value $(\mathbf{x}, \mathbf{v})$ at time s.

Using (3.21)-(3.22), the Vlasov equation (3.13) can be expressed equivalently

$$\frac{d}{dt}(f(\mathbf{X}(t), \mathbf{V}(t)) = 0.$$

We thus have

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

From this expression, we deduce that f verifies a maximum principle which can be written as f_0 is non negative

$$0 \leq f(\mathbf{x}, \mathbf{v}, t) \leq \max_{(x,v)}(f_0(x, v)).$$

Multiplying the Vlasov equation by (3.13) par f^{p-1} and integrating on the whole phase-space we obtain

$$\frac{d}{dt}\left(\int (f(\mathbf{x},\mathbf{v},t))^p \, d\mathbf{x}\, d\mathbf{v}\right) = 0,$$

so that the L^p norms of f are conserved for all $p \in \mathbb{N}^*$. Let us notice that the L^∞ is also conserved thanks to the maximum principle (3.16).

Integrating on a arbitrary volume Vol of phase space and using that f is conserved along the characteristics we get

$$\begin{aligned}\int_{Vol} f(\mathbf{x},\mathbf{v},t)\, d\mathbf{x}\, d\mathbf{v} &= \int_{Vol} f(\mathbf{X}(t;\mathbf{x},\mathbf{v},t),\mathbf{V}(t;\mathbf{x},\mathbf{v},t),t)\, d\mathbf{x}\, d\mathbf{v} \\ &= \int_{Vol} f(\mathbf{X}(0;\mathbf{x},\mathbf{v},t),\mathbf{V}(0;\mathbf{x},\mathbf{v},t),0)\, d\mathbf{x}\, d\mathbf{v} \\ &= \int_{Vol} f_0(\mathbf{X}(0;\mathbf{x},\mathbf{v},t),\mathbf{V}(0;\mathbf{x},\mathbf{v},t)\, d\mathbf{x}\, d\mathbf{v},\end{aligned}$$

now making the change of variables $(\mathbf{y},\mathbf{u}) = \mathbf{F}(\mathbf{x},\mathbf{v})$ defined by $\mathbf{y} = \mathbf{X}(0;\mathbf{x},\mathbf{v},t), \mathbf{u} = \mathbf{V}(0;\mathbf{x},\mathbf{v},t)$ whose jacobian is equal to 1 thanks to proposition 4 as

$$\nabla_{(x,v)} \cdot \begin{pmatrix} \mathbf{v} \\ -\mathbf{E} \end{pmatrix} = 0,$$

we obtain

$$\int_{Vol} f(\mathbf{x},\mathbf{v},t)\, d\mathbf{x}\, d\mathbf{v} = \int_{\mathbf{F}^{-1}(Vol)} f_0(\mathbf{y},\mathbf{u})\, d\mathbf{y}\, d\mathbf{u}.$$

Let us now proceed to the conservation of momentum (or total current density). We shall use the following equality that is verified for any vector $\mathbf{u}$ depending on $\mathbf{x}$ in a periodic domain

$$\int (\nabla \times \mathbf{u}) \times \mathbf{u}\, d\mathbf{x} = -\int \left(\mathbf{u}(\nabla\cdot\mathbf{u}) + \frac{1}{2}\nabla u^2\right) d\mathbf{x} = -\int \mathbf{u}(\nabla\cdot\mathbf{u})\, d\mathbf{x}. \quad (3.23)$$

Let us notice in particular that taking $\mathbf{u} = \mathbf{E}$ in the previous equality with $\mathbf{E}$ solution of the Poisson equation (3.14), we get, as $\nabla \times \mathbf{E} = 0$ and $\nabla \cdot \mathbf{E} = -\Delta\phi = 1 - \rho$, that $\int \mathbf{E}(1-\rho)\, d\mathbf{x} = 0$. As moreover $\mathbf{E} = -\nabla\phi$ and as we integrate on a periodical domain $\int \mathbf{E}\, d\mathbf{x} = 0$. It results that

$$\int \mathbf{E}\rho\, d\mathbf{x} = 0. \quad (3.24)$$

Let us now introduce the Green formula on the divergence:

$$\int_\Omega \nabla\cdot\, \mathbf{F}q + \int_\Omega \mathbf{F}\cdot\nabla q = \int_{\partial\Omega} (\mathbf{F}\cdot\mathbf{n})\, q \quad \forall \mathbf{F} \in H(div,\Omega),\ q \in H^1(\Omega), \quad (3.25)$$

where classically $H^1(\Omega)$ is the subset of $L^2(\Omega)$ the square integrable functions, of the functions whose gradient is in $L^2(\Omega)$; and $H(div,\Omega)$ is the subset of $L^2(\Omega)$ of the functions whose divergence is in $L^2(\Omega)$.

Let's multiply the Vlasov equation (3.13) by $\mathbf{v}$ and integrate in $\mathbf{x}$ and in $\mathbf{v}$

$$\frac{d}{dt}\int \mathbf{v} f \, d\mathbf{x}d\mathbf{v} + \int \nabla_x \cdot (\mathbf{v} \otimes \mathbf{v} f) \, d\mathbf{x}d\mathbf{v} - \int \mathbf{v} \nabla_v \cdot (\mathbf{E} f) \, d\mathbf{x}d\mathbf{v} = 0.$$

The second integral vanishes as the domain is periodic in $\mathbf{x}$ and the Green formula on the divergence (3.25) gives for the last integral

$$-\int \mathbf{v} \nabla_v \cdot (\mathbf{E} f) \, d\mathbf{x}d\mathbf{v} = \int \mathbf{E} f \, d\mathbf{x}d\mathbf{v} = \int \mathbf{E}\rho \, d\mathbf{x} = 0,$$

using (3.24). It finally follows that

$$\frac{d}{dt}\int \mathbf{v} f \, d\mathbf{x}d\mathbf{v} = \frac{d}{dt}\int \mathbf{J} \, d\mathbf{x} = 0.$$

In order to obtain the energy conservation property, we start by multiplying the Vlasov equation by $\mathbf{v} \cdot \mathbf{v} = |\mathbf{v}|^2$ and we integrate on phase space

$$\frac{d}{dt}\int |\mathbf{v}|^2 f \, d\mathbf{x}d\mathbf{v} + \int \nabla_x \cdot (|\mathbf{v}|^2 \mathbf{v} f) \, d\mathbf{x}d\mathbf{v} - \int |\mathbf{v}|^2 \nabla_v \cdot (\mathbf{E} f) \, d\mathbf{x}d\mathbf{v} = 0.$$

As f is periodic in $\mathbf{x}$, we get, integrating in $\mathbf{x}$ that

$$\int \nabla_x \cdot (|\mathbf{v}|^2 \mathbf{v} f) \, d\mathbf{x}d\mathbf{v} = 0$$

and the Green formula on the divergence (3.25) yields

$$\int |\mathbf{v}|^2 \nabla_v \cdot \mathbf{E} \, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{v} \cdot (\mathbf{E} f) \, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{E} \cdot \mathbf{J} \, d\mathbf{x}.$$

So

$$\frac{d}{dt}\int |\mathbf{v}|^2 f \, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{E} \cdot \mathbf{J} \, d\mathbf{x} = 2\int \nabla\phi \cdot \mathbf{J} \, d\mathbf{x}. \tag{3.26}$$

On the other hand, integrating the Vlasov equation (3.13) with respect to $\mathbf{v}$, we get the charge conservation equation, generally called continuity equation: $\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$. Then, using again the Green formula (3.25), the Poisson equation (3.14) and the continuity equation, we obtain

$$\int \nabla\phi \cdot \mathbf{J} \, d\mathbf{x} = \int \phi \nabla \cdot \mathbf{J} \, d\mathbf{x} = -\int \phi \, \frac{\partial \rho}{\partial t} \, d\mathbf{x} = \int \phi \, \frac{\partial \Delta\phi}{\partial t} \, d\mathbf{x} = -\frac{1}{2}\frac{d}{dt}\int \nabla\phi \cdot \nabla\phi \, d\mathbf{x}.$$

And so, plugging this equation in (3.26) and using that $\mathbf{E} = -\nabla\phi$, we get the conservation of energy.

3.3 Solution of the linearised Vlasov-Poisson equations

In a number of physical conditions, one is interested in a small perturbation of the equilibrium plasma. Consider in particular the Vlasov-Poisson 1D near an equilibrium state by a given distribution function $f_0(x,v)$ for electrons and ions of a neutralizing substance such as the associated electric field is zero. We place on a periodic domain of period $L = 2\pi/k_0$ in x and the whole $\mathbb{R}$ in v.

Denoting by $-e$ the charge of an electron and its mass m, the steady state distribution function verifies the Vlasov-Poisson

$$\frac{\partial f^0}{\partial t} + v\frac{\partial f^0}{\partial x} - \frac{e}{m}E^0(x)\frac{\mathrm{d}f^0}{\mathrm{d}v} = 0 \tag{3.27}$$

$$\frac{\mathrm{d}E^0}{\mathrm{d}x} = \frac{e}{\epsilon_0}(n_0 - \int_{-\infty}^{+\infty} f^0(x,v)\,\mathrm{d}v) \tag{3.28}$$

with the initial condition $f^0(x,v,t) = f_0(x,v)$ and where

$$n_0 = \frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f^0(x,v)\,\mathrm{d}x\,\mathrm{d}v$$

is the constant background density of neutralising ions. If f^0 is a stationary solution $\frac{\partial f^0}{\partial t} = 0$ and if moreover $E^0 = 0$, the Vlasov equation reduces to $v\frac{\partial f^0}{\partial x} = 0$ for all v which implies that $f^0(x,v,t) = f_0(x,v) \equiv f^0(v)$. Any function f^0 depending only on v may be an equilibrium solution. A stable solution which corresponds to the thermodynamic equilibrium of the plasma is the case where f^0 is a Maxwellian:

$$f^0(v) = \frac{n_0}{2\pi v_{th}}\, e^{-\frac{v^2}{2v_{th}^2}},$$

denoting by $v_{th} = \frac{T}{m}$ the electron thermal velocity.

One can now linearise Vlasov-Poisson around this equilibrium state by expanding the distribution function and the electric field in the form of the equilibrium solution plus a small perturbation:

$$f(x,v,t) = f^0(x,v) + \epsilon f^1(x,v,t),\ \ E(x,t) = E^0(x) + \epsilon E^1(x,t),\ \text{(with } E^0(x) = 0).$$

The distribution function f verifies the Vlasov-Poisson equations

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - \frac{e}{m}E(x)\frac{\partial f}{\partial v} = 0,$$

$$\frac{dE}{dx} = \frac{e}{\epsilon_0}(n_0 - \int_{-\infty}^{+\infty} f(x,v,t)\,\mathrm{d}v),$$

with an initial condition that we assume of the form

$$f_0(x,v) = f^0(v) + \epsilon f_0^1(x,v).$$

Plugging the expansions of f and E in this equation

$$\epsilon\left(\frac{\partial f^1}{\partial t} + v\frac{\partial f^1}{\partial x}\right) - \frac{e}{m}(E^0 + \epsilon E^1)(\frac{\mathrm{d}f^0}{\mathrm{d}v} + \epsilon\frac{\partial f^1}{\partial v}) = 0,$$
$$\epsilon\frac{\mathrm{d}E^1}{\mathrm{d}x} = \frac{e}{\epsilon_0}(n_0 - \int_{-\infty}^{+\infty}(f^0(v) + \epsilon f^1(x,v,t))\,\mathrm{d}v.$$

Neglecting the terms in ϵ^2, we obtain, knowing that $E^0 = 0$

$$\frac{\partial f^1}{\partial t} + v\frac{\partial f^1}{\partial x} - \frac{e}{m}E^1(x)\frac{\mathrm{d}f^0}{\mathrm{d}v} = 0, \tag{3.29}$$
$$\frac{\mathrm{d}E^1}{\mathrm{d}x} = -\frac{e}{\epsilon_0}\int_{-\infty}^{+\infty} f^1(x,v,t)\,\mathrm{d}v, \tag{3.30}$$

with the initial condition $f^1(x,v,0) = f_0^1(x,v)$. As f^0 is a known function of v, this equation, the unknowns of which are f^1 and E^1, is linear and displays derivatives in x and t. We can thus compute an analytic solution analytic using, as f^1 is periodic in x, a Fourier series in x and a Laplace transform in t. To simplify notations we omit the 1 indices in the sequel of this section. Reminders in complex analysis to understand the calculations below and the definition of the Laplace transform are given in Appendix A.

We define the Fourier series of a continuous and $L-$periodic function g by

$$g(x) = \sum_{k'=-\infty}^{+\infty}\hat{g}(k)e^{ikx} \quad \text{with} \quad \hat{g}(k) = \frac{1}{L}\int_0^L g(x)e^{-ikx}\,\mathrm{d}x \quad \text{et} \quad k = \frac{2\pi}{L}k'.$$

Multiplying (3.29) and (3.30) by e^{-ikx} and integrating between 0 and L, we obtain the following relations between the Fourier coefficients

$$\frac{\partial\hat{f}}{\partial t}(k,v,t) + ikv\hat{f}(k,v,t) - \frac{e}{m}\hat{E}(k,t)\frac{\mathrm{d}f^0}{\mathrm{d}v} = 0, \tag{3.31}$$
$$ik\hat{E}(k,t) = -\frac{e}{\epsilon_0}\int_{-\infty}^{+\infty}\hat{f}(k,v,t)\,\mathrm{d}v. \tag{3.32}$$

Moreover for the initial condition $\hat{f}(k,v,0) = \hat{f}_0(k,v)$. We now perform the Laplace transform in t of these equations. In order to compare our results to the classical plasma physics textbooks, we adopt the convention used by physicists to take $s = -i\omega$ (with $\omega \in \mathbb{C}$) compared to the formulas given in appendix A.3. The Laplace transform of a function $f(t)$ then writes

$$\tilde{f}(\omega) = \int_0^{+\infty} f(t)e^{i\omega t}\,dt \quad \text{for } \Re(s) = \Im(\omega) > R, \tag{3.33}$$

and the inverse Laplace transform

$$f(t) = \int_{-\infty+iu}^{+\infty+iu} \tilde{f}(\omega) e^{-i\omega t}\, d\omega. \tag{3.34}$$

Multiplying $\frac{\partial \hat{f}}{\partial t}(k,v,t)$ with $e^{i\omega t}$ and integrating in t between 0 and $+\infty$, we have

$$\begin{aligned}\int_0^{+\infty} \frac{\partial \hat{f}}{\partial t}(k,v,t) e^{i\omega t}\, dt &= [\hat{f}(k,v,t)e^{i\omega t}]_0^{+\infty} - i\omega \int_0^{+\infty} \hat{f}(k,v,t) e^{i\omega t}\, dt \\ &= -\hat{f}(k,v,0) - i\omega \tilde{f}(k,v,\omega),\end{aligned}$$

where we denote by $\tilde{f}(k,v,\omega)$ Laplace transform of $\hat{f}(k,v,t)$. We then obtain from (3.31)

$$(-i\omega + ikv)\tilde{f}(k,v,\omega) - \frac{e}{m}\tilde{E}(k,\omega)\frac{\mathrm{d}f^0}{\mathrm{d}v} = \hat{f}_0(k,v), \tag{3.35}$$

and the Laplace transform of the Poisson equation yields

$$\tilde{E}(k,\omega) = \frac{ie}{k\epsilon_0} \int_{-\infty}^{+\infty} \tilde{f}(k,v,\omega)\, \mathrm{d}v. \tag{3.36}$$

Plugging (3.35) into (3.36) we obtain

$$\begin{aligned}\tilde{E} = \frac{ie}{k\epsilon_0} \int_{-\infty}^{+\infty} \frac{\hat{f}_0(k,v) + \frac{e}{m}\tilde{E}\frac{\mathrm{d}f^0}{\mathrm{d}v}}{-i\omega + ikv}\, \mathrm{d}v = \\ \frac{ie^2}{k\epsilon_0 m}\tilde{E}\int_{-\infty}^{+\infty} \frac{\frac{\mathrm{d}f^0}{\mathrm{d}v}}{-i\omega + ikv}\, \mathrm{d}v + \frac{ie}{k\epsilon_0}\int_{-\infty}^{+\infty} \frac{\hat{f}_0(k,v)}{-i\omega + ikv}\, \mathrm{d}v.\end{aligned}$$

Let

$$D(k,\omega) = 1 - \frac{e^2}{k^2\epsilon_0 m}\int_{-\infty}^{+\infty} \frac{\frac{\mathrm{d}f^0}{\mathrm{d}v}}{v - \frac{\omega}{k}}\, \mathrm{d}v, \tag{3.37}$$

$$N(k,\omega) = \frac{e}{k^2\epsilon_0}\int_{-\infty}^{+\infty} \frac{\hat{f}_0(k,v)}{v - \frac{\omega}{k}}\, \mathrm{d}v. \tag{3.38}$$

We then obtain the following expression for $\tilde{E}$:

$$\tilde{E}(k,\omega) = \frac{N(k,\omega)}{D(k,\omega)}.$$

The conditions for using the inverse Laplace transform, Theorem 8, are satisfied if the function $\tilde{E}(k,\cdot)$ is analytic in the stripe $\Re(s) = \Im(\omega) > R$. We can then use this expression to calculate the electric field by inverse Laplace transform. Note that $D(k,\omega)$ and $N(k,\omega)$ are well defined for $\Im(\omega) > 0$ and are analytic. in this case provided f^0 and f_0 are, as the integration is performed

on the real axis and the denominator never vanishes. The inverse Laplace transform is thus well defined. Nonetheless, in order to compute it, it is convenient to use the residue theorem, Theorem 7, which requires the function to be integrated to be analytic apart from isolated points. This is the case with the initial expression of the Laplace transform only in the half-plane $\Im(\omega) > 0$.

In order to be able to deal also with the case $\Im(\omega) \leq 0$ which is physically meaningful when damping phenomena are considered, we need to define an continuation of these functions for $\Im(\omega) \leq 0$. Let us consider a function of the form

$$G(\omega) = \int_{-\infty}^{+\infty} \frac{g(v)}{v - \frac{\omega}{k}} \, \mathrm{d}v.$$

Assume that g is analytic and that the contour integrals are well defined. Then, as v and k are real, the function $G(\omega)$ is analytic for $\Im(\omega) > 0$. Our objective now is to defined an continuation of G for $\Im(\omega) \leq 0$. For that, we are going to modify the definition of G by modifying the integration contour which in the original definition is the real axis, so that G is analytic in the half-plane $\Im(\omega) > a$, with $a < 0$ and keeps its original value for $\Im(\omega) > 0$.

Let γ the real axis parametrised by $v \in [-\infty, +\infty]$. Then we have

$$G(\omega) = \int_{\gamma} \frac{g(z)}{z - \frac{\omega}{k}} \, dz.$$

For $\Im(\omega) > 0$ the integrand has no singularity and hence $G(\omega)$ is analytic. As g tends to infinity, the integral on two contours going to infinity on each side is identical if there is no pole between the two contours. The poles are $\frac{\omega}{k}$. For $\Im(\omega) > 0$, they are above the real axis if $k > 0$ and below if $k < 0$. We thus need to distinguish two cases.

Thus for $\Im(\omega) > 0$ and $k > 0$, there is no pole below the real axis. Hence, using instead of the real axis any line below the real axis as an integration contour, we can redefine $G(\omega)$ without changing its value for $\Im(\omega) > 0$. And G so defined is analytic for ω strictly above the chosen line which includes some ω with negative imaginary part. It follows that to define an analytic continuation of $G(\omega)$ for $\Im(\omega) \leq 0$ in the case $k > 0$, it suffices to define $G(\omega)$ as an integral on a line parallel to the real axis, which passes below the pole or any continuous deformation of this line passing below the pole, chose to simplify the computation of the integral. The contours we take in practice are displayed on Figure 3.1. Hence, take the adapte contour γ, i.e. the real axis for $\Im(\omega) > 0$, the contour on the left-hand side of Figure 3.1 for $\Im(\omega) = 0$ and the contour on the right-hand side of Figure 3.1 for $\Im(\omega) < 0$, the function

$$G(\omega) = \int_{\gamma} \frac{g(z)}{z - \frac{\omega}{k}} \, dz$$

is analytic in the complex plane.

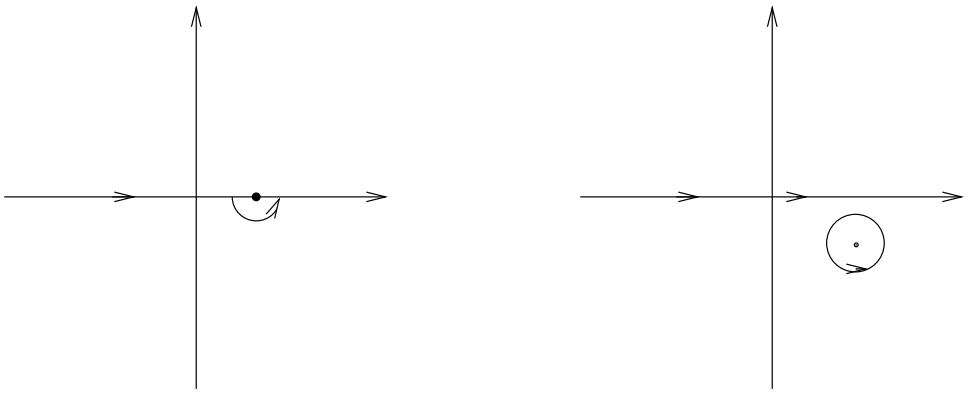

Fig. 3.1. Path for the pole on the real axis on the left, and below the real axis on the right.

In the case when $k < 0$, we define in the same way an analytic continuation of G for $\Im(\omega) \leq 0$ from its value for $\Im(\omega) > 0$ by taking as an integration contour instead of the real axis any line, or adapted contour, that passes above the poles.

We can in this way define an analytic continuation of $D(k, \omega)$ and of $N(k, \omega)$ on the whole complex plane if necessary and express their values. In general the functions f^0 and f_0 are linear combination of Maxwellians in v. The Maxwellian functions in v play a major role in the computation of kinetic plasma dispersion relations. For this reason Fried and Conte [62] have introduce a function called the plasma dispersion function, denoted by Z and defined by

$$Z_{\pm}(\zeta) = \frac{1}{\sqrt{\pi}} \int_{\gamma} \frac{e^{-z^2}}{z - \zeta} \, dz, \tag{3.39}$$

where γ, for Z_-, is any open contour parallel to the real axis at infinity that passes below the pole $z = \zeta$ or similar to the one displayed on the right hand side of Figure 3.1 and for Z_+, γ is a contour passing above the pole. A physical derivation of many more complex kinetic plasma dispersion relations can be found in the book of Stix [98].

Proposition 6 *The plasma dispersion functions Z_- (resp. Z_+, defined by* (3.39) *are independent of the contour γ of the form $t \mapsto t+iu$ for $|t|$ sufficiently large and passing below (resp. above) the pole $z = \zeta$. Moreover we have the following expressions for $Z_\pm$:*

$$Z_{\pm}(\xi) = \frac{1}{\sqrt{\pi}} \left[\mathrm{Pr} \int_{-\infty}^{+\infty} \frac{e^{-(u+\zeta)^2}}{u} \, du \mp i\pi e^{-\zeta^2} \right], \tag{3.40}$$

$$= \sqrt{\pi} e^{-\zeta^2} [\mp i - \operatorname{erfi}(\zeta)], \tag{3.41}$$

where $\operatorname{erfi}(\zeta) = \frac{2}{\sqrt{\pi}} \int_0^\zeta e^{t^2}\, dt$ *is the complex error function. One can also express the derivatives of* $Z_\pm$ *as a function of* $Z_\pm$ *using the following relations:*

$$Z'_\pm(\zeta) = \pm\zeta Z(\zeta) - 2, \tag{3.42}$$

$$Z''(\zeta) = \mp 4\zeta \pm 2Z(\zeta) + 4\zeta^2 Z(\zeta). \tag{3.43}$$

Note that for $b \in \mathbb{R}$

$$\Pr \int_{-\infty}^{+\infty} \frac{g(u)}{u-b}\, du = \lim_{\delta \to 0} \left[\int_{-\infty}^{b-\delta} \frac{g(u)}{u-b}\, du + \int_{b+\delta}^{+\infty} \frac{g(u)}{u-b}\, du \right]$$

denotes the Cauchy principal value.

Remark 4 *The complex error function* erfi *is a classical special function available in most of standard numeric and symbolic computation software.*

Proof. In order to prove that the integral is independent of the contour of the given form, it suffices to take to contours of this form. This contours are parallel to the real axis outside $[-A, A]$ for A large enough. We then can join them at $-A$ and A by lines parallel to the imaginary axis so as to obtain a closed contour. As the two chosen contours are either both below the pole, or both above, the closed contour that we constructed contains no pole so that the integral on this contour vanishes. Moreover, given the form of the function to be integrated, it is clear that the integrals on the line segments parallel to the imaginary axis tend to 0 when $A \to +\infty$. It follows that the integral on the two initial contours is the same. To obtain the expression (3.40), we choose a contour passing through the pole ζ and winding aroung from below for Z_- as displayed on the left hand side of Figure 3.1, and above for Z_+. This contour can be parametrised by $\gamma_1 : t \mapsto t + i\Im(\zeta)$ for $|t - \Re(\zeta)| \geq \delta$ which is the same for Z_- and Z_+ and $\gamma_2 : \theta \mapsto \zeta - \delta e^{\mp i\theta}$ for $\theta \in [0, \pi]$. And we have

$$\begin{aligned}
\int_{\gamma_1} \frac{e^{-z^2}}{z-\zeta} &= \int_{-\infty}^{\Re(\zeta)-\delta} \frac{e^{-(t+i\Im(\zeta))^2}}{t+i\Im(\zeta)-\zeta}\, dt + \int_{\Re(\zeta)+\delta}^{+\infty} \frac{e^{-(t+i\Im(\zeta))^2}}{t+i\Im(\zeta)-\zeta}\, dt \\
&= \int_{-\infty}^{\Re(\zeta)-\delta} \frac{e^{-(t+i\Im(\zeta))^2}}{t-\Re(\zeta)}\, dt + \int_{\Re(\zeta)+\delta}^{+\infty} \frac{e^{-(t+i\Im(\zeta))^2}}{t-\Re(\zeta)}\, dt, \\
&= \int_{-\infty}^{-\delta} \frac{e^{-(u+\zeta)^2}}{u}\, du + \int_{\delta}^{+\infty} \frac{e^{-(u+\zeta)^2}}{u}\, du,
\end{aligned}$$

Making the change of variables $u = t - \Re(\zeta)$. And then letting δ go to 0, we have

$$\int_{\gamma_1} \frac{e^{-z^2}}{z-\zeta} \to Pr \int_{-\infty}^{+\infty} \frac{e^{-(u+\zeta)^2}}{u}\, du.$$

On the other hand

$$\int_{\gamma_2} \frac{e^{-z^2}}{z-\zeta} = \int_0^{\pi} \frac{e^{-(\zeta-\delta e^{\mp i\theta})^2}}{-\delta e^{\mp i\theta}} (\pm i\delta e^{\mp i\theta})\, d\theta \to \mp i\pi e^{-\zeta^2} \text{ quand } \delta \to 0.$$

Adding the limit integrals on the two contours, we obtain the expression (3.40).

We have

$$\begin{aligned}\int_{-\infty}^{+\infty} \frac{e^{-(u+\zeta)^2}}{u}\, du &= e^{-\zeta^2} \int_{-\infty}^{+\infty} e^{-|\mathbf{u}|^2} e^{-2\zeta u} \frac{du}{u}, \\ &= e^{-\zeta^2} \int_0^{+\infty} e^{-|\mathbf{u}|^2} (e^{-2\zeta u} - e^{2\zeta u}) \frac{du}{u}, \\ &= -2e^{-\zeta^2} \int_0^{+\infty} e^{-u^2} \frac{\operatorname{sh}(2\zeta u)}{u}\, du.\end{aligned}$$

Note that as $\operatorname{sh}(2\zeta u) \sim 2\zeta u$ in the neighbourhood of $u = 0$, there is no singularity and so the integral is equal to its Cauchy principal value. Let us introduce

$$y(\zeta) = \int_0^{+\infty} e^{-u^2} \frac{\operatorname{sh}(2\zeta u)}{u}\, du.$$

Then

$$y'(\zeta) = 2 \int_0^{+\infty} e^{-u^2} \operatorname{ch}(2\zeta u)\, du,$$

taking the derivative and integrating by parts

$$y''(\zeta) = 4 \int_0^{+\infty} e^{-u^2} u \operatorname{sh}(2\zeta u)\, du = 4\zeta \int_0^{+\infty} e^{-u^2} \operatorname{ch}(2\zeta u)\, du = 2\zeta y'(\zeta).$$

It follows that $y'(\zeta) = y'(0)e^{\zeta^2}$. Or $y'(0) = 2\int_0^{+\infty} e^{-u^2}\, du = \sqrt{\pi}$. Whence $y'(\zeta) = \sqrt{\pi} e^{\zeta^2}$. Then, as $y(0) = 0$, we have

$$y(\zeta) = \sqrt{\pi} \int_0^{\zeta} e^{x^2}\, \mathrm{d}x = \frac{\pi}{2} \operatorname{erfi}(\zeta),$$

using the definition $\operatorname{erfi}(\zeta) = \frac{2}{\sqrt{\pi}} \int_0^{\zeta} e^{x^2}\, \mathrm{d}x$. We then deduce the expression (3.41) from (3.40).

We obtain the expression of Z' taking the derivative of (3.41) :

$$\begin{aligned}Z'(\zeta) &= \pm 2\zeta Z(\zeta) - \operatorname{erfi}'(\zeta)\sqrt{\pi} e^{-\zeta^2}, \\ &= \pm 2\zeta Z(\zeta) - \frac{2}{\sqrt{\pi}} e^{\zeta^2} \sqrt{\pi} e^{-\zeta^2}, \\ &= \pm 2\zeta Z(\zeta) - 2.\end{aligned}$$

Taking again the derivative we obtain

$$\begin{aligned} Z''(\zeta) &= \pm 2(Z(\zeta) + \zeta Z'(\zeta)), \\ &= \pm 2(Z(\zeta) - 2\zeta \pm 2\zeta^2 Z(\zeta)), \\ &= \mp 4\zeta \pm 2Z(\zeta) + 4\zeta^2 Z(\zeta). \end{aligned}$$

We can now continue the computation of $D(k,\omega)$. We have

$$f^0(v) = \frac{n_0}{\sqrt{2\pi} v_{th}} e^{-\frac{v^2}{2v_{th}^2}},$$

and thus

$$\frac{\mathrm{d}f^0}{\mathrm{d}v}(v) = -\frac{n_0}{\sqrt{2\pi} v_{th}} \frac{v}{v_{th}^2} e^{-\frac{v^2}{2v_{th}^2}}.$$

Plugging this expression into the expression of $D(k,\omega)$ (3.37) and introducing the plasma frequency $\omega_p^2 = \frac{n_0 e^2}{\epsilon_0 m}$, we obtain

$$\begin{aligned} D(k,\omega) &= 1 + \frac{\omega_p^2}{k^2 v_{th}^2} \frac{1}{\sqrt{2\pi} v_{th}} \int_{-\infty}^{+\infty} \frac{v e^{-\frac{v^2}{2v_{th}^2}}}{v - \frac{\omega}{k}}, \\ &= 1 + \frac{\omega_p^2}{k^2 v_{th}^2} \frac{1}{\sqrt{2\pi} v_{th}} \left[\int_{-\infty}^{+\infty} \frac{(v - \frac{\omega}{k}) e^{-\frac{v^2}{2v_{th}^2}}}{v - \frac{\omega}{k}} \, dv + \frac{\omega}{k} \int_{-\infty}^{+\infty} \frac{e^{-\frac{v^2}{2v_{th}^2}}}{v - \frac{\omega}{k}} \, \mathrm{d}v \right]. \end{aligned}$$

But

$$\frac{1}{\sqrt{2\pi} v_{th}} \int_{-\infty}^{+\infty} e^{-\frac{v^2}{2v_{th}^2}} \, \mathrm{d}v = 1.$$

Then making the change of variables $u = \frac{v}{\sqrt{2} v_{th}}$, it comes

$$D(k,\omega) = 1 + \frac{\omega_p^2}{k^2 v_{th}^2} \left[1 + \frac{\omega}{k} \frac{1}{\sqrt{2\pi} v_{th}} \int_{-\infty}^{+\infty} \frac{e^{-u^2}}{u - \frac{\omega}{k\sqrt{2} v_{th}}} \, \mathrm{d}v \right].$$

Using the expression of the plasma dispersion function Z this relation writes

$$D(k,\omega) = 1 + \frac{\omega_p^2}{k^2 v_{th}^2} \left[1 + \frac{\omega}{\sqrt{2} v_{th} k} Z(\frac{\omega}{\sqrt{2} v_{th} k}) \right], \tag{3.44}$$

which becomes with the expression of Z given by (3.41)

$$D(k,\omega) = 1 + \frac{\omega_p^2}{k^2 v_{th}^2} \left[1 + \frac{\sqrt{\pi}\omega}{\sqrt{2} v_{th} k} e^{-\frac{\omega^2}{2v_{th}^2 k^2}} \left(\mp i - \operatorname{erfi}(\frac{\omega}{\sqrt{2} v_{th} k}) \right) \right], \tag{3.45}$$

and using the expression of Z' as a function of Z (3.42) we obtain the following simpler form

$$D(k,\omega) = 1 - \frac{1}{2} \frac{\omega_p^2}{k^2 v_{th}^2} Z'(\frac{\omega}{\sqrt{2} v_{th} k}). \tag{3.46}$$

These computations have been performed on the initial expression valid for $\Im(\omega) > 0$, but taking the analytic continuation as we have seen, we obtain the same expressions by choosing the adequate integration contour for any ω.

We can now perform in the same way the computation of $N(k,\omega)$ assuming that

$$f_0^1(x,v) = g(x)\frac{n_0}{\sqrt{2\pi}v_{th}}e^{-\frac{v^2}{2v_{th}^2}},$$

where g is a given function often of the form $g(x) = \cos(kx)$. We then have

$$\hat{f}_0^1(k,v) = \hat{g}(k)\frac{n_0}{\sqrt{2\pi}v_{th}}e^{-\frac{v^2}{2v_{th}^2}}.$$

Then

$$\begin{aligned} N(k,\omega) &= -\frac{e}{k^2\epsilon_0}\int_{-\infty}^{+\infty}\frac{\hat{f}_0^1}{v-\frac{\omega}{k}}\,\mathrm{d}v, \\ &= -g(k)\frac{n_0 e}{k^2\epsilon_0}\frac{1}{\sqrt{2\pi}v_{th}}\int_{-\infty}^{+\infty}\frac{e^{-\frac{v^2}{2v_{th}^2}}}{v-\frac{\omega}{k}}\,\mathrm{d}v. \end{aligned}$$

Making the change of variables $u = \frac{v}{\sqrt{2}v_{th}}$ to obtain

$$N(k,\omega) = -\hat{g}(k)\frac{n_0 e}{k^2\epsilon_0}\frac{1}{\sqrt{2\pi}v_{th}}\int_{-\infty}^{+\infty}\frac{e^{-u^2}}{u-\frac{\omega}{\sqrt{2}v_{th}k}}\,du.$$

We recognise the plasma dispersion function Z and thus have

$$N(k,\omega) = -\hat{g}(k)\frac{n_0 e}{k^2\epsilon_0}\frac{1}{\sqrt{2}v_{th}}Z(\frac{\omega}{\sqrt{2\pi}v_{th}k}). \tag{3.47}$$

Finally from expressions (3.46) and (3.47), we obtain an explicit formula for the Laplace transform of the electric field $\tilde{E}$. We then deduce the electric field itself using the inverse Fourier and Laplace transforms. Starting with the inverse Laplace transform (A.4), we have

$$\hat{E}(k,t) = \frac{1}{2i\pi}\int_{-\infty+iu}^{+\infty+iu}\tilde{E}(k,\omega)e^{-i\omega t}\,d\omega.$$

We can compute this integral using the residue theorem, Theorem 7 by closing the contour by a half-circle towards the bottom of the complex plane and of radius going to infinity. Assuming that $\tilde{E}(k,\omega)$ is analytic apart from a finite number of poles and the the integral on the half-circle tend to 0 when the radius goes to infinity we get

$$\hat{E}(k,t) = \sum_j \mathrm{Res}_{\omega=\omega_j}(\tilde{E}(k,\omega))e^{-i\omega_j t},$$

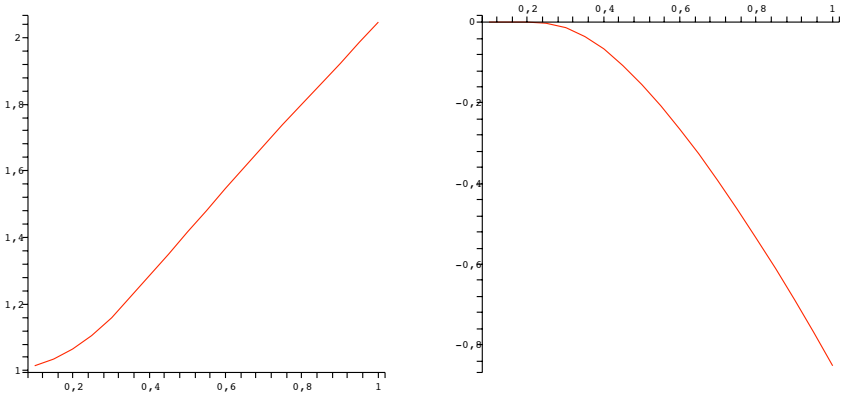

Fig. 3.2. Real part (left) and imaginary part (right) of the zeros ω of $D(k,\omega)$ as a function of k.

where the sum is taken over the poles.

To obtain an explicit value of this expression of the electric field, it remains to compute numerically for k fixed the values of ω for which $D(k,\omega)$ vanishes. The simplest way to this is to use the Newton method, but this needs a good initial guess. We obtain the following values ω for different k:

k	ω
0.5	$1,4156 - 0,1533i$
0.4	$1,2850 - 0,0661i$
0.3	$1,1598 - 0,0126i$
0.2	$1,0640 - 5,510 \times 10^{-5}i$

Newton's method is very sensitive to the initial guess and gives no insurance to find the most unstable or the least damped mode. A more robust method to compute the zeros of an analytic function will be given in the next section.

Another option, applicable in some cases, is to perform an asymptotic expansion. When considering waves such that $v_\phi = \omega/k \gg v_{th}$ which corresponds to the limit $\omega/(kv_{th}) \to +\infty$, we can simplify the dispersion relation by making an asymptotic expansion of the function erfi in the neighbourhood of $+\infty$. For this we start by expressing an asymptotic expansion of its derivative $\text{erfi}'(x) = 2e^{x^2}/\sqrt{\pi}$. Consider the function

$$g(x) = \frac{e^{x^2}}{\sqrt{\pi}}\left(\frac{1}{x} + \frac{1}{2x^3} + \frac{3}{4x^5}\right).$$

We then have

$$g'(x) = \frac{e^{x^2}}{\sqrt{\pi}}\left(2 - \frac{1}{x^2} + \frac{1}{x^2} - \frac{3}{2x^4} + \frac{3}{2x^4} - \frac{15}{4x^6}\right).$$

It follows that in a neighbourhood of $+\infty$,

$$\mathrm{erfi}\,'(x) = g'(x) + O(\frac{e^{x^2}}{x^6}),$$

and as $\mathrm{erfi}\,(x) \to +\infty$ when $x \to +\infty$, the constant appearing in the integration is negligible and it comes

$$\mathrm{erfi}\,(x) = g(x) + O(\frac{e^{x^2}}{x^7}).$$

Finally remplacing erfi by this expression in the expression of $D(k,\omega)$, we obtain

$$\begin{aligned} D(k,\omega) &= 1 + \frac{\omega_p^2}{k^2 v_{th}^2}\left[1 - \left(1 + \frac{k^2 v_{th}^2}{\omega^2} + \frac{3k^4 v_{th}^4}{\omega^4}\right) + i\sqrt{\frac{\pi}{2}}\frac{\omega}{k v_{th}} e^{-\frac{\omega^2}{2k^2 v_{th}^2}}\right] \\ &= 1 - \frac{\omega_p^2}{\omega^2}(1 + 3\frac{k^2 v_{th}^2}{\omega^2}) + i\sqrt{\frac{\pi}{2}}\frac{\omega_p^2 \omega}{k^3 v_{th}^3} e^{-\frac{\omega^2}{2k^2 v_{th}^2}}. \end{aligned} \tag{3.48}$$

This expression corresponds to the classical expression found in the introductory plasma physics textbooks for example [32] which is in general derived making the hypothesis $\frac{\omega}{k} \gg v_{th}$ to compute the integral

$$\mathrm{Pr}\int_{-\infty}^{+\infty} \frac{\frac{\mathrm{d}f^0}{\mathrm{d}v}}{v - \omega/k}\,\mathrm{d}v,$$

and taking an asymptotic expansion at the denominator.

From expression (3.48) of D, we can obtain an explicit formula for the real part ω_r et imaginaire ω_i of ω assuming $\omega_i \ll \omega_r$. We then have to first order$D_r(\omega_r, k) = 0$ et

$$\omega_i = -\frac{D_i(\omega_r, k)}{\frac{\partial D_r}{\partial \omega_r}(\omega_r, k)}.$$

The dispersion relation that we derived in this section can also be used for other equilibrium distributions. A useful case is for example the superposition of several Maxwellians centred at different velocities with possibly different thermal velocities. We have in this case

$$f^0(v) = \frac{n_0}{N\sqrt{2\pi}} \sum_{i=1}^{N} \frac{1}{v_{th_i}} e^{-\frac{(v-v_i)^2}{2v_{th_i}^2}},$$

and the dispersion relation writes

$$\begin{aligned} D(k,\omega) = 1 + \frac{\omega_p^2}{Nk^2} \sum_{i=1}^{N} \frac{1}{v_{th_i}^2} \Bigg[1 + \\ \sqrt{\frac{\pi}{2}}\left(\frac{\omega}{k v_{th_i}} - \frac{v_i}{v_{th_i}}\right) e^{-\frac{(\omega/k - v_i)^2}{2v_{th_i}^2}} \left(i - \mathrm{erfi}\,(\frac{\omega/k - v_i}{\sqrt{2} v_{th_i}})\right)\Bigg]. \end{aligned} \tag{3.49}$$

3.4 The Vlasov-Maxwell system

3.4.1 The equations

We consider here the dimensionless Vlasov-Maxwell equations for one species of particles:

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + (\mathbf{E} + \mathbf{v}\times\mathbf{B})\cdot\nabla_v f = 0, \tag{3.50}$$

$$-\frac{\partial \mathbf{E}}{\partial t} + \nabla\times\ \mathbf{B} = \mathbf{J}, \tag{3.51}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla\times\ \mathbf{E} = 0, \tag{3.52}$$

$$\nabla\cdot\ \mathbf{E} = \rho, \tag{3.53}$$

$$\nabla\cdot\ \mathbf{B} = 0, \tag{3.54}$$

with

$$\rho(\mathbf{x},t) = \int f(\mathbf{x},\mathbf{v},t)\,d\mathbf{v}, \quad \mathbf{J}(\mathbf{x},t) = \int f(\mathbf{x},\mathbf{v},t)\mathbf{v}\,d\mathbf{v}.$$

We first notice that the Vlasov equation (3.50) can also be written in conservative form

$$\frac{\partial f}{\partial t} + \nabla_{\mathbf{x},\mathbf{v}}\cdot(\mathbf{F}f) = 0, \tag{3.55}$$

with $\mathbf{F} = (\mathbf{v}, \mathbf{E} + \mathbf{v}\times\mathbf{B})^T$ such that $\nabla_{\mathbf{x},\mathbf{v}}\cdot\mathbf{F} = 0$.

3.4.2 Conservation properties

Proposition 7 *The Vlasov-Maxwell equations verify following properties:*

- *Maximum principle*

$$0 \le f(\mathbf{x},\mathbf{v},t) \le \max_{(\mathbf{x},\mathbf{v})}(f_0(\mathbf{x},\mathbf{v})). \tag{3.56}$$

- *Conservation of L^p, norms p integer, $1 \le p \le \infty$*

$$\frac{d}{dt}\left(\int (f(\mathbf{x},\mathbf{v},t))^p\,d\mathbf{x}\,d\mathbf{v}\right) = 0 \tag{3.57}$$

- *Conservation of volume. For any volume V of phase space*

$$\int_V f(\mathbf{x},\mathbf{v},t)\,d\mathbf{x}\,d\mathbf{v} = \int_{F^{-1}(V)} f_0(\mathbf{y},\mathbf{u})\,d\mathbf{y}\,d\mathbf{u}. \tag{3.58}$$

- *Conservation of momentum*

$$\frac{d}{dt}\left[\int \mathbf{v}f\,d\mathbf{x}d\mathbf{v} - \int \mathbf{E}\times\mathbf{B}\,d\mathbf{x}\right] = \frac{d}{dt}\left[\int \mathbf{J}\,d\mathbf{x} - \int \mathbf{E}\times\mathbf{B}\,d\mathbf{x}\right] = 0. \tag{3.59}$$

- *Conservation of energy*

$$\frac{d}{dt}\left[\frac{1}{2}\int |\mathbf{v}|^2 f\, d\mathbf{x}d\mathbf{v} + \frac{1}{2}\int (E^2 + B^2)\, d\mathbf{x}\right] = 0. \tag{3.60}$$

Proof. The characteristic curves of our Vlasov equation are the solutions of the following system of ordinary differential equations (ODE):

$$\frac{d\mathbf{X}}{dt} = \mathbf{V}(t), \tag{3.61}$$

$$\frac{d\mathbf{V}}{dt} = \frac{q}{m}(\mathbf{E}(\mathbf{X}(t), t) + \mathbf{V}(t) \times \mathbf{B}(\mathbf{X}(t), t)). \tag{3.62}$$

We denote by $(\mathbf{X}(t; \mathbf{x}, \mathbf{v}, s), \mathbf{V}(t; \mathbf{x}, \mathbf{v}, s))$, or more concisely $(\mathbf{X}(t), \mathbf{V}(t))$ when the dependence with respect to the initial conditions is not explicitly needed, the unique solution at time t of this system which takes the value $(\mathbf{x}, \mathbf{v})$ at time s.

As

$$\nabla_{(x,v)} \cdot \begin{pmatrix} \mathbf{v} \\ \mathbf{E} + \mathbf{v} \times \mathbf{B} = 0 \end{pmatrix},$$

The first three conservation properties are obtained exactly as in the case the Vlasov-Poisson system that we considered in Section 3.2.2

Let us proceed to the proof of the conservation of momentum. Multiply the Vlasov equation (3.50) by $\mathbf{v}$ and integrate in $\mathbf{x}$ and in $\mathbf{v}$

$$\frac{d}{dt}\int \mathbf{v} f\, d\mathbf{x}d\mathbf{v} + \int \nabla_x \cdot (\mathbf{v} \otimes \mathbf{v} f)\, d\mathbf{x}d\mathbf{v} - \int \mathbf{v}\nabla_v \cdot ((\mathbf{E} + \mathbf{v} \times \mathbf{B}) f)\, d\mathbf{x}d\mathbf{v} = 0.$$

The second integral vanishes as the domain is periodic in $\mathbf{x}$ and the Green formula on the divergence (3.25) yields for the last integral

$$-\int \mathbf{v}\nabla_v \cdot ((\mathbf{E}+\mathbf{v}\times\mathbf{B}) f)\, d\mathbf{x}d\mathbf{v} = \int (\mathbf{E}+\mathbf{v}\times\mathbf{B}) f\, d\mathbf{x}d\mathbf{v} = \int \mathbf{E}\rho\, d\mathbf{x} + \int \mathbf{J}\times\mathbf{B}\, d\mathbf{x}.$$

Now taking the cross product of the Ampre equation (3.51) with $\mathbf{B}$ and integrating, it follows

$$\int \mathbf{J} \times \mathbf{B}\, d\mathbf{x} = -\int \frac{\partial \mathbf{E}}{\partial t} \times \mathbf{B}\, d\mathbf{x} + \int (\nabla \times \mathbf{B}) \times \mathbf{B}\, d\mathbf{x} = -\int \frac{\partial \mathbf{E}}{\partial t} \times \mathbf{B}\, d\mathbf{x},$$

as the second integral vanishes using the equality (3.23) and the fact that $\nabla \cdot \mathbf{B} = 0$. In the same way, taking the cross product of the Faraday equation (3.52) with $\mathbf{E}$ and integrating, we get

$$\int \frac{\partial \mathbf{B}}{\partial t} \times \mathbf{E}\, d\mathbf{x} = -\int (\nabla \times \mathbf{E}) \times \mathbf{E}\, d\mathbf{x} = \int \rho \mathbf{E}\, d\mathbf{x}$$

using the equality (3.23) and the Poisson equation (3.53). It finally results that

$$\frac{d}{dt}\int \mathbf{v} f \, d\mathbf{x}d\mathbf{v} - \frac{d}{dt}\int \mathbf{E}\times\mathbf{B}\, d\mathbf{x} = 0.$$

Finally, in order to obtain the energy conservation for the Vlasov-Maxwell system, we start by multiplying the Vlasov equation by $\mathbf{v}\cdot\mathbf{v} = |\mathbf{v}|^2$ then we integrate on phase space

$$\frac{d}{dt}\int |\mathbf{v}|^2 f \, d\mathbf{x}d\mathbf{v} + \int \nabla_x\cdot(|\mathbf{v}|^2\mathbf{v}f)\, d\mathbf{x}d\mathbf{v} + \int |\mathbf{v}|^2\nabla_v\cdot((\mathbf{E}+\mathbf{v}\times\mathbf{B})f)\, d\mathbf{x}d\mathbf{v} = 0.$$

But as f is periodic in $\mathbf{x}$ and tends to 0 at infinity in $\mathbf{v}$ we have

$$\int \nabla_x\cdot(|\mathbf{v}|^2\mathbf{v}f)\, d\mathbf{x}d\mathbf{v} = 0$$

and

$$\int |\mathbf{v}|^2\nabla_v\cdot((\mathbf{E}+\mathbf{v}\times\mathbf{B})f)\, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{v}\cdot((\mathbf{E}+\mathbf{v}\times\mathbf{B})f)\, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{E}\cdot\mathbf{J}\, d\mathbf{x}.$$

So

$$\frac{d}{dt}\int |\mathbf{v}|^2 f \, d\mathbf{x}d\mathbf{v} = 2\int \mathbf{E}\cdot\mathbf{J}\, d\mathbf{x}. \tag{3.63}$$

We now take the dot product of the Ampre equation with $\mathbf{E}$ and we integrate

$$\frac{1}{2}\frac{d}{dt}\int E^2\, d\mathbf{x} - \int \nabla\times\mathbf{B}\cdot\mathbf{E}\, d\mathbf{x} = -\int \mathbf{J}\cdot\mathbf{E}\, d\mathbf{x}, \tag{3.64}$$

then we take the dot product of the Faraday equation with $\mathbf{B}$ and we integrate

$$\frac{1}{2}\frac{d}{dt}\int B^2\, d\mathbf{x} + \int \nabla\times\mathbf{E}\cdot\mathbf{B}\, d\mathbf{x} = 0. \tag{3.65}$$

Let us recall here the Green formula on the curl

$$\int_\Omega \mathbf{F}\cdot\nabla\times\mathbf{G} - \int_\Omega \nabla\times\mathbf{F}\cdot\mathbf{G} = \int_{\partial\Omega}(\mathbf{F}\times\mathbf{n})\cdot\mathbf{G} \quad \forall\mathbf{F}\in H(curl,\Omega),\ \mathbf{G}\in H^1(\Omega)^3, \tag{3.66}$$

with

$$H(curl,\Omega) = \{\mathbf{F}\in L^2(\Omega)\mid \nabla\times\mathbf{F}\in L^2(\Omega)\}.$$

Using the Green formula on the curl in a periodic domain we get

$$\int \nabla\times\mathbf{B}\cdot\mathbf{E}\, d\mathbf{x} = \int \nabla\times\mathbf{E}\cdot\mathbf{B}\, d\mathbf{x},$$

Thus adding equations (3.63) and equations (3.64) and (3.65) multiplied by two, we obtain the conservation of total energy

$$\frac{d}{dt}\left[\frac{1}{2}\int |\mathbf{v}|^2 f\, d\mathbf{x}d\mathbf{v} + \frac{1}{2}\int (E^2+B^2)\, d\mathbf{x}\right] = 0,$$

the first term stands for the kinetic energy of the particles and the second for the potential energy.

4

Code verification

4.1 Introduction

Codes developed to simulation the Vlasov-Maxwell equations or one of their approximations are quite complicated. This is why a systematic procedure should be used exclude as much as possible the sources of errors. Such a procedure is called *verification* and *validation* or *V & V*. Let us try and define more precisely these terms. Simulation codes provide an approximate solution to a physical model, expressed in the form of a set of equations that represents the reality that can be measured in an experiment. *Verification* consists in making sure that the simulation code is indeed a good approximation of the initial model and *validation* consists in comparing directly the simulation results and the experiment. Given the complexity of the problem and the experiments, *verification* is an essential step to make sure that a code correctly implements a given model.

Rather than comparing codes based on a more or less loose physics problem, we advocate the necessity to introduce *verification* problems as clearly defined and well-posed mathematical models. Such a model should consist in a set of equations, with boundary and initial conditions, including the needed profiles, that admit a unique solution and possibly an analytical solution or at least some known analytical features, like conservation of some quantities. Such information can then be used for *verification*. A good introduction to the Verification and Validation processes for magnetic fusion can be found in [67].

As some unknown quantities, be it only round-off errors, are necessarily introduced into a simulation, this should also be taken into account when examining the results of a numerical simulation. Uncertainty Quantification can be used to determine the effects of these uncertainties on the final results, but this is beyond the scope of this book.

In this chapter we are going to introduce some models for which an analytical theory can is available, which can be used to verify our codes. These

will provide the so-called test cases that will be used later to verify our codes and compare our numerical schemes on.

The verification of Vlasov-Poisson or Vlasov-Maxwell solver starts naturally with the testing of the different building bricks of the code: low level routines like interpolation for example, field solver *i.e.* Poisson or Maxwell in most cases, Vlasov solver.

Sufficiently many verification tests need to be found so that as much of the potential problems that can occur will be detected. These tests will rely on analytic solutions, convergence order and other known mathematical properties of the equations being validated. In particular, it is important to check that exact conservation properties of the scheme are indeed satisfied up to round off errors. It is also important to monitor the conservation properties of the equations that are not exactly verified by the scheme. This can be a good indicator of the accuracy of the code.

4.2 Test of the field solver

The field solvers, if they are not coupled with the Vlasov equation are linear partial differential equations for which the method of manufactured solutions is very convenient to find verification tests. The method of manufactured solutions consists in adapting the data for a given well posed partial differential equation to a well chosen smooth solution. This enables to generate easily an analytical solution, which then is used to do convergence tests and compare the numerical order of accuracy with the theoretical order.

If only some specific boundary conditions are available in the code, one needs to make sure that the analytical solution that is chosen verifies these boundary conditions.

4.2.1 Test of the Poisson solver

For the Poisson equation in 1D or on a tensor product domain, with periodic, homogeneous Dirichlet or Neumann boundary conditions the simplest test consist in taking the Fourier modes as the analytical solution. For a spectral method one should then obtain the exact solution, up to round-off errors, for spectral method and be able to check the theoretical order of accuracy for all other methods.

4.2.2 Test of the Maxwell solver

For the Maxwell equations, one can also use the method of manufactured solutions as well as simple physics problems for which analytical solutions are available, like wave-guides or cavity modes.

As we will see later, coupling Maxwell with Vlasov can be unstable in long time. This problem has been known since the first electromagnetic particle in

cell simulations. But actually, it can only be seen for the Maxwell equations on their own for some specific sources. One good verification problem in 2D has been proposed in [72, 52] and also extended in [99] to assess the stability of 3D DG solvers with hyperbolic field correction.

Here we consider the 2D Maxwell equations (2.48)–(2.50) in a metallic cavity $\Omega = [0,1]^2$, with articifial permittivity ϵ_0 and light speed c equal to one. The given current source is

$$\mathbf{J}(t,x,y) = (\cos(t)-1)\begin{pmatrix}\pi\cos(\pi x)+\pi^2 x\sin(\pi y)\\ \pi\cos(\pi y)+\pi^2 y\sin(\pi x)\end{pmatrix} - \cos(t)\begin{pmatrix}x\sin(\pi y)\\ y\sin(\pi x)\end{pmatrix}$$

and an exact solution for this source is

$$\mathbf{E}(t,x,y) = \sin(t)\begin{pmatrix}x\sin(\pi y)\\ y\sin(\pi x)\end{pmatrix}$$

and

$$B(t,x,y) = (\cos(t)-1)\big(\pi y\cos(\pi x) - \pi x\cos(\pi y)\big).$$

Note that for this solution the associated charge density reads

$$\rho(t,x,y) = \sin(t)\big(\sin(\pi x)+\sin(\pi y)\big).$$

4.3 Test of the Vlasov solver with a given advection field

The first test that should be performed with the Vlasov solver, is to take a given force field, reduced in the Vlasov-Poisson case to the electric field. In particular, for the 1D Vlasov-Poisson equations one can take $E(x) = x$. The Vlasov equation then reads

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

The corresponding characteristics are

$$\frac{dX}{dt} = V, \qquad \frac{dV}{dt} = -X,$$

which corresponds to a rotation field, so that the solution after one turn $t = 2\pi$ is exactly back at the initial condition, so that one use the initial solution as an analytical solution after a complete number of terms. One can then test the solver, using first smooth initial conditions (like Gaussians) to test convergence and convergence order and then non smooth initial conditions, like the slotted cylinder or a cone to test the appearance of oscillations. See Figure 4.1.

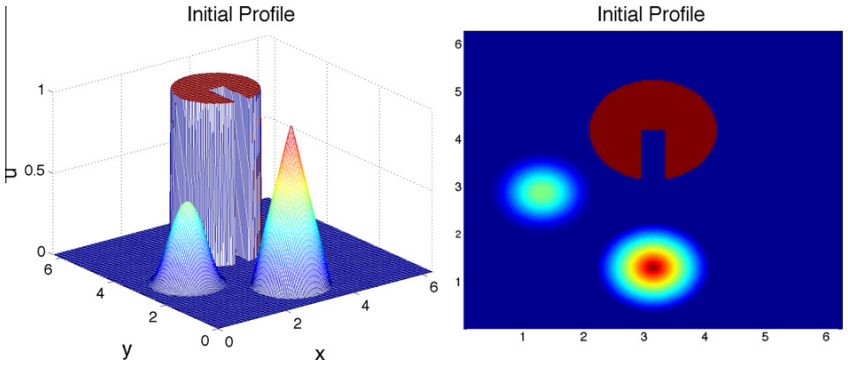

Fig. 5.2. Plots of the initial profile.

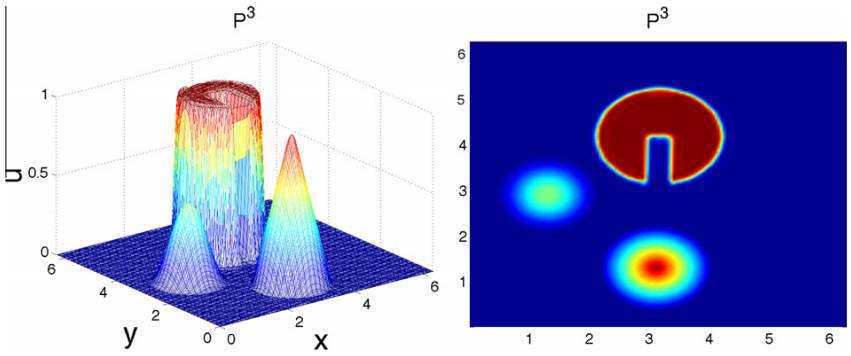

Fig. 4.1. Possible initial conditions for a rotating advection field and their evolution. From Qiu-Shu J. Comput. Phys. 2011

4.4 Comparison with analytical solution of the linearised Vlasov-Poisson system

Once all the building bricks are validated, one can come to the validation of the global code which can be done by using known solutions of the linearized Vlasov-Poisson equations. We shall here consider three classical test cases: Landau damping, two-stream instability and bump on tail.

As we saw in Chapter 3, the electric field solution of the linearised Vlasov-Poisson equations writes

$$\hat{E}(k,t) = \sum_j Res_{\omega=\omega_j} \tilde{E}(k,\omega) e^{-i\omega t},$$

where

$$\tilde{E}(k,\omega) = \frac{N(k,\omega)}{D(k,\omega)}$$

and the ω_j are the roots of the dispersion relation $D(k,\omega) = 0$ for fixed k. There are in general several roots of $D(k,\omega)$. These roots can be computed using techniques from complex analysis enabling to find all roots within some close contour, which will be described below. In the linear phase only the dominant root matters (i.e. the one with the largest eigenvalue). Only the dominant root can be observed after some time, the contribution of the others becoming negligible.

In order to compute the residues we can write the Taylor expansion of $D(k,\omega)$ in the neighbourhood of ω_j

$$D(k,\omega) = D(k,\omega_j) + (\omega - \omega_j)\frac{\partial D}{\partial \omega}(k,\omega_j) + O((\omega-\omega_j)^2),$$

and so, if ω_j is a simple root, we have $D(k,\omega_j) = 0$ and $\frac{\partial D}{\partial \omega}(k,\omega_j) \neq 0$. It then results that

$$Res_{\omega=\omega_j}\left(\frac{N(k,\omega)}{D(k,\omega)}e^{-i\omega t}\right) = \lim_{\omega\to\omega_j}((\omega-\omega_j)\frac{N(k,\omega)}{D(k,\omega)}e^{-i\omega t}) = \frac{N(k,\omega_j)}{\frac{\partial D}{\partial \omega}(k,\omega)}e^{-i\omega_j t}.$$

4.4.1 Computation of the zeros of an analytic function

For a given k the dispersion function $D(\omega,k)$ is analytic in ω. We are then reduced to the computation of the zeros of an analytic function that we shall denote by f in this section. An algorithm for doing this in the case f and its derivative f' can be evaluated at any complex point, which is the case for us, was proposed by Delves and Lyness [50] and a slightly more robust version of the algorithm was implemented by Kravanja *et al.* in the freely available Zeal package [74]. Another implementation was also proposed by Davies [46]. It is again based on the theorem of residues Theorem 7.

If an analytic function f has a zero at a complex point z_j, then f'/f has a simple pole at z_j with multiplicity given by the corresponding residue. Hence, assuming that γ is a closed contour in the complex plane that does not go through any zero of f the theorem of residues implies that N the number of poles of f enclosed in the contour γ, including multiplicities, is

$$N = \frac{1}{2i\pi}\int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z.$$

Moreover the same theorem also yields, for any $p \in \mathbb{N}$

$$s_p := z_1^p + \cdots + z_N^p = \frac{1}{2i\pi}\int_\gamma z^p \frac{f'(z)}{f(z)}\,\mathrm{d}z.$$

Delves and Lyness then introduced the polynomial P_N with zeros $z_1,\ldots,z_N$

$$P_N(z) = \Pi_{j=1}^N (z - z_j) = z^N + \sigma_1 z^{N-1} + \ldots \sigma_N.$$

The coefficients $\sigma_1, \ldots, \sigma_N$ of P_N are related to the values $s_1, \ldots, s_N$ by the following Newton identities:

$$\begin{aligned} s_1 + \sigma_1 &= 0, \\ s_2 + s_1\sigma_1 + 2\sigma_2 &= 0, \\ &\vdots \\ s_N + s_{N-1}\sigma_1 + \cdots + s_1\sigma_{N-1} + N\sigma_N &= 0. \end{aligned}$$

Then the algorithm is straightforward. The values s_j can be computing by numerically computing the contour integrals $\int_\gamma z^p f'(z)/f(z)\,\mathrm{d}z$. Then the coefficients of the polynomial σ_j can be computed using Newton's identities, so that the problem reduces to computing the roots of a known polynomial, for which there are many known procedures especially for small degrees. One needs to be careful however that the mapping associating the σ_j to the s_j can be ill condition if zeros are too close or if N is too large. For this we advocate an adaptive procedure dividing the chosen contour into four smaller contours if there are convergence problems.

The algorithm then proceeds as follows: 1) The user gives a rectangular box in which the zeros are to be found. 2) Compute the number of zeros with multiplicities N if $N > 5$ subdivide the box into four sub boxes and start again for each box.

Convergence problems can still occur if a zero is too close to the contour or if two distinct zeros are too close. This can be solved by shifting slightly the contour and/or further subdividing the box. All this is done automatically in the code.

An application of this method for finding all the zeros with imaginary part larger than -2 of the Landau dispersion relation for different values of k is displayed in Figure 4.2.

4.4.2 Landau damping

The initial conditions of Landau damping corresponds to

$$f_0(x,v) = (1 + \epsilon\cos(kx))\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}, \quad L = 4\pi.$$

In our numeric simulations we take $\epsilon = 0.001$. Physically this represents a small perturbation of a Maxwellian equilibrium. This equilibrium is stable and the distribution comes back to its equilibrium after the perturbation.

In this case the equilibrium function f^0 and the initial perturbation f_0^1 are defined by

$$f^0(v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}, \qquad f_0^1(x,v) = \epsilon\cos(kx))\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}.$$

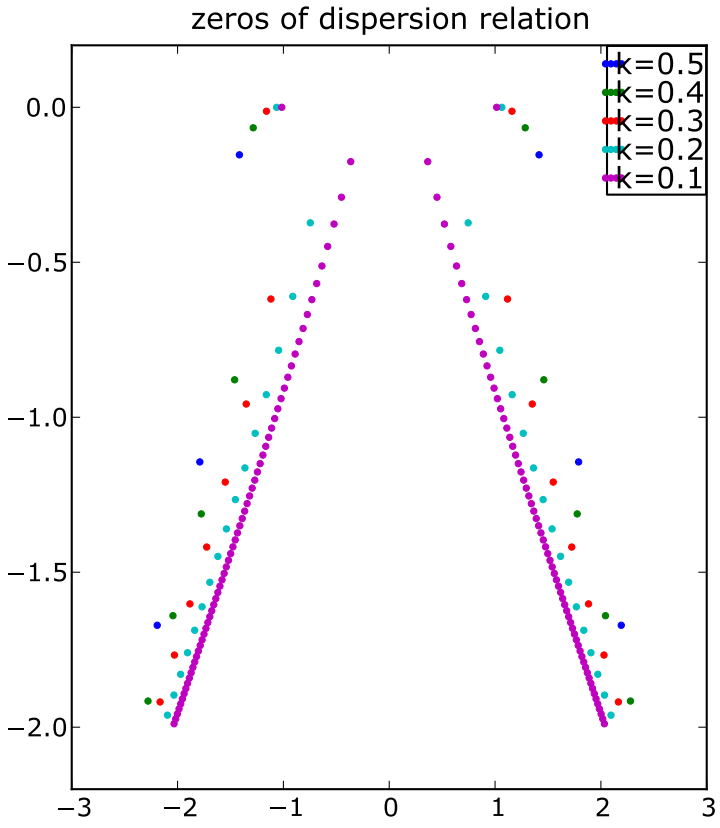

Fig. 4.2. All zeros with imaginary part larger than -2 of the Landau dispersion relation for different values of k.

It follows that associated perturbation of the charge density is $\rho_0^1(x) = \int f_0^1(x,v)\,dv = \epsilon\cos(kx))$, so that only the Fourier modes -1 and 1 associated to $k = \pm\frac{2\pi}{L}$ are non zero and their value is $\frac{1}{2}$. Thus $g(k)$ appearing in (3.47) is $\frac{1}{2}$ for these values of k and 0 for others. Then, as the electric field satisfies $\frac{dE}{dx}(x,0) = \rho_0^1(x)$, on a $E(x,0) = \frac{\epsilon}{k}\sin(kx)$ which has the same non vanishing Fourier modes with values $\frac{1}{2}$ for $k' = 1$ and $-\frac{1}{2}$ for $k' = -1$.

We recall (3.46)

$$D(k,\omega) = 1 - \frac{1}{2}\frac{\omega_p^2}{k^2 v_{th}^2} Z'(\frac{\omega}{\sqrt{2}v_{th}k}),$$

and (3.47)

$$N(k,\omega) = g(k)\frac{n_0 e}{k^2\epsilon_0}\frac{1}{\sqrt{2}v_{th}}Z(\frac{1}{\sqrt{2\pi}v_{th}}).$$

It follows that

$$\frac{N(k,\omega)}{\frac{\partial D}{\partial\omega}(k,\omega)} = -2g(k)\frac{m}{e}kv_{th}^2\frac{Z(\frac{\omega}{\sqrt{2}v_{th}k})}{Z''(\frac{\omega}{\sqrt{2}v_{th}k})} = -g(k)\frac{m}{e}kv_{th}^2\frac{Z(\frac{\omega}{\sqrt{2}v_{th}k})}{2\frac{\omega}{\sqrt{2}v_{th}k} - (1-\frac{\omega^2}{v_{th}^2k^2})Z(\frac{\omega}{\sqrt{2}v_{th}k})}. \tag{4.1}$$

Let us write down a table with the roots of $D(k,\omega) = 0$ with largest imaginary part and the corresponding value of $N(k,\omega_j)/\frac{\partial D}{\partial\omega}(k,\omega_j)$ for several values of k:

k	ω_j	$N(k,\omega_j)/\frac{\partial D}{\partial\omega}(k,\omega_j)$
0.5	$\pm 1.4156 - 0.1533i$	$0.3677\, e^{\pm i 0.536245}$
0.4	$\pm 1.2850 - 0.0661i$	$0.424666\, e^{\pm i 0.3357725}$
0.3	$\pm 1.1598 - 0.0126i$	$0.63678\, e^{\pm i 0.114267}$
0.2	$\pm 1.0640 - 5.510\times 10^{-5}i$	$1.129664\, e^{\pm i 0.00127377}$

We denote by $\omega_r = \Re(\omega_j)$, $\omega_i = \Im(\omega_j)$, r the amplitude of $N(k,\omega_j)/\frac{\partial D}{\partial\omega}(k,\omega_j)$ and φ its phase. Note that we always have a root of the form $\omega_r + i\omega_j$ associated to $re^{i\varphi}$, and a root of the form $-\omega_r + i\omega_j$ associated to $re^{-i\varphi}$. So considering only the two root for which ω_i is the largest, we have

$$\begin{aligned}\hat{E}(k,t) &\approx re^{i\varphi}e^{-i(\omega_r+i\omega_i)t} + re^{-i\varphi}e^{-i(-\omega_r+i\omega_i)t},\\ &= re^{\omega_i t}(e^{-i(\omega_r t-\varphi)} + e^{i(\omega_r t-\varphi)}),\\ &= 2re^{\omega_i t}\cos(\omega_r t-\varphi).\end{aligned}$$

Then, we can verify that $\hat{E}(-k,t) = -\hat{E}(k,t)$, so that

$$E(x,t) \approx 4\epsilon re^{\omega_i t}\sin(kx)\cos(\omega_r t-\varphi).$$

Note that this is not the full solution, as we have only considered the dominating mode. However, as the others are decaying much faster and become negligible, this is a very good approximation of E after a short time.

Let us apply this formula by considering for example the first line of the table. The electric field for $k = 0,5$ will quickly, after the terms associated to the non dominating roots have decayed enough, be of the form

$$E(x,t) = 4\epsilon\times 0.3677e^{-0.1533t}\sin(0.5x)\cos(1.4156t-0.536245).$$

We overlap in Figure 4.3 the analytical solution including the dominating mode only of the electric field for $k = 0.4$ ($\hat{E}(k,t) = 0.002\times 0.424666e^{0.0661t}\cos(1.2850t-0.3357725)$) and the solution computed numerically by a semi-Lagrangian method with a grid of 128 points in x and in v. The v mesh has be truncated to the interval $[-10,10]$.

For the same numerical parameters, we display in Figure 4.4 the numerical solution and the slope corresponding to the damping rate, rather that the full

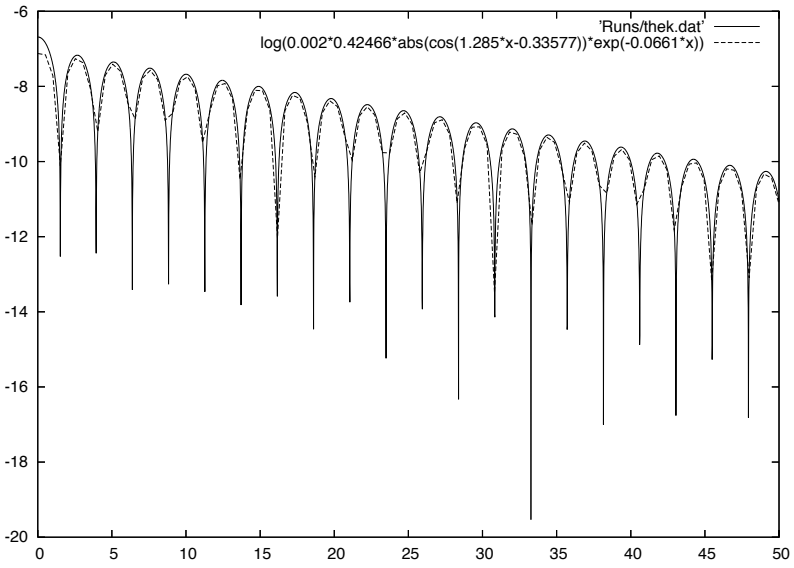

Fig. 4.3. Landau damping for $k = 0.4$.

solution. We notice that for a time around 80, the damping stops and the electric field comes almost back to its initial amplitude. This is a purely numerical phenomenon linked to the use of a uniform velocity mesh for a problem, which is periodic in space. This phenomenon is known as the Poincaré recurrence. In our case, it can be analysed by considering the free streaming equation

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

The domain being periodic in x, we take the Fourier series. Mode k verifies then $\frac{df_k}{dt} - \frac{2i\pi k}{L} v f_k = 0$. Whence $f_k(v,t) = f_k(v,0)e^{\frac{2i\pi k}{L}vt}$. And as $v = j\Delta v$, f_k will be periodic in t of period $T_R = \frac{L}{\Delta v}$. In our case, we have $L = 4\pi$ and $\Delta v = 20/127 = 0.1574$. It follows that $T_R \approx 79.8$, which corresponds to the observed value.

4.4.3 The two stream instability

The two stream instability corresponds to two streams of different mean velocities, which can become linearly unstable. We consider here mean velocities v_0 and $-v_0$. Depending on k and v_0 this configuration can be stable or unstable.

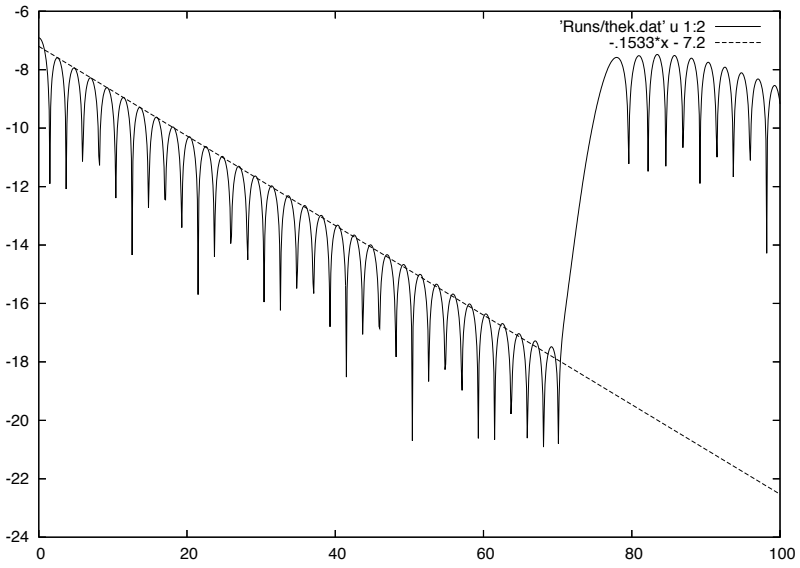

Fig. 4.4. Landau damping for $k = 0.5$.

The initial condition corresponding to this problem

$$f_0(x,v) = (1 + 0.001\cos(kx))\frac{1}{2\sqrt{2\pi}}(e^{-\frac{(v-v0)^2}{2}} + e^{-\frac{(v+v0)^2}{2}}), \quad L = \frac{2\pi}{k}.$$

Let us use the dispersion relation fro many beams (3.49) in the case of two beams of the same thermal velocity and with opposite mean velocities. It then becomes

$$D(k,\omega) = 1 + \frac{\omega_p^2}{2k^2v_{th}^2}\left[2 + \sqrt{\frac{\pi}{2}}\left(\frac{\omega}{v_{th}k} - \frac{v0}{v_{th}}\right)e^{-\frac{(\frac{\omega}{k}-v_0)^2}{2v_{th}^2}}(i - \text{erfi}\,(\frac{\frac{\omega}{k}-v_0}{\sqrt{2}v_{th}}) \right.$$
$$\left. + \sqrt{\frac{\pi}{2}}\left(\frac{\omega}{v_{th}k} + \frac{v0}{v_{th}}\right)e^{-\frac{(\frac{\omega}{k}+v_0)^2}{2v_{th}^2}}(i - \text{erfi}\,(\frac{\frac{\omega}{k}+v_0}{\sqrt{2}v_{th}})\right],$$

For given values of k and v_0, $D(k,\omega)$ admits several roots in ω. We display the dominant roots for $k = 0.2$ and several values of v_0 in the table 4.1 . Depending on the value of v_0 the instability is more or less strong, and for some values the configuration is stable.

We notice that the stable solutions oscillate with a very light damping and that the unstable roots correspond to purely imaginary roots and do not oscillate.

v_0	ω	ω
1.3	$0.02115i$	$1.1648 - 0.00104i$
2.4	$0.2258i$	$1.3390 - 0.00242i$
3.0	$0.2845i$	$1.446 - 0.00299i$

Table 4.1. Roots of the dispersion relation for the two-stream instability.

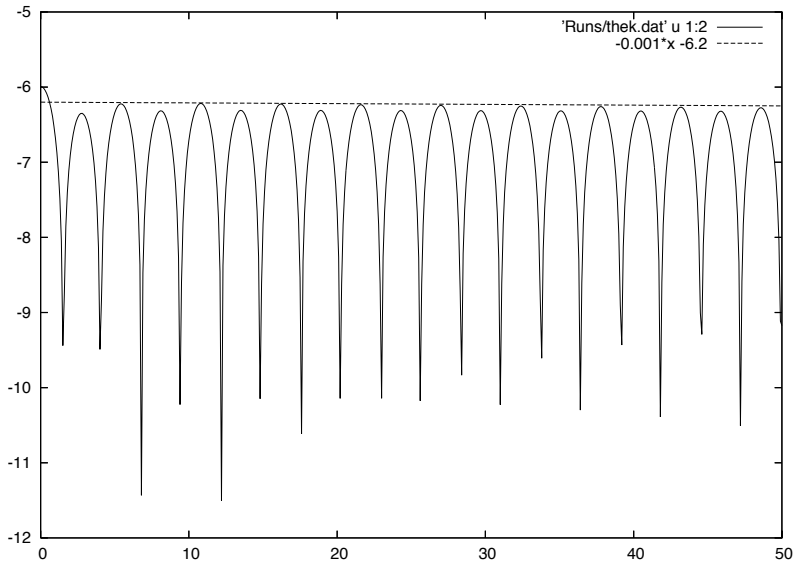

Fig. 4.5. Two-stream instability for $k = 0.2$ and $v_0 = 1.3$.

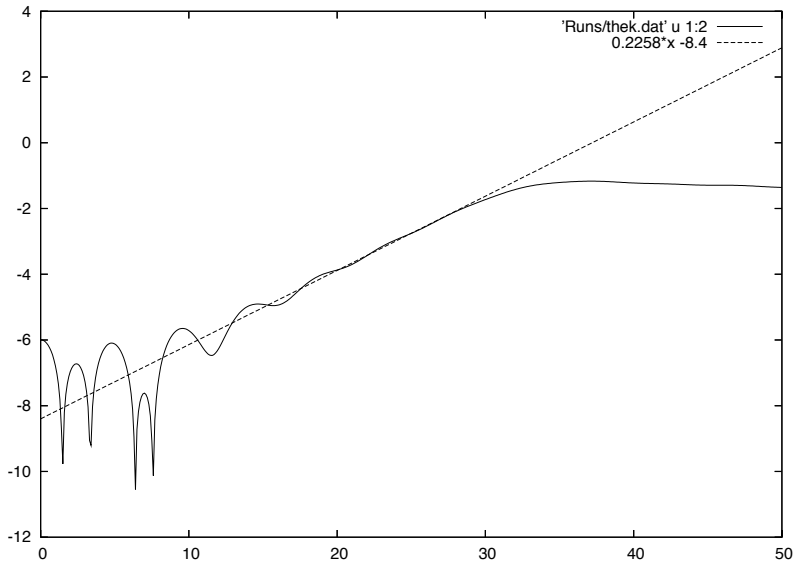

Fig. 4.6. Two-stream instability for $k = 0.2$ and $v_0 = 2.4$.

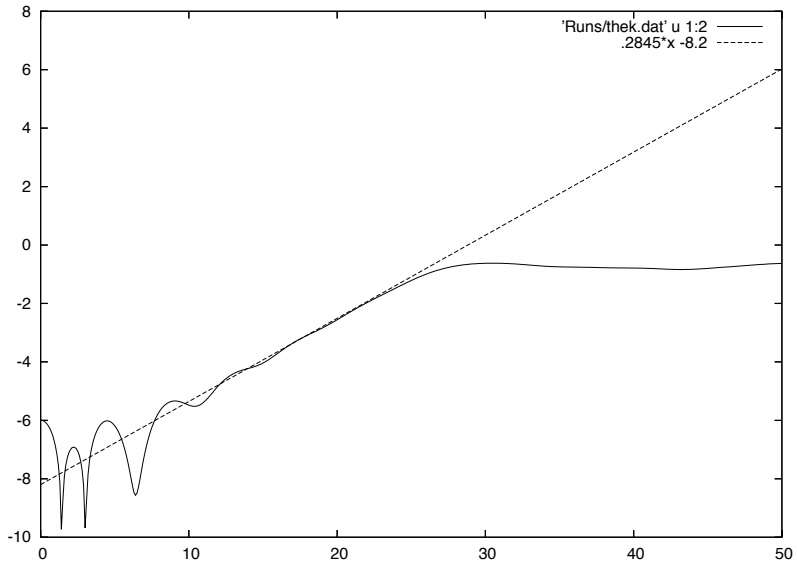

Fig. 4.7. Two-stream instability for $k = 0.2$ and $v_0 = 3$.

5

Basic numerical tools

5.1 Operator splitting

In the Vlasov equation without a magnetic field, the advection field in $\mathbf{x}$, which is $\mathbf{v}$, does not depend on $\mathbf{x}$ and the advection field in $\mathbf{v}$, which is $\mathbf{E}(\mathbf{x}, t)$, does not depend on $\mathbf{x}$. Therefore it is often convenient to decompose these two parts, using the technique called *operator splitting*.

Let us consider the non relativistic Vlasov-Poisson equation which reads

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_x f + \frac{q}{m} \mathbf{E} \cdot \nabla_v f = 0,$$

coupled with the Poisson equation $-\Delta\phi = 1 - \rho(t, \mathbf{x}) = 1 - \int f(t, \mathbf{x}, \mathbf{v})\, d\mathbf{v}$, $\mathbf{E}(\mathbf{x}, t) = -\nabla\phi$. Throught this coupling, $\mathbf{E}$ depends on f, which makes the Vlasov-Poisson system non linear.

We shall split the equation into the following two pieces:

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_x f = 0, \tag{5.1}$$

with $\mathbf{v}$ fixed and

$$\frac{\partial f}{\partial t} + \frac{q}{m} \mathbf{E}(\mathbf{x}, t) \cdot \nabla_v f = 0, \tag{5.2}$$

with $\mathbf{x}$ fixed. We then get two constant coefficient advections that can are easier to solve. This is obvious for (5.1) as $\mathbf{v}$ does not depend on t and x. On the other hand, integrating (5.2) with respect to $\mathbf{v}$, we get that $\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial t} \int f(t, \mathbf{x}, \mathbf{v})\, d\mathbf{v} = 0$, so that ρ and consequently $\mathbf{E}$ does not change when this equation is advanced in time. So that $\mathbf{E}(t, x)$ needs to be computed with the initial f for this equation and does then depend neither on t, nor $\mathbf{x}$.

Remark 5 *When the starting equation has some features which are important for the quality of the numerical solution, it is essential not to remove then when doing operator splitting. In particular, if the initial equation is conservative, it is generally a good idea to split such that each of the split equation is conservative.*

In order to analyze the error resulting from operator splitting, let us consider the following system of equations

$$\frac{du}{dt} = (A + B)u, \tag{5.3}$$

where A and B are any two differential operators (in space), that are assumed constant between t_n and t_{n+1}. The formal solution of this equation on one time step reads:

$$u(t + \Delta t) = e^{\Delta t(A+B)}u(t).$$

Let us split the equation (5.3) into

$$\frac{du}{dt} = Au, \tag{5.4}$$

$$\frac{du}{dt} = Bu. \tag{5.5}$$

The formal solutions of these equations taken separately are

$$u(t + \Delta t) = e^{\Delta tA}u(t) \text{ and } u(t + \Delta t) = e^{\Delta tB}u(t).$$

The standard operator splitting method consists in solving successively on one time step first (5.4) and then (5.5). Then one gets on one time step

$$\tilde{u}(t + \Delta t) = e^{\Delta tB}e^{\Delta tA}u(t).$$

If the operators A and B commute $e^{\Delta tB}e^{\Delta tA} = e^{\Delta t(A+B)}$ and the splitting is exact. This is the case in particular when considering a constant coefficient advection equation of the form

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = 0.$$

This can be checked using the method of characteristics. Note that such an equation is also a good first test case to validate a Vlasov code.

In the case when A and B do not commute, the splitting error can be decreased by solving first (5.4) on a half time step, and then (5.5) on a full time step and again (5.4) on a half time step. This method is known as the *Strang splitting method*. It corresponds to the formal solution

$$\bar{u}(t + \Delta t) = e^{\frac{\Delta t}{2}A}e^{\Delta tB}e^{\frac{\Delta t}{2}A}u(t).$$

The error committed at each time step by the operator splitting method when the operators do not commute is given by

Proposition 8 • *The standard splitting method is of order 1 in time.*
• *The Strang splitting method is of order 2 in time.*

Proof. In order to find the error we need to expand the matrix exponential. On the one hand we have

$$e^{\Delta t(A+B)} = I + \Delta t(A+B) + \frac{\Delta t^2}{2}(A+B)^2 + O(\Delta t^3),$$

and on the other hand

$$\begin{aligned} e^{\Delta tB}e^{\Delta tA} &= (I + \Delta tB + \frac{\Delta t^2}{2}B^2 + O(\Delta t^3))(I + \Delta tA + \frac{\Delta t^2}{2}A^2 + O(\Delta t^3)) \\ &= I + \Delta t(A+B)\frac{\Delta t^2}{2}(A^2 + B^2 + 2BA) + O(\Delta t^3)). \end{aligned}$$

But as A and B do not commute, we have $(A+B)^2 = A^2 + AB + BA + B^2$. It follows that $e^{\Delta t(A+B)} - e^{\Delta tB}e^{\Delta tA} = O(\Delta t^2)$, which leads to a local error of order 2 and a global error of order 1.

For the Strang splitting method, we have

$$\begin{aligned} e^{\frac{\Delta t}{2}A}e^{\Delta tB}e^{\frac{\Delta t}{2}A} =& (I + \frac{\Delta t}{2}A + \frac{\Delta t^2}{4}A^2 + O(\Delta t^3)))(I + \Delta tB + \frac{\Delta t^2}{2}B^2 + O(\Delta t^3)) \\ & (I + \frac{\Delta t}{2}A + \frac{\Delta t^2}{4}A^2 + O(\Delta t^3))) \\ =& I + \Delta t(A+B) + \frac{\Delta t^2}{2}(A^2 + B^2 + BA + AB) + O(\Delta t^3). \end{aligned}$$

We thus obtain a local error of order 3 and thus a global error of order 2 for the method of Strang.

Remark 6 *It is possible to obtain splitting methods of order as high as desired by taking adequate compositions of the two operators. Details on high order splitting methods can be found in [107].*

Remark 7 *The Strang splitting method can also be generalized to more that two operators. If $A = A_1 + \cdots + A_n$, the following decomposition will be of global order 2:*

$$e^{\frac{\Delta t}{2}A_1} \ldots e^{\frac{\Delta t}{2}A_{n-1}} e^{\Delta tA_n} e^{\frac{\Delta t}{2}A_{n-1}} \ldots e^{\frac{\Delta t}{2}A_1}.$$

5.2 Discrete Fourier Transform

5.2.1 Definition

Let P be the symmetric matrix formed with the powers of the n^{th} roots of unity the coefficients of which are given by $P_{jk} = \frac{1}{\sqrt{n}}e^{\frac{2i\pi jk}{n}}$. Denoting by $\omega_n = e^{\frac{2i\pi}{n}}$, we have $P_{jk} = \frac{1}{\sqrt{n}}\omega_n^{jk}$.

Notice that the columns of P, denoted by P_i, $0 \leq i \leq n-1$ are the vectors X_i normalized so that $P_i^* P_j = \delta_{i,j}$. On the other hand the vector X_k

corresponds to a discretization of the function $x \mapsto e^{-2i\pi kx}$ at the grid points $x_j = j/n$ of the interval $[0, 1]$. So the expression of a periodic function in the base of the vectors X_k is thus naturelly associated to the Fourier series of a periodic function.

Definition 2 (Discrete Fourier Transform (DFT)) .

- *The **discrete Fourier transform (DFT)** of a vector $x \in \mathbb{C}^n$ is the vector $y = P^* x$.*
- *La **inverse discrete Fourier transform** of a vector $y \in \mathbb{C}^n$ is the vector $x = P^{*-1}x = Px$.*

Lemma 1 *The matrix P is unitary and symmetric,* i.e. $P^{-1} = P^* = \bar{P}$.

Proof. We clearly have $P^T = P$, so $P^* = \bar{P}$. There remains to prove that $P\bar{P} = I$. But we have

$$(P\bar{P})_{jk} = \frac{1}{n}\sum_{l=0}^{n-1} \omega^{jl}\omega^{-lk} = \frac{1}{n}\sum_{l=0}^{n-1} e^{\frac{2i\pi}{n}l(j-k)} = \frac{1}{n}\frac{1 - e^{\frac{2i\pi}{n}n(j-k)}}{1 - e^{\frac{2i\pi}{n}(j-k)}},$$

and so $(P\bar{P})_{jk} = 0$ si $j \neq k$ and $(P\bar{P})_{jk} = 1$ if $j = k$.

Corollary 1 *Let $F, G \in \mathbb{C}^n$ and denote by $\hat{F} = P^*F$ and $\hat{G} = P^*G$, their discrete Fourier transforms. Then we have*

- *the discrete Parseval identity:*

$$(F, G) = F^T\bar{G} = \hat{F}^T\bar{\hat{G}} = (\hat{F}, \hat{G}), \tag{5.6}$$

- *The discrete Plancherel identity:*

$$\|F\| = \|\hat{F}\|, \tag{5.7}$$

where $(.,.)$ and $\|.\|$ denote the usual euclidian dot product and norm in $\mathbb{C}^n$.

Proof. The dot product in $\mathbb{C}^n$ of $F = (f_1, \ldots, g_n)^T$ and $G = (g_1, \ldots, g_n)^T$ is defined by

$$(F, G) = \sum_{i=1}^{N} f_i\bar{g}_i = F^T\bar{G}.$$

The using the definition of the inverse discrete Fourier transform, we have $F = P\hat{F}$, $G = P\hat{G}$, we get

$$F^T\bar{G} = (P\hat{F})^T\overline{P\hat{G}} = \hat{F}^T P^T\bar{P}\bar{\hat{G}} = \hat{F}^T\bar{\hat{G}},$$

as $P^T = P$ and $\bar{P} = P^{-1}$. The Plancherel identity follows from the Parseval identity by taking $G = F$.

Remark 8 *The discrete Fourier transform is defined as a matrix-vector multiplication. Its computation hence requires a priori n^2 multiplications and additions. But because of the specific structure of the matrix there exists a very fast algorithm, called Fast Fourier Transform (FFT) for performing it in $O(n \log n)$ operations. This makes it particularly interesting for many applications, and many fast PDE solvers make use of it.*

5.2.2 Approximation of the coefficients of a Fourier series using the FFT

A $L-$periodic function f can be expressed by its Fourier series. More precisely the classical Dirichlet theorem states that if f is C^1 its Fourier series converges uniformly towards $f(x)$ pointwise for any x, where the Fourier series is defined by

$$f(x) = \sum_{k=-\infty}^{+\infty} \hat{f}_k e^{i\frac{k2\pi}{L}x},$$

where the Fourier coefficients c_k are defined by

$$\hat{f}_k = \frac{1}{L}\int_0^L f(x)e^{-i\frac{k2\pi}{L}x}\,dx.$$

In order to compute a numerical approximation of the Fourier coeffients, we define a mesh with N points on one period $[0, L[$ such that $x_j = jL/N$, $0 \leq j \leq N-1$. We denote by $f_j = f(x_j)$. We then have

$$\hat{f}_k = \frac{1}{L}\sum_{j=0}^{N-1}\int_{x_j}^{x_{j+1}} f(x)e^{-i\frac{k2\pi}{L}x}\,dx.$$

As f is known only at the grid points the integral is approximated by the trapezoidal rule on each grid interval. Note that due to the Euler-MacLaurin formula the composed trapezoidal rule is very accurate for periodic function. More precisely, if $f \in C^{2p}([0, L])$ and $L-$periodic, the composed trapezoidal rule is of order $2p$. We then get

$$\hat{f}_k \approx \frac{1}{L}\frac{L}{N}\sum_{j=0}^{N} f_j e^{-i\frac{k2\pi}{L}\frac{jL}{N}} = \frac{1}{N}\sum_{j=0}^{N} f_j e^{-i\frac{k2\pi j}{N}}.$$

Such a numerical approximation involves a sampling of the initial function at N points x_j. Because of this sampling some information on the initial function is lost and the Fourier series only contains N distinct values $\hat{f}_k$. Indeed, for any $k \in \mathbb{Z}$ we have

$$\hat{f}_{k+N} = \frac{1}{N}\sum_{j=0}^{N} f_j e^{-i\frac{(k+N)2\pi j}{N}} = \hat{f}_k.$$

These N distinct values approximate $\hat{f}_k$ for $-N/2 \leq k \leq N/2-1$. The other modes are not represented by the discrete Fourier tranform. Notice that the corresponding frequencies $\omega = \frac{2\pi k}{L}$ lie in the interval $[-\pi/L, \pi/L[$.

Note that the discrete Fourier transform gives $\hat{f}_{k+N}$ for $0 \leq k \leq N-1$. In order to use it for approximating Fourier series, we use the N-periodicity of the coefficients to define $\hat{f}_k$ for $-N/2 \leq k < 0$ from $\hat{f}_k$ for $N/2 \leq k \leq N-1$. Matlab and other numerical software provide the funciton `fftshift` to transfer the $N/2-1$ last modes provided by the FFT to the beginning of the array.

5.3 Circulant matrices

Definition 3 *A matrix of the form*

$$M = \begin{pmatrix} c_0 & c_1 & c_2 & \dots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & & c_{n-3} \\ \vdots & & & \ddots & \vdots \\ c_1 & c_2 & c_3 & \dots & c_0 \end{pmatrix}$$

with $c_0, c_1, \dots, c_{n-1} \in \mathbb{R}$ *is called* circulant.

Proposition 9 *The eigenvalues of the circulant matrix* M *are given by*

$$\lambda_k = \sum_{j=0}^{n-1} c_j \omega^{jk}, \tag{5.8}$$

where $\omega = e^{2i\pi/n}$.

Proof. Let J be the circulant matrix obtained from M by taking $c_1 = 1$ and $c_j = 0$ for $j \neq 1$. We notice that M can be written as a polynomial in J

$$M = \sum_{j=0}^{n-1} c_j J^j.$$

As $J^n = I$, the eigenvalues of J are the n-th roots of unity that are given by $\omega^k = e^{2ik\pi/n}$. Looking for X_k such that $JX_k = \omega^k X_k$ we find that an eigenvector associated to the eigenvalue λ_k is

$$X_k = \begin{pmatrix} 1 \\ \omega^k \\ \omega^{2k} \\ \vdots \\ \omega^{(n-1)k} \end{pmatrix}.$$

We then have that

$$MX_k = \sum_{j=0}^{n-1} c_j J^j X_k = \sum_{j=0}^{n-1} c_j \omega^{jk} X_k,$$

and so the eigenvalues of M associated to the eigenvectors X_k are

$$\lambda_k = \sum_{j=0}^{n-1} c_j \omega^{jk}.$$

Proposition 10 *Any circulant matrix C can be written in the form $C = P\Lambda P^*$ where P is the matrix of the discrete Fourier transform and Λ is the diagonal matrix of the eigenvalues of C. In particular all circulant matrices have the same eigenvectors (which are the columns of P), and any matrix of the form $P\Lambda P^*$ is circulant.*

Corollary 2 *We have the following properties:*

- *The product of two circulant matrix is circulant matrix.*
- *A circulant matrix the eigenvalues of which are all non vanishing is invertible and its inverse is circulant.*

Proof. The key point is that all circulant matrices can be diagonalized in the same basis of eigenvectors. If C_1 and C_2 are two circulant matrices, we have $C_1 = P\Lambda_1 P^*$ and $C_2 = P\Lambda_2 P^*$ so $C_1 C_2 = P\Lambda_1 \Lambda_2 P^*$.

If all eigenvalues of $C = P\Lambda P^*$ are non vanishing, Λ^{-1} is well defined and $P\Lambda P^* P\Lambda^{-1} P^* = I$. So the inverse of C is the circulant matrix $P\Lambda^{-1} P^*$.

Because of the fact that all circulant matrices diagonalise in the Fourier basis, the FFT provides a fast and convenient tool for solving linear systems or performing products involving circulant matrices. In particular performing spline interpolation at a constant displacement from each grid point can be writen in matrix form as: $MC = F^n$ (compute spline coefficients from point values), then $F^{n+1} = DC$ where M and D are circulant matrices. Combining both relations we obtain $F^{n+1} = DM^{-1}F^n$. Denoting λ_D and λ_M the diagonal matrices of the eigenvalues of respectively D and M that can be computing easily with formula (5.8) using the coefficients of the circulant matrix. Indeed we have

$$F^{n+1} = P\Lambda_D \lambda_M^{-1} P^* F^n,$$

And multiplying by P^* consists in performing a Discrete Fourier Tranform. So that the algorithm for the computation becomes:

1. $\hat{F}^n = FFT(F^n)$,
2. $\hat{G}^n_j = \hat{F}^n \lambda_{D,j} / \lambda_{M,j}$ for $i = 0, n-1$,
3. $F^{n+1} = iFFT(\hat{G}^n)$, where $iFFT$ denotes an inverse FFT.

5.4 Interpolation

One of the main buiding blocks of a semi-Lagrangian method is interpolation. In order for the method not to be too diffusive it is important to use a good interpolation method which is accurate enough. Typically, linear interpolation is way too diffusive. The method of choice in many semi-Lagrangian codes is cubic splines which proves very robust and accurate.

5.4.1 Splines

General definition

Consider a 1D grid of an interval $[a, b]$ $a = x_1 < x_2 < ... < x_N = b$. We define $\mathcal{S}^p(a, b)$ the linear space of splines of degree p on $[a, b]$ by

$$\mathcal{S}^p(a,b)\{S \in C^{p-1}([a,b]) \mid S_{|[x_i,x_{i+1}]} \in \mathbb{P}^p([x_i, x_{i+1}])\},$$

where $\mathbb{P}^p([x_i, x_{i+1}])$ denotes the space of polynomials of degree p on $[x_i, x_{i+1}]$, which is of dimension is $p + 1$.

Let us first consider periodic splines which are very usefull for our applications as the special case. In this case we consider periodic functions of period $b - a$. These functions take the same value at a and b so that x_1 and x_N can be considered the same to be the same grid point. Let us compute the dimention of $\mathcal{S}^p$ in this case. $\mathcal{S}^p$ is included in the space of $N - 1$ piecewise polynomials of dimension $(N - 1)(p + 1)$. Its dimension is reduced by the continuity requirements on the spline and its derivatives at each grid point, which is altogether $(N-1)p$. So that the dimension of $\mathcal{S}^p$ is $N-1$ for periodic splines. In this case a spline can be determined by an interpolation condition at each grid point.

Now on a regular interval this is slightly modified. Indeed the spline is still a subspace of the space of $N - 1$ piecewise polynomials of dimension $(N - 1)(p + 1)$, but now the continuity requirements are only at the $N - 2$ interior points. So that the dimension of $\mathcal{S}^p$ in this case is $(N - 1)(p + 1) - (N - 2)p = N + p - 1$. This is larger than N for $p \geq 2$ so that boundary conditions are needed in addition to the interpolation conditions at the grid points to uniquely determine the spline. A classical boundary condition is Hermite boundary conditions, which state that all derivatives up to order $(p - 1)/2$ are given at each of the two boundaries of the interval for odd degree splines that are used in practise for interpolation.

Following the arguments to compute the dimension of the spline space, a natural way of computing the spline would be to compute the local polynomials $a_i^p x^p + a_i^{p-1} x^{p-1} + \cdots + a_i^1 x + a_i^0$ on each interval i, $1 \leq i \leq N - 1$, using the interpolation values at the grid points and the continuity relations.

This can be used in practise to compute spline interpolations but it is in general more efficient to use a set of basis functions called $B - splines$.

B-splines

We define $B-splines$ as follows: Let $T = (t_i)_{1 \leqslant i \leqslant N+k}$ be a non-decreasing sequence points. In the splines jargon these points are called knots. This is more general than standard spline interpolation as considered previously. In particular repeated knots can be considered.

Definition 4 (B-Spline) *The i-th B-Spline of degree p is defined by the recurrence relation:*

$$N_j^{p+1} = w_j^{p+1} N_j^p + (1 - w_{j+1}^{p+1}) N_{j+1}^p \tag{5.9}$$

where,

$$w_j^{p+1}(x) = \frac{x - t_j}{t_{j+p} - t_j} \qquad\qquad N_j^0(x) = \chi_{[t_j, t_{j+1}[}(x)$$

We note some important properties of a B-splines basis:

- B-splines are piecewise polynomial of degree p.
- Positivity: $N_j^p(x) \geq 0$ for all x.
- Compact support; the support of N_j^{p+1} is contained in $[t_j, .., t_{j+k}]$.
- Partition of unity : $\sum_{i=1}^{N} N_i^p(x) = 1, \forall x \in \mathbb{R}$
- Local linear independence.
- If a knot t has a multiplicity m then the B-spline is $\mathcal{C}^{(p-m)}$ at t.

The derivative of a B-spline of degree p can be computed as a simple difference of B-splines of degree $p-1$

$$N_i^{p\prime}(x) = p \left(\frac{N_i^{p-1}(x)}{t_{i+p} - t_i} - \frac{N_{i+1}^{p-1}(x)}{t_{i+p+1} - t_{i+1}} \right). \tag{5.10}$$

An important special case is the case of uniformly spaced knots on an infinite or periodic grid. In this case the splines are often called cardinal splines and all the B-splines are translates of each other, so that the basis can be defined with only one element denoted by N^p for the degree p, if h is the spacing between successive knots then the full basis is composed of the translates $(N^p(\cdot - jh))_{j \in I}$, where the index set is $I = \mathbb{Z}$ or a finite subset of $\mathbb{Z}$ in the periodic case. The cardinal splines are generally defined with the integers as knots. In this case formula (5.9) becomes

$$N^{p+1}(x) = \frac{x}{p} N^p(x) + \frac{p + 1 - x}{p} N^p(x - 1), \tag{5.11}$$

with $N^0(x) = 1$ if $x \in [0, 1[$ and 0 else. It easily follows that the support of N^p is $[0, p+1]$.

Remark 9 *B-Splines N_h^p on the uniform grid jh, $j \in \mathbb{Z}$ verify $N_h^p(x) = N^p(x/h)$. It is thus sufficient to define B-splines on integer knots.*

Note that in addition to the general properties of the splines, the cardinal splines also verify

$$N^{p+1}(x) = \int_0^1 N^p(x-t)\,dt$$

and

$$N^p(\frac{p+1}{2}+x) = N^p(\frac{p+1}{2}-x) \quad \forall x \in \mathbb{R}.$$

See the book of Chui [34] for proofs of these properties and more on cardinal splines.

Using this properties we can prove the following lemma on the first moment of a cardinal spline. Similar properties can also be obtained for higher order moments [100].

Lemma 2 *For all $x \in \mathbb{R}$, if N^p is the cardinal spline of degree p we have*

$$\sum_j (j-x)N^p(j-x) = \int_0^{p+1} tN^p(t)\,dt =: M^p.$$

In other words, the sum is independent on x and is equal to the moment of the cardinal spline that we denote by M^p.

Proof. Let us first denote for any given p $M^p(x) = \sum_j (j-x)N^p(j-x)$. Using (5.11) we have

$$\begin{aligned} p\sum_j (j-x)N^{p+1}(j-x) &= \sum_j (j-x)^2 N^p(j-x) \\ &\quad + \sum_j (j-x)(p+1-j+x)N^p(j-x-1) \\ &= \sum_j (j-x)^2 N^p(j-x) \\ &\quad + \sum_j (j+1-x)(p-j+x)N^p(j-x) \end{aligned}$$

making a change of index in the last sum. Then combining both sums

$$\begin{aligned} p\sum_j (j-x)N^{p+1}(j-x) &= p\sum_j N^p(j-x) + (p-1)\sum_j (j-x)N^p(j-x) \\ &= p + (p-1)\sum_j (j-x)N^p(j-x), \end{aligned}$$

due to the partition of unity property. So that we get the recurrence relation

$$M^{p+1}(x) = 1 + \frac{p-1}{p}M^p(x).$$

For $p = 1$, $M^1(x)$ involves only two non vanishing terms. Denoting by $\lfloor x \rfloor$ the floor of x, *i.e.* the greatest integer smaller than x, only the terms corresponding to $j = \lfloor x \rfloor + 1$ and $j = \lfloor x \rfloor + 2$ do not vanish in the sum. Then denoting by $\alpha = x - \lfloor x \rfloor$ the fractional part of x, we have $0 \leq \alpha \leq 1$ and

$$M^1(x) = (1-\alpha)N^1(1-\alpha) + (2-\alpha)N^1(2-\alpha) = (1-\alpha)^2 + (2-\alpha)\alpha = 1,$$

as $N^1(x) = x$ on $[0,1]$ and $N^1(x) = 2-x$ on $[1,2]$. So $M^1(x)$ does not depend on x and then by induction using the recurrence relation previously derived $M^p(x)$ does not depend on x, only on p.

On the other hand let us directly compute $M^p = \int_0^{p+1} tN^p(t)\,dt$. Using (5.11) we get

$$\begin{aligned}
pM^{p+1} &= \int_0^{p+2} xN^{p+1}(x)\,dt \\
&= \int_0^{p+1} x^2 N^p(x)\,dx + \int_1^{p+2} (p+1-x)xN^p(x-1)\,dx \\
&= \int_0^{p+1} x^2 N^p(x)\,dx + \int_0^{p+1} (p-x)(x+1)N^p(x)\,dx \\
&= \int_0^{p+1} (x^2 + (p-x)(x+1))N^p(x)\,dx \\
&= \int_0^{p+1} (p + (p-1)x)N^p(x)\,dx \\
&= p + (p-1)M^p,
\end{aligned}$$

using that $\int_0^{p+1} N^p(x)\,dx = 1$ and the definition of M_p, we hence get the same recurrence formula for M^p as we had for $M^p(x)$ it thus remains to check that $M^1 = 1$ to conclude. For this a straightforward computation yields

$$M^1 = \int_0^2 xN^1(x)\,dx = \int_0^1 x^2\,dx + \int_1^2 x(2-x)\,dx = \frac{1}{3} + \frac{2}{3} = 1.$$

5.4.2 Using B-splines for spline interpolation

The B-splines form a basis of the spline space $\mathcal{S}_N^p$. In the case of a periodic domain, we saw that the dimension of $\mathcal{S}_N^p$ is exactly the number of grid points. In this case the knots can be taken to be exactly the grid points. The situation is a little bit more complicated for a bounded interval, in which case it more knots than grid points are needed to define the B-splines that will generate $\mathcal{S}_N^p$. Two natural possibilities exist, the first one is to replicate the knots at the two extremities of the interval. This has the advantage that the spline is interpolatory at the boundary so that Dirichlet boundaries are easily handled. On the other hand, this solution has the drawback that the shape of the B-splines changes a both ends of the domain which might be unwanted in

particular if the grid is uniform so that the shape of the splines is the same for all the inner splines. In this case, another option, is to mirror the points close to the boundary.

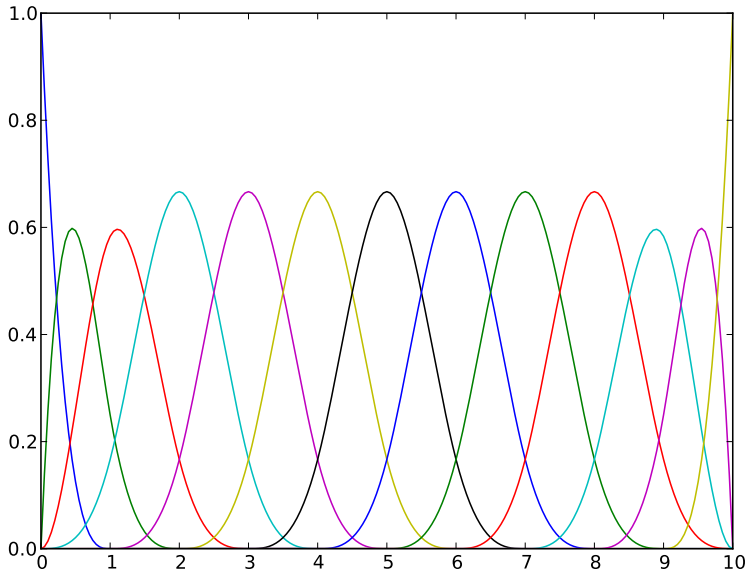

Fig. 5.1. Cubic splines with open boundary conditions with knots at integers

In any case let us denote $M = \dim \mathcal{S}_N^p$. Recall that $M = N-1$ for periodic splines and $M = N+p-1$ for bounded splines. Then a spline $S \in \mathcal{S}_N^p$ can be written, using the B-spline basis

$$S(x) = \sum_{i=1}^{M} c_i N_i^p(x).$$

In order to use this formula for interpolation we first need to determine the spline coefficients $(c_i)_{1\leq i\leq M}$. These can be determined by the interpolation conditions $S(x_k) = f(x_k)$ at the grid points and the boundary conditions in the case of non periodic splines.

5.4.3 Cubic spline interpolation

Let us consider a uniform mesh of the interval $[a, b]$ defined by $x_i = a + ih$, $i = 0, \ldots, N$, with $h = \frac{b-a}{N}$. Let $f \in C^k([a, b])$, $k \geq 0$. Its cubic spline

interpolant f_h on this mesh is defined by $f_h(x_i) = f(x_i)$ for $i = 0, \ldots, N$, $f_h \in \mathbb{P}_3([x_i, x_{i+1}])$ and $f_h \in C^2([a, b])$.

In the case of a periodic domain, i.e., if $[a, b]$ corresponds to one period of the periodic functions f and f_h, these conditions are sufficient to determine uniquely f_h. Else boundary conditions are needed, often Hermite type boundary conditions, consisting in giving the values of $f_h'(a)$ and of $f_h'(b)$ at the ends of the interval or the so-called natural boundary conditions, consisting in setting $f_h''(a) = f_h''(b) = 0$ are used.

It is convenient to have an expression of f_h using the cubic B-splines basis, which are the translations of the function S^3. Let us recall the expression of S^3 on our mesh.

$$S^3(x) = \frac{1}{6} \begin{cases} (2 - \frac{|x|}{h})^3 & \text{if } h \le |x| < 2h, \\ 4 - 6\left(\frac{x}{h}\right)^2 + 3\left(\frac{|x|}{h}\right)^3 & \text{if } 0 \le |x| < h, \\ 0 & \text{else.} \end{cases}$$

Let us first deal with the periodic case. We assume that all functions we consider are periodic of period $b - a$. Then in particular $f_h^{(p)}(a) = f_h^{(p)}(b)$ for $p = 0, 1, 2$. The point x_N of the mesh corresponds to the point x_0 and no additional value of the unknown is defined there.

The expression of f_h on the B-splines basis then reads

$$f_h(x) = \sum_{j=0}^{N-1} \alpha_j S^3(x - x_j),$$

and the coefficients α_i are determined by the interpolation conditions.

$$f(x_i) = f_h(x_i) = \sum_{j=0}^{N-1} \alpha_j S^3(x_i - x_j).$$

But $S^3(x_i - x_i) = \frac{2}{3}$, $S^3(x_i - x_{i+1}) = S^3(x_i - x_{i-1}) = \frac{1}{6}$ and $S^3(x_i - x_j) = 0$ if $|x_i - x_j| \ge 2h$.

We thus get a linear system with unknowns α_i, $i = 0, N - 1$:

$$\alpha_{i-1} + 4\alpha_i + \alpha_{i+1} = 6f(x_i), \quad 0 \le i \le N - 1,$$

with because of periodicity $\alpha_{-1} = \alpha_{N-1}$ and $\alpha_N = \alpha_0$. This system can be written in matrix form $A\alpha = b$ with

$$A = \begin{pmatrix} 4 & 1 & 0 & \ldots & 0 & 1 \\ 1 & 4 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 & 4 & 1 \\ 1 & 0 & \ldots & 0 & 1 & 4 \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{N-1} \end{pmatrix}, \quad b = 6h \begin{pmatrix} f(x_0) \\ \vdots \\ f(x_{N-1}) \end{pmatrix}.$$

As the matrix A is strictly diagonally dominant it is non singular and allows to determine the unknows α and hence the function f_h uniquely.

5.4.4 Fast local spline interpolation

Unser in [102] introduces a fast algorithm for computing spline coefficients on a uniform grid based on ideas from signal processing, which needs fewer operations than the traditional way of doing a Cholesky decomposition and does not require to store a Cholesky decomposition.

Let us introduce here this idea using simple linear algebra tools rather than the language of signal processing.

It has the potential of being faster as there is no need for storing and reading the terms of the Cholesky decomposition, moreover a it is easy to express the algorithm using three or four terms (or more) recurrence relations with constant coefficients for vectorization or thread parallelization.

Periodic boundary conditions

Let us first consider a periodic uniform grid such that $x_i = ih$ $0 \leq i \leq N-1$.

The cubic spline interpolant of a function f on this grid is defined by

$$S(x) = \sum_{i=0}^{N-1} c_i B(x - x_i),$$

where B is the cubic B-spline centered on 0. It is defined by having the value 1 at 0, and 0 at all the other grid points. Moreover it is a cubic polynomial on each cell and has two continuous derivatives. To define the spline interpolant, we only need to know that $B(0) = \frac{2}{3}$ and $B(h) = B(-h) = \frac{1}{6}$. Then the interpolation conditions give us the following system enabling to solve for the coefficients c_i. Let us denote by $C = (c_0, \ldots, c_{N-1})^T$, $F = (f(x_0), \ldots, f(x_{N-1})^T$ and A the symmetric tridiagonal circulant matrix having $2/3$ on the diagonal and $1/6$ on the upper and lower diagonals. The we have

$$AC = F.$$

$$A = \frac{1}{6}\begin{pmatrix} 4 & 1 & 0 & \ldots & 0 & 1 \\ 1 & 4 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & 1 & 4 & 1 \\ 1 & 0 & \ldots & 0 & 1 & 4 \end{pmatrix}, \qquad L = \begin{pmatrix} a & 0 & 0 & \ldots & 0 & b \\ b & a & 0 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & b & a & 0 \\ 0 & 0 & \ldots & 0 & b & a \end{pmatrix},$$

If we look for an expression of A as a product of LL^T, with L a lower circulant matrix with a on the diagonal and b on the lower diagonal, we find that it is possible provided

$$a^2 + b^2 = \frac{2}{3}, \quad \text{and } ab = \frac{1}{6},$$

which yields $b = \frac{1}{6a}$ and a solution of

$$36x^4 - 24x^2 + 1 = 0, \tag{5.12}$$

resulting in $a^2 = \frac{2+\sqrt{3}}{6}$, $b^2 = \frac{2-\sqrt{3}}{6}$ as we want $b < a$. We shall take the positive roots of these expressions for a and b.

Now, we can compute C as the solution of $LL^TC = F$. To this aim, we first solve for D such that $LD = F$. We first have

$$\begin{aligned} ad_0 &= f(x_0) - bd_{N-1} \\ &= f(x_0) - \frac{b}{a}(f(x_{N-1)} - bd_{N-2}) \\ &= f(x_0) - \frac{b}{a}f(x_{N-1)} + \left(\frac{b}{a}\right)^2 f(x_{N-2)} + \cdots + (-1)^i \left(\frac{b}{a}\right)^i (f(x_{N-i}) - bd_{N-i-1}). \end{aligned}$$

As $b/a < 1$, the series $\sum_{i\geq 0} b^i$ absolutely converges and can be approximated by its partial sum. Hence we can approximate d_0, for M large enough by

$$d_0 = \frac{1}{a}\left(f(x_0) + \sum_{i=1}^{M}(-1)^i \left(\frac{b}{a}\right)^i f(x_{N-i})\right).$$

We have $(b/a)^{10} = 1.90 \times 10^{-6}$ and $(b/a)^{25} = 5.03 \times 10^{-15}$, so that M should probably be taken between 10 and 25 depending on desired precision. Then for $i = 1, \ldots, N-1$, we get

$$d_i = \frac{1}{a}(f(x_i) - bd_{i-1}).$$

Once D is known, the same procedure can be used for the computation of C solution of $L^TC = D$. Here we approximate

$$c_{N-1} = \frac{1}{a}\left(d_{N-1} + \sum_{i=1}^{M}(-1)^i \left(\frac{b}{a}\right)^i d_{i-1}\right),$$

and for $i = N-2, \ldots, 0$

$$c_i = \frac{1}{a}(d_i - bc_{i+1}).$$

Hermite boundary conditions

Let us here consider the mesh $x_i = ih$ $0 \leq i \leq N-1$ and assume Hermite boundary conditions to close the cubic B-spline approximation, i.e. we assume that the approximating spline S verifies $S'(x_0) = S'(x_{N-1}) = 0$. In this case the spline approximation has $N+2$ terms on the B-spline basis and can be written

$$S(x) = \sum_{i=-1}^{N} c_i B(x - x_i).$$

The Hermite boundary conditions are handled in a different manner on the two ends of the interval. At x_0, the idea will be to use a finite sum as an approximation of a series of the same type as the one used for periodic boundary conditions. To this aim, in order to approximate a spline with vanishing derivative we extend the function and associated spline at the left of x_0 by symmetry, i.e. we define $f(x_{-i}) = f(x_i)$ for $i = 1, \ldots$. On the right hand side of the interval, we derive the expression of S. Then $S'(x_{N-1}) = 0$ yields $c_{N-2} = c_N$. On the other hand, the last interpolation condition reads

$$\frac{1}{6}c_{N-2} + \frac{2}{3}c_{N-1} + \frac{1}{6}c_N = f(x_{N-1}).$$

Replacing c_N by c_{N-2} and dividing by 2, we get

$$\frac{1}{6}c_{N-2} + \frac{1}{3}c_{N-1} = \frac{1}{2}f(x_{N-1}).$$

We can then write the system, in a semi-infinite form at the top. This means that the system is not closed at the top. We just assume that we have enough terms there to define the approximation of the series, we need. Then the spline coefficients $C = (\ldots, c_0, \ldots c_{N-1})^T$ can be computed from $F = (\ldots, f(x_0), \ldots, \frac{1}{2}f(x_{N-1}))^T$ by solving the semi-infinite system $AC = F$, and the matrix A can be written as $A = LL^T$:

$$A = \frac{1}{6}\begin{pmatrix} \ddots & \ddots & \ddots & \ldots & \ldots & 0 & 0 \\ \ddots & 4 & 1 & 0 & \ldots & 0 & 0 \\ \ddots & 1 & 4 & 1 & 0 & \ldots & 0 \\ & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ & \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ & 0 & & \ddots & 1 & 4 & 1 \\ & 0 & 0 & \ldots & 0 & 1 & 2 \end{pmatrix}, \qquad L = \begin{pmatrix} \ddots & \ddots & 0 & 0 & \ldots & 0 & 0 \\ \ddots & a & 0 & 0 & \ldots & 0 & 0 \\ \ddots & b & a & 0 & 0 & \ldots & 0 \\ & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ & \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ & 0 & & \ddots & b & a & 0 \\ & 0 & 0 & \ldots & 0 & b & \alpha \end{pmatrix},$$

A straightforward computation yields $a^2 + b^2 = \frac{2}{3}$, $ab = \frac{1}{6}$ and $b^2 + \alpha^2 = \frac{1}{3}$, such that, like for the periodic boundary conditions, $a = \sqrt{\frac{2+\sqrt{3}}{6}}$, $b = \sqrt{\frac{2-\sqrt{3}}{6}}$, $\alpha = \sqrt{\frac{\sqrt{3}}{6}}$.

The algorithm to compute the spline coefficients, will then be as follows. We first compute the coefficients of $D = (d_0, \ldots, d_{N-1})^T$ solution of $LD = F$. As in the periodic case, d_0 is computed as a series expansion, for M large enough, using the mirrored values of f ($f(x_{-i}) = f(x_i)$):

$$d_0 = \frac{1}{a}\left(f(x_0) + \sum_{i=1}^{M}\left(-\frac{b}{a}\right)^i f(x_i)\right). \tag{5.13}$$

Then

$$ad_i = f(x_i) - bd_{i-1} \quad \text{i=2,...N-2,} \qquad \alpha d_{N-1} = \frac{1}{2}f(x_{N-1}) - bd_{N-2}.$$

In a second stage C can be computed from D solving $L^T C = D$:

$$c_{N-1} = d_{N-1}/\alpha, \quad ac_i = d_i - bc_{i+1} \quad i = N-2, \ldots, 0.$$

5.4.5 Parallelization

In order to get a purely local algorithm, similar in spirit to our local hermite splines, we can use the formula (5.13) using the values from the subdomain on the left instead of mirrored values to start the iterations with d_0. On the other side one could either use the value for $S'(x_{N-1})$ given by the local splines or use the direct formula for the impulse response given by Unser.

This formula gives an approximation of the spline coefficient c_n

$$c_n = \sqrt{3}\left(f(x_n) + \sum_{i\geq 1}(\sqrt{3}-2)^i[f(x_{n-i}) + f(x_{n+i})]\right). \tag{5.14}$$

If x_n is the last point of a subdomain, this formula requires contribution from the subdomain as well as its neighbour that need to be precomputed and sent before starting the algorithm.

5.4.6 Lagrange interpolation

Lagrange interpolation, although dissipative at low order can be a good alternative to spline interpolation if a high enough order is used. In pratice odd degree Lagrange interpolation starting from degree 7 gives quite good results.

Let us recall how this can be implemented efficiently at arbitrary order. The Lagrange interpolation polynomial of degree N at points $x_0, \ldots, x_N$ of a smooth function f is defined by

$$p(x) = \sum_{j=0}^{N} f_j l_j(x),$$

where $f_j = f(x_j)$ and $l_j(x)$ is he j^{th} Lagrange polynomial of degree N uniquely defined by $l_j(x_i) = \delta_{ij}$, δ_{ij} being the Kronecker symbol which is 1 if $i = j$ and 0 else. The explicit formula for $l_j(x)$ is

$$l_j(x) = \frac{\prod_{i=0, i\neq j}^{N} (x - x_i)}{\prod_{i=0, i\neq j}^{N} (x_j - x_i)}.$$

Computing the Lagrange polynomials for Lagrange interpolation is not very convenient as it involves $O(n)$ multiplications and sums for each point to be interpolated and needs to be started anew when an interpolation point is added. In order to simplify this we introduce the function

$$\omega(x) = \prod_{i=0}^{N}(x - x_i), \text{ and } w_j = \frac{1}{\omega'(x)} = \frac{1}{\prod_{i=0, i\neq j}^{N} (x_j - x_i)}.$$

The the Lagrange interpolating polynomial can be written

$$p(x) = \omega(x)(\sum_{j=0}^{N} f_j \frac{w_j}{x - x_j}).$$

And as the interpolation is exact for $f = 1$, we get an expression for $\omega(x)$

$$1 = \omega(x)(\sum_{j=0}^{N} \frac{w_j}{x - x_j}),$$

so that we get the following simple formula that is convenient and efficient for Lagrange interpolation as the coefficients w_j need to be computed only once for all interpolation points:

$$p(x) = \frac{\sum_{j=0}^{N} f_j \frac{w_j}{x - x_j}}{\sum_{j=0}^{N} \frac{w_j}{x - x_j}}.$$

This is called the barycentric interpolation formula. See the review article by Beirut and Trefethen [10] for further information.

6

Numerical methods for the Maxwell equation

6.1 Spectral method for the Poisson equation

For the approximationg a linear PDE with constant coefficients on a periodic domain the FFT is the simplest and often fastest method. If the solution is smooth it provides moreover spectral convergence, which means that it converges faster than a polynomial approximation of any order, so that very good accuracy can be obtained with relatively few points. The exact number depends of course on the variation of the solution. Let us explain how this works for the Poisson equation on a periodic domain that we shall need for our simulations.

Consider the Poisson equation $-\Delta\phi = \rho$ on a periodic domain of $\mathbb{R}^3$ of period L_1, L_2, L_3 in each direction. The solution is uniquely defined provided we assume that the integral of ϕ on one period vanishes.

We look for an approximation of ϕ of the form in the form of a truncated Fourier series

$$\phi_h(x_1, x_2, x_3) = \sum_{k_1=-N_1/2}^{N_1/2-1} \sum_{k_2=-N_2/2}^{N_2/2-1} \sum_{k_3=-N_3/2}^{N_3/2-1} \hat{\phi}_{k_1,k_2,k_3} e^{i\mathbf{k}\cdot\mathbf{x}},$$

where we denote by $\mathbf{k} = (2\pi k_1/L_1, 2\pi k_2/L_2, 2\pi k_3/L_3)$ and by $\mathbf{x} = (x_1, x_2, x_3)$.

Remark 10 *Note that in principle, it would be natural to truncate the Fourier series in a symmetric way around the origin,* i.e. *from $-N/2$ to $N/2$. However, because the FFT is most efficient when the number of points is a power of 2, we need to use an even number of points which leads to the truncation we use here. See Canuto, Hussaini, Quarteroni and Zang for more details [28].*

We assume the same decomposition for ρ_h. Then taking explicitly the Laplace of ϕ_h we get

$$-\Delta\phi_h(x_1, x_2, x_3) = \sum_{k_1=-N_1/2}^{N_1/2-1} \sum_{k_2=-N_2/2}^{N_2/2-1} \sum_{k_3=-N_3/2}^{N_3/2-1} |\mathbf{k}|^2 \hat{\phi}_{k_1,k_2,k_3} e^{i\mathbf{k}\cdot\mathbf{x}}.$$

Then using the collocation principle, we identify this expression with that of ρ_h at the discretisation points $\mathbf{j} = (j_1 L_1/N_1, j_2 L_2/N_2, j_3 L_3/N_3)$ with $0 \leq j_i \leq N_i - 1$:

$$\sum_{k_1,k_2,k_3} |\mathbf{k}|^2 \hat{\phi}_{k_1,k_2,k_3} e^{i\mathbf{k}\cdot\mathbf{j}} = \sum_{k_1,k_2,k_3} \hat{\rho}_{k_1,k_2,k_3} e^{i\mathbf{k}\cdot\mathbf{j}}.$$

Then as the $(e^{i\mathbf{k}\cdot\mathbf{j}})_{(k_1,k_2,k_3)}$ form a basis of $\mathbf{R}^{N_1} \times \mathbf{R}^{N_2} \times \mathbf{R}^{N_2}$ we can identify the coefficients, so that we have a simple expression of the Fourier coefficients of ϕ_h with respect to those of ρ_h for $|k| \neq 0$:

$$\hat{\phi}_{k_1,k_2,k_3} = \frac{\hat{\rho}_{k_1,k_2,k_3}}{|\mathbf{k}|^2}, \qquad -N_i/2 \leq k_i \leq N_i/2 - 1,$$

and because we have assume that the integral of ϕ is 0, we have in addition $\hat{\phi}_{0,0,0} = 0$.

Now, to complete the algorithm , we shall describe how these coefficients can be computed from the grid values by a 3D discrete Fourier transform.

In 1D, $\hat{\phi}_k$ is defined by $\hat{\phi}_k = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} \phi_i e^{2i\pi jk/N}$ from which it is easy to see that $\hat{\phi}_{k+lN} = \hat{\phi}_k$ for any integer l. So after having computed $\hat{\phi}_k$ using a Fast Fourier Transform for $0 \leq k \leq N-1$ we get the negative coefficients $\hat{\phi}_k$ for $-N/2 \leq k \leq -1$ from the known coefficients by the relation $\hat{\phi}_k = \hat{\phi}_{k+N}$. Finally to go to the 3D case, it is enough to see that the 3D discrete Fourier transform is nothing but a series of 1D transforms in each direction.

Remark 11 *In order to compute the electric field $\mathbf{E} = -\nabla\phi$ in the pseudo-spectral approximation, we just multiply each mode $\hat{\phi}_{k_1,k_2,k_3}$ by the corresponding $i\mathbf{k}$. Beware however, that because we use an unsymmetric truncated Fourier series, the mode $-N/2$ of the electric field needs to be set to 0 in order to get back a real electric field by inverse Fourier transform. Indeed for a real field $(u_j)_{0\leq j\leq N-1}$, the corresponding $N/2$ mode is $\sum_{j=0}^{N-1}(-1)^j u_j$ which is real. Hence, as ρ and then ϕ are real, their corresponding $N/2$ mode is real, and thus the same mode for E would be purely imaginary and not real unless it is 0. Note that setting this mode to 0 introduces an additional error of the order of truncation error of the series, and thus is acceptable.*

6.2 Discretization of the 1D Maxwell equations

The E_x component can be obtained by solving a standard ODE either in space or in time. We shall concentrate on the propagating system coupling E_y and B_z that will give us some interesting insight for the solution of Maxwell's equation in higher dimensions. In order to introduce and analyze our numerical schemes, we shall consider the system

$$\frac{\partial E}{\partial t} + c^2 \frac{\partial B}{\partial x} = -\frac{1}{\varepsilon_0} J, \tag{6.1}$$

$$\frac{\partial B}{\partial t} + \frac{\partial E}{\partial x} = 0, \tag{6.2}$$

where E stands for E_y, B for B_z and J stands for J_y. We shall assume a periodic domain and given initial conditions E_0 and B_0.

6.2.1 Centered finite difference discretization

Let us define a uniform grid of the periodic domain $[0, L[$: $x_j = j\Delta x$, $j = 0, \ldots, N-1$ with $\Delta x = L/N$. All functions are assumed L-periodic. The electric and magnetic fields are approximated on staggered grids, E_j will be an approximation of $E(x_j)$, $j = 0, \ldots, N-1$ and $B_{j+\frac{1}{2}}$ will be an approximation of $B(x_{j+\frac{1}{2}})$, $j = 0, \ldots, N-1$, with $x_{j+\frac{1}{2}} = (j+\frac{1}{2})\Delta x$. A centered finite difference approximation, will then yield for $j = 0, \ldots, N-1$

$$\frac{dE_j}{dt} + c^2 \frac{B_{j+\frac{1}{2}} - B_{j-\frac{1}{2}}}{\Delta x} = 0, \tag{6.3}$$

$$\frac{dB_{j+\frac{1}{2}}}{dt} + \frac{E_{j+1} - E_j}{\Delta x} = 0. \tag{6.4}$$

It is convenient to write this differential system in matrix form. For this we introduce

$$\mathbf{E} = \begin{pmatrix} E_0 \\ \vdots \\ E_{N-1} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} B_{\frac{1}{2}} \\ \vdots \\ B_{N-\frac{1}{2}} \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ -1 & 1 & 0 & \ldots & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix}.$$

In the case of periodic boundary conditions, D is a square matrix with N lines and columns, it is generally rectangular for other boundary conditions.

Then the differential system (6.3)-(6.4) becomes

$$\frac{d\mathbf{E}}{dt} = -\frac{c^2}{\Delta x} D\mathbf{B}, \tag{6.5}$$

$$\frac{d\mathbf{B}}{dt} = \frac{1}{\Delta x} D^T \mathbf{E}. \tag{6.6}$$

A second order centered time scheme can be obtained also by computing the electric and magnetic fields at staggered times using a leap-frog time stepping schemes, that we can derive using a Strmer-Verlet formulation, in order to start from time $t = 0$ for both fields. This reads

$$\frac{\mathbf{B}^{n+\frac{1}{2}} - \mathbf{B}^n}{\frac{\Delta t}{2}} = \frac{1}{\Delta x} D^T \mathbf{E}^n, \tag{6.7}$$

$$\frac{\mathbf{E}^n + 1 - \mathbf{E}^n}{\Delta t} = -\frac{c^2}{\Delta x} D\mathbf{B}^{n+\frac{1}{2}}, \tag{6.8}$$

$$\frac{\mathbf{B}^{n+1} - \mathbf{B}^{n+\frac{1}{2}}}{\frac{\Delta t}{2}} = \frac{1}{\Delta x} D^T \mathbf{E}^{n+1}, \tag{6.9}$$

adding (6.9) and (6.7) for the next time step, we get the traditional leap-frog scheme which combines (6.8) with

$$\frac{\mathbf{B}^{n+\frac{3}{2}} - \mathbf{B}^{n+\frac{1}{2}}}{\Delta t} = \frac{1}{\Delta x} D^T \mathbf{E}^{n+1}. \tag{6.10}$$

6.2.2 Mixed Finite Element formulation

We shall construct an arbitrary order mixed Finite Element approximation of the 1D Maxwell equations (6.1)-(6.2). For this we define a mesh $0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < L$ of the periodic interval $[0, L[$. As we consider periodic boundary conditions the values at L will be the values at 0.

A Finite Element method is based on a variational formulation. In the case of mixed Finite Elements, one of the equation is kept in strong form (in our case (6.1)) and the other is put in weak form by integrating by parts after having multiplied by a test function. This yields the following variational formulation:
Find $(E, B) \in H^1_\sharp([0, L]) \times L^2_\sharp([0, L])$ *such that*

$$\begin{aligned} \frac{d}{dt}\int_0^L E(x,t)F(x)\,dx - c^2 \int_0^L B(x,t)\frac{\partial F}{\partial x}(x)\,dx \\ = -\frac{1}{\varepsilon_0}\int_0^L J(x,t)F(x)\,dx \quad \forall F \in H^1_\sharp([0,L]), \end{aligned} \tag{6.11}$$

$$\frac{d}{dt}\int_0^L B(x,t)C(x)\,dx + \int_0^L \frac{\partial E}{\partial x}(x,t)C(x)\,dx = 0 \quad \forall C \in L^2_\sharp([0,L]). \tag{6.12}$$

The $\sharp$ subscript stands for spaces of periodic functions.

We define the discrete subpaces $V_k \subset H^1_\sharp([0, L])$ and $W_k \subset L^2_\sharp([0, L])$ as follows

$$V_k = \{F \in C^0_\sharp([0,L]) \mid F_{|[x_i,x_{i+1}]} \in \mathbb{P}_k\},$$
$$W_k = \{C \in L^2_\sharp([0,L]) \mid C_{|[x_i,x_{i+1}]} \in \mathbb{P}_{k-1}\},$$

where we denote by $\mathbb{P}_k$ the space of polynomials of degree less or equal to k. The functions of V_k are continuous and piecewise polynomials of degree k and the functions of W_k piecewise polynomials of degree $k-1$ with no continuity requirements at the grid points. The dimension of $\mathbb{P}_k$ the space of polynomials of one variable of degree less or equal to k is $k+1$. It follows that, due to the continuity requirement at the grid points, the dimension of V_k is Nk for N cells, and the dimension of W_0 is also Nk. Notice that the derivatives of functions of V_k are in W_k.

The discrete variational formulation in the spaces V_k and W_k then reads
Find $(E_h, B_h) \in V_k \times W_k$ *such that*

$$\frac{d}{dt}\int_0^L E_h(x,t)F(x)\,dx - c^2\int_0^L B_h(x,t)\frac{\partial F}{\partial x}(x)\,dx = -\frac{1}{\varepsilon_0}\int_0^L J(x,t)F(x)\,dx \quad \forall F \in V_k, \tag{6.13}$$

$$\frac{d}{dt}\int_0^L B_h(x,t)C(x)\,dx + \int_0^L \frac{\partial E_h}{\partial x}(x,t)C(x)\,dx = 0 \quad \forall C \in W_k. \tag{6.14}$$

We shall now express this variational formulations in finite dimensional spaces in matrix form using appropriate basis functions of the spaces V_k and W_k. In the case of high order methods we need to keep in mind that the condition number of the elementary mass matrices should be kept as low as possible in order to avoid problems coming from round-off errors. We shall consider here for simplicity only Lagrange Finite Elements, where the degrees of freedom are point values at Lagrange interpolating points.

It is well known in particular that Lagrange Finite Elements with uniformly distributed interpolation points (degrees of freedom) lead to very ill conditioned matrices for moderately large values of k. A much better option is to use Lagrange polynomials at Gauss points. This is the best choice for W_k. For V_k we have the additional continuity condition at the grid points which forces us to put degrees of freedom at the grid points. Then the best choice is the Gauss-Lobatto points.

Remark 12 *Note that if one does not want to stick to Lagrange Finite Elements a natural choice of basis functions for W_k would be the orthonormal Legendre polynomials for which the mass matrix would be identity.*

Let us denote by $(\varphi_i)_{0\leq kN-1}$ the basis of V_k and $(\psi_j)_{0\leq j\leq kN-1}$ the basis of W_k.

Let us know compute the different integrals appearing in the variational formulation (6.13)-(6.14) using the basis functions. Expressing E_h and F using the basis (φ_i) and B_h and C in the basis (ψ_j). We get

$$E_h(x,t) = \sum_{j=0}^{kN-1} E_j(t)\varphi_j(x), \quad F(x) = \sum_{i=0}^{kN-1} F_i\varphi_i(x),$$

$$B_h(x,t) = \sum_{j=0}^{kN-1} B_j(t)\psi_j(x), \quad C(x) = \sum_{i=0}^{kN-1} F_i\psi_i(x).$$

Note that as both bases are Lagrange bases the coefficients E_i, F_i, B_i, C_i are simply the values of the corresponding functions E_h, F, B_h and C at the corresponding points.

Plugging these expression into (6.13)-(6.14) we obtain

$$\sum_{i=0}^{kN-1}\sum_{j=0}^{kN-1}\left[\frac{dE_j(t)}{dt}F_i\int_0^L \varphi_i(x)\varphi_j(x)\,dx - B_j(t)F_i\int_0^L \varphi_i'(x)\psi_j(x)\,dx\right] \tag{6.15}$$
$$= \sum_{i=0}^{N-1} F_i\int_0^L J(x)\varphi_i(x)\,dx$$

$$\sum_{i=0}^{kN-1}\sum_{j=0}^{kN-1}\left[\frac{dB_j(t)}{dt}C_i\int_0^L \psi_i(x)\psi_j(x)\,dx + E_j(t)C_i\int_0^L \varphi_j'(x)\psi_i(x)\,dx\right] = 0. \tag{6.16}$$

Denote by

$$\mathbb{E}(t) = \begin{pmatrix} E_0(t) \\ \vdots \\ E_{N-1}(t)\end{pmatrix}, \quad \mathbb{B}(t) = \begin{pmatrix} B_0(t) \\ \vdots \\ B_{N-1}(t)\end{pmatrix}, \quad \mathbb{F} = \begin{pmatrix} F_0 \\ \vdots \\ F_{N-1}\end{pmatrix}, \quad \mathbb{C} = \begin{pmatrix} C_0 \\ \vdots \\ C_{N-1}\end{pmatrix},$$

and

$$\mathbb{J}(t) = \begin{pmatrix} \int_0^L J(t,x)\varphi_0(x)\,dx \\ \vdots \\ \int_0^L J(t,x)\varphi_{N-1}(x)\,dx\end{pmatrix}.$$

Let us now introduce the mass matrices

$$M_E = ((\int_0^L \varphi_i(x)\varphi_j(x)))_{0\le i\le Nk-1, 0\le j\le Nk-1},$$

$$M_B = ((\int_0^L \psi_i(x)\psi_j(x)))_{0\le i\le Nk-1, 0\le j\le Nk-1},$$

and the derivative matrix

$$K = ((\int_0^L \varphi_j'(x)\psi_i(x)))_{0\le i\le Nk-1, 0\le j\le Nk-1}.$$

The variational formulations (6.15)-(6.16) then become

$$\mathbb{F}^T M_E\frac{d\mathbb{E}(t)}{dt} - c^2\mathbb{F}^T K^T\mathbb{B} = -\frac{1}{\varepsilon_0}\mathbb{F}^T\mathbb{J} \quad \forall \mathbb{F}\in\mathbb{R}^{N-1},$$

$$\mathbb{C}^T M_B\frac{d\mathbb{B}(t)}{dt} + \mathbb{C}^T K\mathbb{E} = 0 \quad \forall \mathbb{C}\in\mathbb{R}^{N-1}.$$

As M_E and M_B are non singular matrices this can be written equivalently

$$\frac{d\mathbb{E}(t)}{dt} - M_E^{-1}K^T\mathbb{B} = 0, \tag{6.17}$$

$$\frac{d\mathbb{B}(t)}{dt} + M_B^{-1}K\mathbb{E} = 0. \tag{6.18}$$

As usual for Finite Elements the matrices M_B, M_E and K are computed from the corresponding elementary matrices that are obtained by change of variables onto the reference element $[-1,1]$ for each cell. So

$$\int_0^L \varphi_i(x)\varphi_j(x)\,dx = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varphi_i(x)\varphi_j(x)\,dx,$$

and doing the change of variable $x = \frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2}$, we get

$$\int_{x_n}^{x_{n+1}} \varphi_i(x)\varphi_j(x)\,dx = \frac{x_{n+1}-x_n}{2} \int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})\,d\hat{x},$$

where $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2})$. The local indices α on the reference element go from 0 to k and the global numbers of the basis functions not vanishing on element n are $j = kn+\alpha$. The $\hat{\varphi}_\alpha$ are the Lagrange polynomials at the Gauss-Lobatto points in the l'interval $[-1,1]$.

The mass matrix in V_k can be approximated with no loss of order of the finite element approximation using the Gauss-Lobatto quadrature rule. Then because the products $\hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})$ vanish for $\alpha \neq \beta$ at the Gauss-Lobatto points by definition of the $\hat{\varphi}_\alpha$ which are the Lagrange basis functions at these points, the elementary matrix $\hat{M}_E$ is diagonal and we have

$$\int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})^2\,d\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL} \varphi_\alpha(\hat{x}_\beta)^2 = w_\alpha^{GL}$$

using the quadrature rule, where w_α^{GL} is the Gauss-Lobatto weight at Gauss-Lobatto point $(\hat{x}_\alpha) \in [-1,1]$. So that finally $\hat{M}_E = diag(w_0^{GL}, \dots w_k^{GL})$ is the matrix with $k+1$ lines and columns with the Gauss-Lobatto weights on the diagonal.

In the same spirit we use Gauss quadrature to compute the mass matrix in W_k. In this case, the quadrature is exact and we get in the same way an elementary matrix which is diagonal of order k where the diagonal terms are the Gauss weights: $\hat{M}_B = diag(w_0^G, \dots w_{k-1}^{GL})$.

Let us now compute the elements of K. As previously we go back to the interval $[-1,1]$ with the change of variables $x = \frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2}$ and we define $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2})$ and the same for ψ_i. Note however that a global basis function φ_i associated to a grid point has a support which overlaps two cells and is associated to two local basis functions whereas the global basis functions ψ_i have all only a support on one cell and are only associated to one local basis functions. This is the reason why there are $k+1$ local basis functions $\hat{\varphi}_\alpha$ and only k local basis functions $\hat{\psi}_\alpha$ even though there are in our case the same number of global basis functions for V_k and W_k.

We then get $\hat{\varphi}'_\alpha(\hat{x}) = \frac{x_{n+1}-x_n}{2}\varphi'_i(\frac{x_{n+1}-x_n}{2}(\hat{x}+1) + x_n)$. It follows that

$$\int_{x_n}^{x_{n+1}} \varphi_j'(x)\psi_i(x)\,dx = \int_{-1}^{1} \frac{2}{x_{n+1}-x_n}\hat{\varphi}_\beta'(\hat{x})\hat{\psi}_\alpha(\hat{x})\frac{x_{n+1}-x_n}{2}\,d\hat{x}$$
$$= \int_{-1}^{1} \hat{\varphi}_\beta'(\hat{x})\hat{\psi}_\alpha(\hat{x})\,d\hat{x}.$$

Both $\hat{\varphi}_\beta'(\hat{x})$ and $\hat{\psi}_\alpha(\hat{x})$ are of degree $k-1$ so that the Gauss-Lobatto quadrature rule with $k+1$ points is exact. Using this rule

$$\int_{-1}^{1} \hat{\varphi}_\beta'(\hat{x})\hat{\psi}_\alpha(\hat{x})\,d\hat{x} = \sum_{m=0}^{k} w_m^{GL}\hat{\varphi}_\beta'(\hat{x}_m^{GL})\hat{\psi}_\alpha(\hat{x}_m^{GL}).$$

In this case there is no vanishing term, but this expression can be used to compute the elementary matrix $\hat{K}$ along with the formula for the Lagrange polynomial at the Gauss points

$$\hat{\psi}_\alpha(\hat{x}_m^{GL}) = \frac{\pi_{\beta\neq\alpha}(\hat{x}_m^{GL}-\hat{x}_\beta^G)}{\pi_{\beta\neq\alpha}(\hat{x}_\alpha^G-\hat{x}_\beta^G)}.$$

On the other hand evaluating the derivatives of the Lagrange polynomial at the Gauss-Lobatto points at these Gauss-Lobatto points can be done using the formula

$$\hat{\varphi}_\alpha'(\hat{x}_\beta^{GL}) = \frac{p_\beta/p_\alpha}{\hat{x}_\beta^{GL}-\hat{x}_\alpha^{GL}} \text{for } \beta\neq\alpha \text{ and } l_\alpha'(\hat{x}_\alpha) = -\sum_{\beta\neq\alpha} l_\beta'(\hat{x}_\alpha).$$

Note that the support of a function ψ_j is restricted to only one cell $[x_n, x_{n+1}]$, of the mesh. Therefore matrix K is consists in blocks of size $k\times(k+1)$ with only one block by group of k lines and with a common column corresponding to a grid point for two successive blocks.

6.2.3 B-spline Finite Elements

Let us now construct a different kind of Finite Element discretization using B-Splines as basis functions.

In order to define a family of n B-splines of degree k, we need $(x_i)_{0\leqslant i\leqslant n+k}$ a non-decreasing sequence of points on the real line called *knots* in the spline terminology. There can be several knots at the same position. In the case when there are m knots at the same point, we say that the knot has multiplicity m.

Definition 5 (B-Spline) *Let $(x_i)_{0\leqslant i\leqslant n+k}$ be a non-decreasing sequence of knots. Then the j-th B-Spline ($0\leq j\leq n-1$) denoted by N_j^k of degree k is defined by the recurrence relation:*

$$N_j^k(x) = w_j^k(x)N_j^{k-1}(x) + (1-w_{j+1}^k(x))N_{j+1}^{k-1}(x)$$

where,

$$w_j^k(x) = \frac{x-x_j}{x_{j+k}-x_j} \qquad N_j^0(x) = \chi_{[x_j,x_{j+1}[}(x)$$

We note some important properties of a B-splines basis:

- B-splines are piecewise polynomial of degree k,
- B-splines are non negative
- Compact support; the support of N_j^k is contained in $[t_j, .., t_{j+k+1}]$
- Partition of unity: $\sum_{i=0}^{n-1} N_i^k(x) = 1, \forall x \in \mathbb{R}$
- Local linear independence
- If a knot x_i has a multiplicity m then the B-spline is $\mathcal{C}^{(k-m)}$ at x_i.

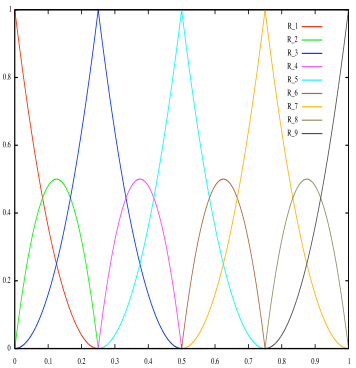

Fig. 6.1. All B-splines functions associated to a knot sequence defined by $n = 9$, $k = 2$, $T = \{000, \frac{1}{4}\frac{1}{4}, \frac{1}{2}\frac{1}{2}, \frac{3}{4}\frac{3}{4}, 111\}$

A key point for constructing discrete Finite Element spaces for the Maxwell equation comes from the recursion formula for the derivatives:

$$N_i^{k\prime}(x) = k\left(\frac{N_i^{k-1}(x)}{x_{i+k} - x_i} - \frac{N_{i+1}^{k-1}(x)}{x_{i+k+1} - x_{i+1}}\right). \tag{6.19}$$

It will be convenient to introduce the notation $D_i^k(x) = k\frac{N_i^{k-1}(x)}{x_{i+k}-x_i}$. Then the recursion formula for derivative simply becomes

$$N_i^{k\prime}(x) = D_i^k(x) - D_{i+1}^k(x). \tag{6.20}$$

Remark 13 *In the case where all knots, except the boundary knots are of multiplicity 1, the set $(N_i^k)_{0\leq i\leq n-1}$ of B-splines of degree k forms a basis of the spline space defined by*

$$\mathcal{S}^k = \{v \in C^{k-1}([x_0, x_n]) \; | v_{|[x_i,x_{i+1}]} \in \mathbb{P}_k([x_i, x_{i+1}]).$$

The boundary knots are chosen to have multiplicity $k + 1$ so that the spline becomes interpolatory on the boundary in order to simplify the application of Dirichlet boundary conditions.

Then due to the definitions if follows immediately that $(D_i^k)_{1\leq i\leq n-1}$ *is a basis of* $\mathcal{S}^{k-1}$*. Note that if the first knot has multiplicity* $k+1$*,* D_0^k *will have a support restricted to one point and be identically 0.*

Remark 14 *Splines can be easily defined in the case of periodic boundary conditions by taking a periodic knot sequence.*

Assuming only knots of multiplicity 1 and denoting by $\mathcal{S}_\sharp^k$ the set of periodic splines associated to a periodic knot sequences, we can take $V_k = \mathcal{S}_\sharp^k$ whose basis functions are the N_i^k and $W_k = \mathcal{S}_\sharp^{k-1}$ with basis functions the D_i^k. This defines the Finite Element spaces that we can use with the discrete variational formulation of Maxwell's equations (6.13)-(6.14). We can then construct the mass and derivative matrices like for the Lagrange Finite Elements, just replacing the basis functions by their spline counterparts, *i.e.* φ_i by N_i^k and ψ_i by D_i^k. The matrices can be computed with no quadrature error using adequate Gauss or Gauss-Lobato formulas. We then get a linear system which has the form (6.17)-(6.18). Note that in this case the mass matrices will not become diagonal thanks to the quadrature formula. However, as one of the two variational formulations, the Faraday equation (6.14) in our case (but it could have been the other one), has not been integrated by part it can be written in strong form in the discretization space thanks to the choice of the spaces. Indeed, we have that the derivatives of the functions of V_k are in W_k. Hence $\frac{\partial E_h}{\partial x} \in W_k$ and (6.14) implies that $\frac{\partial B_h}{\partial t} + \frac{\partial E_h}{\partial x}$ is a function of W_k which is orthogonal to all functions of W_k and thus vanishes. The second variational formulation is thus equivalent to the strong form

$$\frac{\partial B_h}{\partial t} + \frac{\partial E_h}{\partial x} = 0.$$

Let us now write the expressions of B_h and E_h in their respective basis: $B_h = \sum_i \mathbb{B}_i(t) D_i^k(x)$ and $E_h = \sum_i \mathbb{E}_i(t) N_i^k(x)$. Note that unlike for Lagrange finite elements, the coefficients $\mathbb{B}_i(t)$ and $\mathbb{E}_i(t)$ on the bases are not values of the discrete functions at some given points. Then taking the derivative with respect to x of E_h and using the B-spline derivative formula (6.20) we find

$$\frac{\partial E_h}{\partial x} = \sum_i \mathbb{E}_i(t)(N_i^k)'(x) = \sum_i \mathbb{E}_i(t)(D_i^k(x) - D_{i+1}^k(x)) = \sum_i (\mathbb{E}_i(t) - \mathbb{E}_{i-1}) D_i^k(x).$$

Hence identifying the coefficients of the basis, the discrete Faraday equation reduces in this case to

$$\frac{d\mathbb{B}_i}{dt} = \mathbb{E}_i(t) - \mathbb{E}_{i-1}(t).$$

On the other hand, as in the case of the mixed Lagrange Finite Elements the discrete matrix form of Ampère's law writes

$$M_V \frac{d\mathbb{E}}{dt} = K^T \mathbb{B},$$

where

$$M_V = ((\int N_i^k(x) N_j^k(x)\,dx))_{0\le i,j\le n-1}, \quad K = ((\int (N_i^k)'(x) D_j^k(x)\,dx))_{0\le i,j\le n-1}.$$

6.3 High-order time schemes

6.3.1 Schemes based on a Taylor expansion

When working with linear homogeneous equations, like Maxwell's equations in vacuum with no source term, the simplest way to derive high order time schemes is to use a Taylor expansion in time and plug in the expression of the successive time derivatives obtained from the differential system resulting from the semi-discretization in space. Consider for example that after semi-discretization in space using Finite Elements or Finite Differences we obtain the differential systems

$$\frac{dU}{dt} = AU, \text{ with } U = \begin{pmatrix} U_0 \\ \vdots \\ U_{M-1} \end{pmatrix},$$

and A the appropriate matrix coming from the semi-discretization in space. Then a Taylor expansion in time up to order p yields

$$U(t_{n+1}) = U(t_n) + \Delta t \frac{dU}{dt}(t_n) + \cdots + \frac{\Delta t^p}{p!}\frac{d^pU}{dt^p}(t_n) + O(\Delta t^p).$$

Now as A does not depend on time and $\frac{dU}{dt} = AU$, we get that

$$\frac{d^pU}{dt^p} = A^pU, \text{ for any integer} p.$$

Hence, denoting U^n an approximation of $U(t_n)$, we get a time scheme of order p using the formula

$$U^{n+1} = U^n + \Delta t A U^n + \cdots + \frac{\Delta t^p}{p!} A^p U^n = (I + \Delta t A + \cdots + \frac{\Delta t^p}{p!} A^p) U^n. \quad (6.21)$$

For $p = 1$ this boils down to the standard explicit Euler scheme.

6.3.2 Leap-frog schemes

For systems resulting from Maxwell's equations, the semi-discretized equation have the form

$$\frac{dE}{dt} = M_E^{-1} K^T B. \quad (6.22)$$

$$\frac{dB}{dt} = -M_B^{-1} K E, \quad (6.23)$$

In this case, it is possible to go to high order with fewer operations using a leap-frog technique, with for example E computed at integer time steps and B computed at half integer time steps. Indeed, summing the Taylor expansions at $t_{n+\frac{1}{2}}$ and $t_{n-\frac{1}{2}}$ which write

$$B^{n\pm\frac{1}{2}} = B^n \pm \frac{\Delta t}{2}\frac{dB}{dt}(t_n) + \frac{\Delta t^2}{8}\frac{d^2B}{dt^2}(t_n) \pm \frac{\Delta t^3}{48}\frac{d^3B}{dt^3}(t_n) + \dots$$

we get

$$B^{n+\frac{1}{2}} = B^{n-\frac{1}{2}} + \Delta t\frac{dB}{dt}(t_n) + \frac{\Delta t^3}{24}\frac{d^3B}{dt^3}(t_n) + \dots$$

where only the odd order derivatives remain. These can be expressed using the semi-discrete equations (6.23)-(6.22) which yield $\frac{dB}{dt} = -M_B^{-1}KE$, $\frac{d^3B}{dt^3} = M_B^{-1}KM_E^{-1}K^TM_B^{-1}KE$ and so on. So that

$$B^{n+\frac{1}{2}} = B^{n-\frac{1}{2}} - \Delta t M_B^{-1}KE^n + \frac{\Delta t^3}{24}M_B^{-1}KM_E^{-1}K^TM_B^{-1}KE^n + O(\Delta t^5). \tag{6.24}$$

In the same way we get for E

$$E^{n+1} = E^n + \Delta t M_E^{-1}K^TB^{n+\frac{1}{2}} - \frac{\Delta t^3}{24}M_E^{-1}K^TM_B^{-1}KM_E^{-1}K^TB^{n+\frac{1}{2}} + O(\Delta t^5). \tag{6.25}$$

The scheme (6.24)-(6.25) is globally fourth order in time, the local truncation error being $O(\Delta t^5)$. Higher (even) order time schemes can be obtained in the same way by taking higher order Taylor expansions. Note that if only the first term in the expansion is kept, we get the standard second order leap-frog scheme.

Notice that all the high-order leap-frog schemes in the same line as (6.25)-(6.24) have the form

$$B^{n+\frac{1}{2}} = B^{n-\frac{1}{2}} - \Delta t M_B^{-1}K_pE^n, \tag{6.26}$$

$$E^{n+1} = E^n + \Delta t M_E^{-1}K_p^TB^{n+\frac{1}{2}}, \tag{6.27}$$

with $K_p = K$ for the second order Leap-Frog scheme,

$$K_p = K - \frac{\Delta t^2}{24}KM_E^{-1}K^TM_B^{-1}K$$

for the fourth order,

$$K_p = K - \frac{\Delta t^2}{24}KM_E^{-1}K^TM_B^{-1}K + \frac{\Delta t^4}{1920}KM_E^{-1}K^TM_B^{-1}KM_E^{-1}K^TM_B^{-1}K$$

for the sixth order and so on.

In order to initialize a high order leap-frog scheme, one needs to compute $B^{\frac{1}{2}}$ from B^0 and E^0 with the same order. For example for the fourth order leap-frog scheme

$$B^{\frac{1}{2}} = B^0 + \frac{\Delta t}{2} M_B^{-1} K E^0 + \frac{\Delta t^2}{8} M_B^{-1} K M_E^{-1} K^T B^0 + \frac{\Delta t^3}{48} M_B^{-1} K M_E^{-1} K^T M_B^{-1} K B^0 + \frac{\Delta t^4}{384} M_B^{-1} K M_E^{-1} K^T M_B^{-1} K M_E^{-1} K^T B^0 + O(\Delta t^5).$$

In the same way it is possible to express B^{n+1} from $B^{n+\frac{1}{2}}$ if needed.

Remark 15 *Taylor expansions are particularly simple for getting high order time scheme when no source terms are present. They can still be used with source terms provided their time derivatives are readily available.*

6.3.3 Conservation of a discrete energy for the leap-frog scheme

Starting from the general form of the high-order leap-frog schemes (6.26)-(6.27) we have

$$B^{n+\frac{1}{2}} = B^{n-\frac{1}{2}} - \Delta t M_B^{-1} K_p E^n, \tag{6.28}$$

$$E^{n+1} = E^n + \Delta t M_E^{-1} K_p^T B^{n+\frac{1}{2}}. \tag{6.29}$$

Lemma 3 *The discrete energy*

$$\mathcal{E}^n = \frac{1}{2}(E_n^T M_E E_n + B_{n-\frac{1}{2}}^T M_B B_{n+\frac{1}{2}})$$

does not depend on n.

Proof. The idea is to mimic the energy conservation proof at the semi-discrete level which uses the structure of the right-hand-side with $-K_p$ and K_p^T.

We first take the dot product of (6.28) multiplied by M_B with $B_{n+\frac{1}{2}}$ (which amounts to multiplying by $B_{n+\frac{1}{2}}^T$ on the left)

$$B_{n+\frac{1}{2}}^T M_B B_{n+\frac{1}{2}} = B_{n+\frac{1}{2}}^T M_B B_{n-\frac{1}{2}} - \Delta t B_{n+\frac{1}{2}}^T K_p E_n,$$

then we perform the same operation for (6.28) where n is replaced by $n+1$

$$B_{n+\frac{1}{2}}^T M_B B_{n+\frac{3}{2}} = B_{n+\frac{1}{2}}^T M_B B_{n+\frac{1}{2}} - \Delta t B_{n+\frac{1}{2}}^T K_p E_{n+1},$$

and now we take the dot productof (6.29) multiplied by M_E with $E_n + E_{n+1}$

$$(E_n + E_{n+1})^T M_E E_{n+1} = (E_n + E_{n+1})^T M_E E_n + \Delta t (E_n + E_{n+1})^T K_p^T B^{n+\frac{1}{2}}.$$

Finally we sum the three relations to get

$$B_{n+\frac{3}{2}}^T M_B B_{n+\frac{1}{2}} + E_{n+1}^T M_E E_{n+1} = B_{n+\frac{1}{2}}^T M_B B_{n-\frac{1}{2}} + E_n^T M_E E_n,$$

as $E_{n+1}^T M_E E_n = E_n^T M_E E_{n+1}$ and $B_{n+\frac{1}{2}}^T K_p (E_n + E_{n+1}) = (E_n + E_{n+1})^T K_p^T B_{n+\frac{1}{2}}$.

We now want to prove that $\mathcal{E}^n$ is a definite positive quadratic form of $(E_n, B_{n-\frac{1}{2}})$ provided the time step Δt is sufficiently small. To this aim we will need a lemma on block matrix determinants which will also be useful later.

Lemma 4 *Consider a matrix defined by blocks*

$$\mathcal{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Assume that A is non singular and that A and C commute. Then

$$\det \mathcal{A} = \det(AD - CB).$$

Proof. By block multiplication we get that

$$\begin{pmatrix} A^{-1} & 0 \\ -C & A \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & A^{-1}B \\ -CA + AC & AD - CB \end{pmatrix}.$$

Then as A and C commute, the last matrix is block triangular, so that its determinant is the product of the diagonal blocks. Therefore as the determinant of a product is the product of the determinants we obtain

$$1 \times \det \mathcal{A} = 1 \times \det(AD - CB).$$

Let us define the matrix norm $\|A\|_2 = \sqrt{\rho(A^T A)}$ where $\rho(A^T A)$ is the spectral radius of $A^T A$.

Lemma 5 *Assume that $\frac{1}{2}\|M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}}\|_2 \Delta t < 1$ The energy function $\mathcal{E}^n$ defines a definite quadratic form of $(E_n, B_{n-\frac{1}{2}})$.*

Proof. Using the definition of $\mathcal{E}^n$ and (6.28) we get

$$\mathcal{E}^n = \frac{1}{2}(E_n^T M_E E_n + B_{n-\frac{1}{2}}^T M_B (B^{n-\frac{1}{2}} - \Delta t M_B^{-1} K_p E^n).$$

Writing this in matrix form

$$2\mathcal{E}^n = \begin{pmatrix} E_n \\ B_{n-\frac{1}{2}} \end{pmatrix}^T \begin{pmatrix} M_E & -\frac{\Delta t}{2} K_p^T \\ -\frac{\Delta t}{2} K_p & M_B \end{pmatrix} \begin{pmatrix} E_n \\ B_{n-\frac{1}{2}} \end{pmatrix}$$

M_B and M_E are symmetric positive definite matrices, so we can introduce their square roots, which are defined as the matrices whose eigenvalues are the square root of the positive eigenvalues of the initial matrix. Then our expression rewrites

$$2\mathcal{E}^n = \begin{pmatrix} M_E^{\frac{1}{2}} E_n \\ M_B^{\frac{1}{2}} B_{n-\frac{1}{2}} \end{pmatrix}^T \begin{pmatrix} I & -\frac{\Delta t}{2} M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}} \\ -\frac{\Delta t}{2} M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} & I \end{pmatrix} \begin{pmatrix} M_E^{\frac{1}{2}} E_n \\ M_B^{\frac{1}{2}} B_{n-\frac{1}{2}} \end{pmatrix}$$

Hence the energy $\mathcal{E}^n$ is positive definitive provided the matrix

$$\mathcal{A} = \begin{pmatrix} I & -\frac{\Delta t}{2} M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}} \\ -\frac{\Delta t}{2} M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} & I \end{pmatrix}$$

is, which is the case if its eigenvalues are all real and non negative. The eigenvalues of $\mathcal{A}$ are the λ such that $\det(\mathcal{A}-\lambda I) = 0$. Applying Lemma (4) we have $\det(\mathcal{A}-\lambda I) = \det((1-\lambda)^2 I - \frac{\Delta t^2}{4} M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}})$. Denoting by μ^2 the eigenvalues of $M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}}$, we find that $(1-\lambda)^2 = \frac{\Delta t^2}{4}\mu^2$ so that $\lambda = 1 \pm \frac{\Delta t}{2}\mu$, so that all the λ are positive if $\frac{\Delta t}{2}\mu_{max} < 1$ where μ_{max} is the square root of the largest eigenvalue of $M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}}$ which is also the 2-norm of the matrix $M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}}$.

6.3.4 Stability of time schemes

Writing U^n the solution in vector form at time t_n, we define the propagation matrix $\mathcal{A}$ such that

$$U^{n+1} = \mathcal{A} U^n.$$

Definition 6 *The numerical scheme defined by the propagation matrix $\mathcal{A}$ is stable if there exists $\tau > 0$ such that U^n is bounded for all $n \in \mathbb{N}$ and $\Delta t < \tau$.*

Proposition 11 *The numerical scheme defined by the propagation matrix $\mathcal{A}$ is stable if there exists $\tau > 0$ such that for all $\Delta t < \tau$ all eigenvalues of $\mathcal{A}$ are of modulus less or equal to 1.*

Stability of Taylor schemes. For a Taylor scheme of ordre p applied to $\frac{dU}{dt} = AU$, we have $\mathcal{A} = I + \Delta t A + \cdots + \frac{\Delta t}{p!} A^p$. Then denoting by λ an eigenvalue of A, the corresponding eigenvalue of $\mathcal{A}$ is $\mu = 1 + \lambda\Delta t + \cdots + \lambda^p \frac{\Delta t^p}{p!}$. And one can plot the region of the complex plane in which $|\mu| \leq 1$ using for example `implicitplot` in Maple, which are the stability regions.

Note that for the order 1 and 2 scheme the intersection of the stability zone with the imaginary axis is reduced to the point 0. So that when all eigenvalues are purely imaginary as is the case for non dissipative space discretization schemes like the ones we have introduced up to now for Maxwell's equations, these schemes are not stable for any positive Δt. On the other hand the schemes of order 3 and 4 have a non vanishing stability zone on the imaginary axis, larger for the order 4 scheme. The order 5 and 6 schemes are even more problematic for eigenvalues of A on the imaginary axis as the zooms of Figure 6.4 tell us. Even though there is a part of the imaginary axis in the stability zone, there is also a part in the neighborhood of 0 which is not. Therefore small eigenvalues of A will lead to instability on longer time scales. This is problematic, as unlike usual Courant condition instability problems which reveal themselves very fast, this leads to a small growth in time.

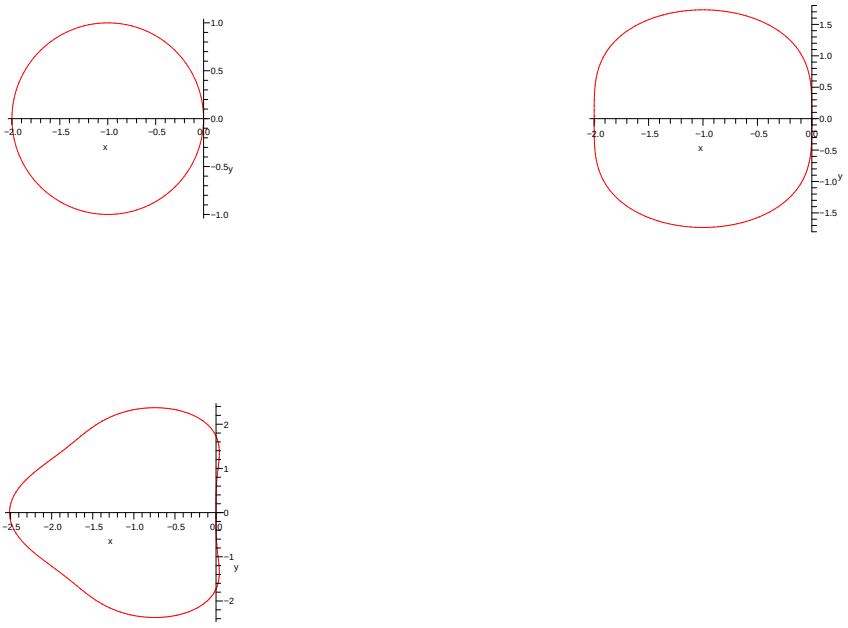

Fig. 6.2. Stability zone for Taylor schemes. From left to right order 1, 2, 3.

There are two ways to stabilize the unstable schemes or to increase the stability region, the first is to introduce some upwinding or hyperviscosity in the space discretization so that the eigenvalues get shifted to the left hand side of the imaginary axis. The second one is to modify the Taylor scheme by adding an additional term which has no influence on the order. That is consider an order p scheme where

$$\mathcal{A} = I + \Delta t A + \cdots + \frac{\Delta t^p}{p!} A^p + \sigma \frac{\Delta t^{p+1}}{(p+1)!} A^{p+1}$$

with σ chosen such that there is a good stability zone. There are choices of σ that maximize the stability zone and also choices that yield a stability zone while minimizing dissipation. Stabilization examples for the order 1 scheme with $\sigma = 2$ and the order 2 scheme with $\sigma = \frac{3}{2}$ are given in Figure 6.5.

Stability of Leap-frog schemes.

Let us consider again the general form of the high-order leap-frog schemes (6.26)-(6.27):

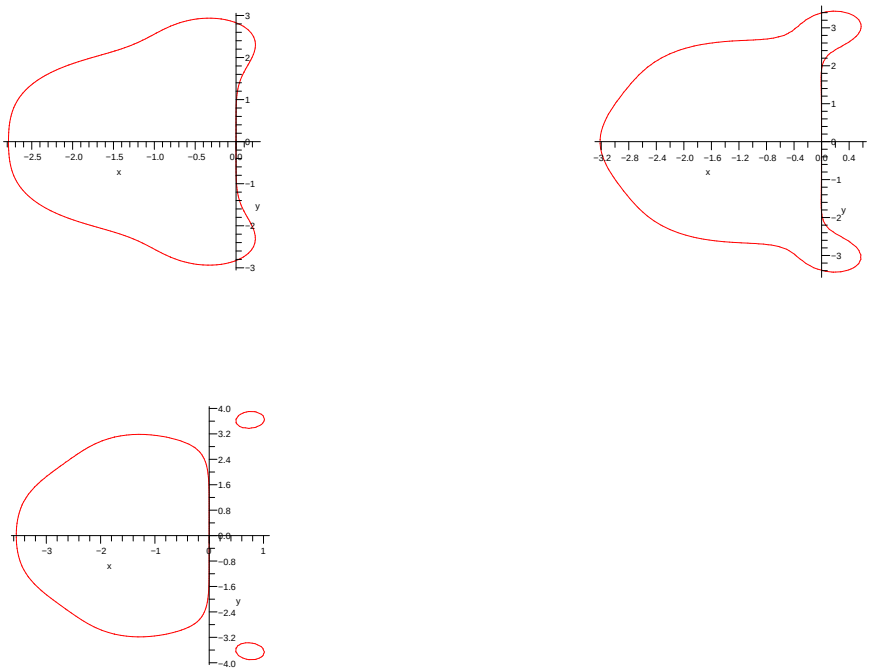

Fig. 6.3. Stability zone for Taylor schemes. From left to right order 4, 5, 6.

$$B^{n+\frac{1}{2}} = B^{n-\frac{1}{2}} - \Delta t M_B^{-1} K_p E^n, \tag{6.30}$$

$$E^{n+1} = E^n + \Delta t M_E^{-1} K_p^T B^{n+\frac{1}{2}}. \tag{6.31}$$

Replacing the expression of $B^{n+\frac{1}{2}}$ in (6.31) using (6.30) we get

$$M_E E^{n+1} = M_E E^n + \Delta t K_p^T B^{n-\frac{1}{2}} - \Delta t^2 K_p^T M_B^{-1} K_p E^n.$$

Inspired by our computations on discrete energy conservation for the leap-frog schemes, we can write a propagation matrix in the form

$$\begin{pmatrix} M_B^{\frac{1}{2}} B_{n+\frac{1}{2}} \\ M_E^{\frac{1}{2}} E_{n+1} \end{pmatrix} = \begin{pmatrix} I & -\Delta t M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} \\ \Delta t M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}} & I - \Delta t^2 M_E^{-\frac{1}{2}} K_p^T M_B^{-1} K_p M_E^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} M_B^{\frac{1}{2}} B_{n-\frac{1}{2}} \\ M_E^{\frac{1}{2}} E_n \end{pmatrix}.$$

We denote by $\mathcal{A}$ the propagation matrix

$$\mathcal{A} = \begin{pmatrix} I & -\Delta t M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} \\ \Delta t M_E^{-\frac{1}{2}} K_p^T M_B^{-\frac{1}{2}} & I - \Delta t^2 M_E^{-\frac{1}{2}} K_p^T M_B^{-1} K_p M_E^{-\frac{1}{2}} \end{pmatrix}.$$

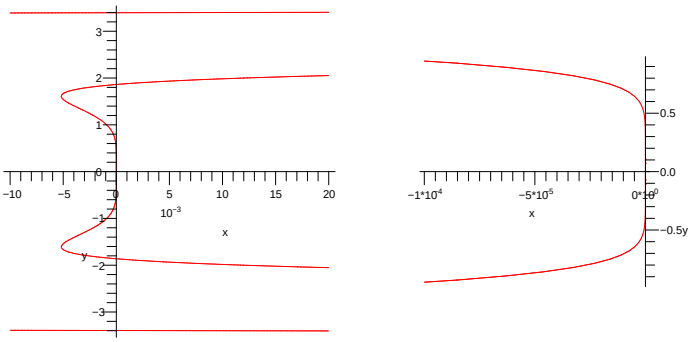

Fig. 6.4. Stability zone for Taylor schemes. Zoom around imaginary axis. Left order 5, right order 6.

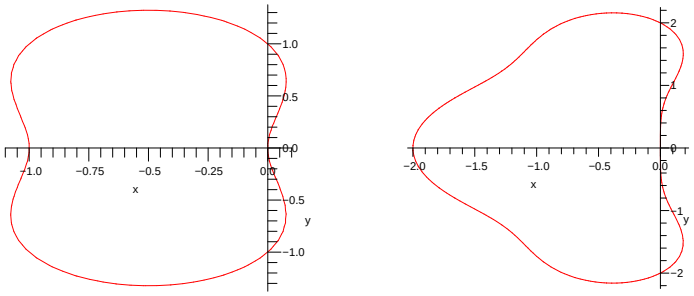

Fig. 6.5. Stability zone for stabilized Taylor schemes. Left order 1, right order 2.

Lemma 6 *A leap-frog scheme of the form* (6.30)-(6.31) *is stable if and only if*

$$\Delta t \| M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}} \|_2 \leq 2.$$

In this case the discrete energy of Lemma 3 is conserved.

Proof. In order to investigate the stability of the scheme, we need to check the eigenvalues of $\mathcal{A}$. Let us denote by $K = M_B^{-\frac{1}{2}} K_p M_E^{-\frac{1}{2}}$. Using Lemma (4)

we get

$$\det(\mathcal{A}-\lambda I) = \det((1-\lambda)((1-\lambda)I-\Delta t^2 K^T K)+\Delta t^2 K^T K) = \det((1-\lambda)^2 I+\lambda\Delta t^2 K^T K).$$

Hence the eigenvalues λ of $\mathcal{A}$ are related to the eigenvalues μ^2 of $K^T K$ by

$$(1-\lambda)^2 + \mu^2\Delta t^2\lambda = 0.$$

This is a degree two polynomial in λ with real coefficients. The constant coefficient being 1, the product of the roots is one, so that both roots satisfy $|\lambda| \leq 1$ if and only if both roots are complex conjugate, which is the case when $(2-\mu^2\Delta t^2)^2 \leq 4$, *i.e.* when $\mu\Delta t \leq 2$. This is verified for all the eigenvalues μ provided it is verified for the largest which is by definition the norm $\|K\|_2$. Note that this condition is the same as the condition for the discrete energy of Lemma 3 to be a positive quadratic form.

6.3.5 Explicit computation of the stability condition when the matrices are circulant

Note that on a uniform grid, the terms of the Finite Element mass matrices on each cell are scaled by the cell size Δx and those of the derivative matrix on each cell do not depend on Δx. In the simplest case where $k = 1$, we even have that $M_V = M_W = \Delta x\mathbb{I}$ and K is the circulant matrix with -1 on the diagonal and 1 on the upper diagonal, so that KK^T is the circulant matrix with 2 on the diagonal and -1 on the upper and lower diagonal. Using the formula for the eigenvalues of a symmetric circulant matrix we find that the eigenvalues of KK^T are $\lambda_j = 2(1+\cos\frac{2\pi j}{N})$ so that the largest eigenvalue is 4 and $\|M_W^{-\frac{1}{2}} K M_V^{-\frac{1}{2}}\|_2 = \frac{2}{\Delta x}$. The stability condition $\|M_W^{-\frac{1}{2}} K M_V^{-\frac{1}{2}}\|_2 \frac{\Delta t}{2} \leq 1$ then becomes the well-known $\Delta t \leq \Delta x$.

6.4 Discretization of the 2D Maxwell equations

6.4.1 The Yee scheme

The Yee scheme is base on centered Finite Differences on staggered meshes. Define a primal uniform 2D mesh by $x_i = x_0 + i\Delta x$, $0 \leq i \leq N_x - 1$, $y_j = y_0 + j\Delta y$, $0 \leq j \leq N_y - 1$ and the associated dual mesh with vertices $x_{i+\frac{1}{2}} = x_0 + (i+\frac{1}{2})\Delta x$, $0 \leq i \leq N_x - 2$, $y_{j+\frac{1}{2}} = y_0 + (j+\frac{1}{2})\Delta y$, $0 \leq j \leq N_y - 2$.

Consider now the TE mode part of Maxwell's equations (2.48)-(2.50) that we recall for convenience

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\, B_z = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{6.32}$$

$$\frac{\partial B_z}{\partial t} + \operatorname{curl} \mathbf{E} = 0, \tag{6.33}$$

$$\operatorname{div} \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{6.34}$$

A condition for well-posedness of the Maxwell equations is that the sources $\mathbf{J}$ and ρ verify the continuity equation

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{J} = 0. \tag{6.35}$$

This is the case when the sources are computed from the Vlasov equation.

In the Yee scheme the discrete values of B_z is taken at the the vertices of the dual mesh, the discrete values of $\mathbf{E}$ and $\mathbf{J}$ at the midpoints of the edges, more precisely the x components at $(i + \frac{1}{2}, j)$ and the y components at $(i, j + \frac{1}{2})$ (see Figure 6.6). Then taking centered differences of the space

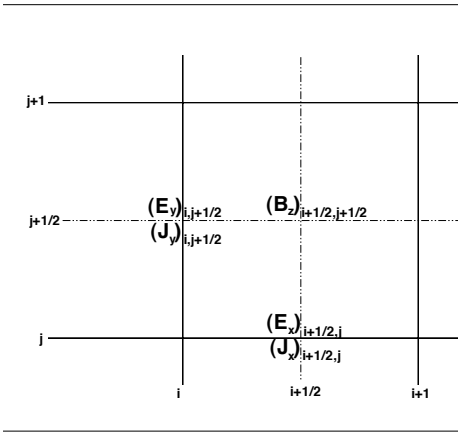

Fig. 6.6. Position of unknowns for the Yee scheme.

derivatives appearing in Maxwell's equations and a leap-frog scheme in time we get

$$\frac{B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}} - B^{n-\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}}}{\Delta t} + \frac{E^n_{x,i+\frac{1}{2},j+1} - E^n_{x,i+\frac{1}{2},j}}{\Delta y} - \frac{E_{y,i+1,j+\frac{1}{2}} - E^n_{y,i,j+\frac{1}{2}}}{\Delta x} = 0, \tag{6.36}$$

$$\frac{E^{n+1}_{x,i+\frac{1}{2},j} - E^n_{x,i+\frac{1}{2},j}}{\Delta t} - c^2 \frac{B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y} = -\frac{1}{\varepsilon_0} J^{n+\frac{1}{2}}_{x,i+\frac{1}{2},j}, \tag{6.37}$$

$$\frac{E^{n+1}_{y,i,j+\frac{1}{2}} - E^n_{y,i,j+\frac{1}{2}}}{\Delta t} + c^2 \frac{B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i-\frac{1}{2},j+\frac{1}{2}}}{\Delta x} = -\frac{1}{\varepsilon_0} J^{n+\frac{1}{2}}_{y,i,j+\frac{1}{2}},. \tag{6.38}$$

Putting ρ at the nodes of the primal cell he Gauss law can be discretized in the same manner using second order finite differences, which yields

$$\frac{E^n_{x,i+\frac{1}{2},j} - E^n_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{E^n_{y,i,j+\frac{1}{2}} - E^n_{y,i,j-\frac{1}{2}}}{\Delta y} = \frac{\rho_{i,j}}{\varepsilon_0}. \tag{6.39}$$

However this equation is redundant provided it is satisfied at the initial time and the following discrete continuity equation is satisfied:

$$\frac{\rho^{n+1}_{i,j} - \rho^n_{i,j}}{\Delta t} + \frac{J^{n+\frac{1}{2}}_{x,i+\frac{1}{2},j} - J^{n+\frac{1}{2}}_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{J^{n+\frac{1}{2}}_{y,i,j+\frac{1}{2}} - J^{n+\frac{1}{2}}_{y,i,j-\frac{1}{2}}}{\Delta y} = 0. \tag{6.40}$$

Proposition 12 *When Maxwell's equation are advanced in time using the Yee scheme* (6.36)-(6.38)*, the discrete Gauss law* (6.39) *is satisfied at time* t_0 *and the discrete continuity equation* (6.40) *is satisfied, then the discrete Gauss law* (6.39) *is satisfied for all discrete times* t_n*.*

Proof. Taking the discrete divergence of Ampère's law (6.37)-(6.38), in the same form in appears in (6.39), we get

$$\begin{aligned}
\frac{1}{\Delta t}\Bigg(& \frac{E^{n+1}_{x,i+\frac{1}{2},j} - E^{n+1}_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{E^{n+1}_{y,i,j+\frac{1}{2}} - E^{n+1}_{y,i,j-\frac{1}{2}}}{\Delta y} - \frac{E^n_{x,i+\frac{1}{2},j} - E^n_{x,i-\frac{1}{2},j}}{\Delta x} \\
& - \frac{E^n_{y,i,j+\frac{1}{2}} - E^n_{y,i,j-\frac{1}{2}}}{\Delta y}\Bigg) \\
& - c^2 \frac{B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j-\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i-\frac{1}{2},j+\frac{1}{2}} + B^{n+\frac{1}{2}}_{z,i-\frac{1}{2},j-\frac{1}{2}}}{\Delta y \Delta x} \\
& + c^2 \frac{B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i-\frac{1}{2},j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{z,i+\frac{1}{2},j-\frac{1}{2}} + B^{n+\frac{1}{2}}_{z,i-\frac{1}{2},j-\frac{1}{2}}}{\Delta x \Delta y} \\
& = -\frac{1}{\varepsilon_0}\left(\frac{J^{n+\frac{1}{2}}_{x,i+\frac{1}{2},j} - J^{n+\frac{1}{2}}_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{J^{n+\frac{1}{2}}_{y,i,j+\frac{1}{2}} - J^{n+\frac{1}{2}}_{y,i,j-\frac{1}{2}}}{\Delta y} \right).
\end{aligned}$$

The terms in B_z vanish and the right hand side can be replaced by the discrete time derivative in ρ using the discrete continuity equation so that it finally remains

$$\frac{E^{n+1}_{x,i+\frac{1}{2},j} - E^{n+1}_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{E^{n+1}_{y,i,j+\frac{1}{2}} - E^{n+1}_{y,i,j-\frac{1}{2}}}{\Delta y} - \frac{\rho^{n+1}_{i,j}}{\varepsilon_0} = \frac{E^{n}_{x,i+\frac{1}{2},j} - E^{n}_{x,i-\frac{1}{2},j}}{\Delta x} + \frac{E^{n}_{y,i,j+\frac{1}{2}} - E^{n+1}_{y,i,j-\frac{1}{2}}}{\Delta y} - \frac{\rho^{n}_{i,j}}{\varepsilon_0},$$

which tells us that if the discrete Gauss' law (6.39) is verified at time t_n it will still be verified at time t_{n+1}.

Remark 16 *The discrete Gauss law plays an important role, as we will see, when Maxwell's equations are coupled with Vlasov equations.*

6.4.2 Variationnal formulations for the 2D Maxwell equations

In order to introduce the Finite Element formulation, we need the variational form of Maxwell's equations (6.32)-(6.34). The unknowns of Maxwell's equations live in related function spaces.

Indeed in two dimensions we have the following diagrams, which are called *exact sequences*:

$$H^1(\Omega) \overset{grad}{\longrightarrow} H(\mathrm{curl}\,,\Omega) \overset{\mathrm{curl}}{\longrightarrow} L^2(\Omega)$$

and

$$H(\mathbf{curl}\,,\Omega) \overset{\mathbf{curl}}{\longrightarrow} H(\mathrm{div}\,,\Omega) \overset{div}{\longrightarrow} L^2(\Omega).$$

Note that $\mathbf{curl}\,\varphi \in L^2(\Omega)^2$ is equivalent to $\nabla\varphi \in L^2(\Omega)^2$ and hence the spaces $H(\mathbf{curl}\,,\Omega)$ and $H^1(\Omega)$ are identical.

The exact sequences mean that the image of the space on the left hand side of the arrow by the operator on top of the arrow is included in the space on the right hand side of the arrow and moreover that it is equal to the kernel of the following operator.

In order to conserve at the discrete level the properties of the continuous equations, it is necessary to look for the unknowns in the spaces associated to one of the exact sequence and introduce discrete spaces satisfying the same exact sequence property. This will be convenient in practice as one of Ampère's law or Faraday's law can be kept in strong form removing the need for inverting a mass matrix. Moreover this setting enables to prove stability and convergence in a very general manner [5, 4].

Thus, we have two choices for the variational form, either to use a weak form of Ampère's law and keep Faraday's law as it is or the opposite. The first option then consists in looking for $\mathbf{E} \in H(\mathrm{curl}\,,\Omega)$ and $B_z \in L^2(\Omega)$ using the last arrow of the first diagram and the second option in looking for $\mathbf{E} \in H(div,\Omega)$ and $B_z \in H^1(\Omega) = H(\mathbf{curl}\,,\Omega)$ using the first arrow of the second diagram. Let us consider here the first option.

We shall need the following 2D Green formulae to derive the variational formulations:

$$\int_\Omega \mathbf{F}\cdot\mathbf{curl}\, C - \int_\Omega \operatorname{curl}\mathbf{F}\, C = \int_{\partial\Omega} (\mathbf{F}\cdot\boldsymbol{\tau})\, C \quad \forall \mathbf{F}\in H(\operatorname{curl},\Omega),\ C\in H^1(\Omega), \tag{6.41}$$

$$\int_\Omega \operatorname{div}\mathbf{F}\, q + \int_\Omega \mathbf{F}\cdot\nabla q = \int_{\partial\Omega} (\mathbf{F}\cdot\mathbf{n})\, q \quad \forall \mathbf{F}\in H(\operatorname{div},\Omega),\ q\in H^1(\Omega). \tag{6.42}$$

We shall look for $\mathbf{E}\in H(\operatorname{curl},\Omega)$ and $B_z\in L^2(\Omega)$. In order to get the weak form, we take the scalar product of (6.32) with a test function $\mathbf{F}\in H(curl,\Omega)$ and integrate on Ω

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} - c^2\int_\Omega \mathbf{curl}\, B_z\cdot\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F}.$$

Using the Green formula (6.41) this becomes, assuming the boundary term vanishes, which is the case for perfect conductor or periodic boundary conditions

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} + c^2\int_\Omega B_z \operatorname{curl}\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F} \quad \forall \mathbf{F}\in H(\operatorname{curl},\Omega). \tag{6.43}$$

For Faraday's law, we multiply by $C\in L^2(\Omega)$ and integrate on Ω which yields

$$\frac{d}{dt}\int_\Omega B_z\, C + \int_\Omega \operatorname{curl}\mathbf{E}\, C = 0 \quad \forall C\in L^2(\Omega). \tag{6.44}$$

For Gauss's law, we multiply by $q\in H^1(\Omega)$ and integrate on Ω which yields

$$\int_\Omega \operatorname{div}\mathbf{E}\, q = \frac{1}{\varepsilon_0}\int_\Omega \rho\, q.$$

And using (3.25) and assuming that $\mathbf{E}\cdot\mathbf{n}=0$ on the boundary or periodic boundary conditions, we obtain that

$$-\int_\Omega \mathbf{E}\cdot\nabla q = \frac{1}{\varepsilon_0}\int_\Omega \rho\, q \quad \forall q\in H^1(\Omega). \tag{6.45}$$

Finally, let us also write the weak form of the continuity equation (6.35), that we multiply by $q\in H^1(\Omega)$ and integrate using (6.42) and the fact that $\mathbf{J}\cdot\mathbf{n}=0$ on the boundary or periodic boundary conditions, like in the previous case. We then get

$$\frac{d}{dt}\int_\Omega \rho\, q = \int_\Omega \mathbf{J}\cdot\nabla q \quad \forall q\in H^1(\Omega). \tag{6.46}$$

Remark 17 *Let us note that if the weak form of Gauss' law (6.45) is verified at $t=0$ and if the weak continuity equation (6.46) is verified, the weak form of Gauss' law is verified for all times. Indeed, for any $q\in H^1(\Omega)$,* $\operatorname{curl}\nabla q = 0$

and hence $\nabla q \in H(\mathrm{curl}\,, \Omega)$*, then we can use* ∇q *as a test function in* (6.43)*, which becomes*

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\nabla q = -\int_\Omega \mathbf{J}\cdot\nabla q \quad \forall q \in H^1(\Omega).$$

Then, using (6.46),

$$\frac{d}{dt}\big(\int_\Omega \mathbf{E}\cdot\nabla q + \int_\Omega \rho\, q\big) = 0 \quad \forall q \in H^1(\Omega),$$

which gives the desired result.

Let us now introduce the second variational formulation which involves $\mathbf{E} \in H(div, \Omega)$ and $B_z \in H^1(\Omega)(= H(\mathbf{curl}\,, \Omega))$. In order to get the variationl form, we take the scalar product of (6.32) with a test function $\mathbf{F} \in H(div, \Omega)$ and integrate on Ω

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} - c^2\int_\Omega \mathbf{curl}\, B_z\cdot\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F}, \quad \forall \mathbf{F} \in H(div, \Omega). \quad (6.47)$$

For Faraday's law, we multiply by $C \in L^2(\Omega)$ and integrate on Ω which yields

$$\frac{d}{dt}\int_\Omega B_z\, C + \int_\Omega \mathrm{curl}\,\mathbf{E}\, C = 0,$$

Using the Green formula (3.66) this becomes

$$\frac{d}{dt}\int_\Omega B_z\, C + \int_\Omega \mathbf{E}\cdot\mathbf{curl}\, C = 0 \quad \forall C \in H^1(\Omega). \quad (6.48)$$

6.4.3 Discretization using conforming finite elements

In order to keep the specific features of Maxwell's equations at the discrete level which are useful in different contexts, we shall consider finite dimensional subspaces endowed with the same exact sequence structure as in the continuous level [4, 5, 69]. Let us first derive the linear system that comes out of this discretization. Let $\{\boldsymbol{\psi}_i\}_{i=1\ldots N}$ be a basis of $W \subset H(\mathrm{curl}, \Omega)$ and $\{\varphi_k\}_{k=1\ldots M}$ a basis of $V \subset L^2(\Omega)$. Then denoting by σ_i^W the degrees of freedom associated to $\{\boldsymbol{\psi}_i\}_{i=1\ldots N}$ and by σ_k^V the degrees of freedom associated to $\{\varphi_k\}_{k=1\ldots M}$. In the case of Lagrange Finite Elements, this degrees of freedom are just point values. We can write elements of W and V respectively

$$\mathbf{E}_h = \sum_{i=1}^{N} \sigma_i^W(\mathbf{E}_h)\boldsymbol{\psi}_i, \qquad B_h = \sum_{k=1}^{M} \sigma_k^V(B_h)\varphi_k.$$

Replacing the continuous spaces by the discrete spaces in the variational formulations (6.43) and (6.44) we get the following discrete problem:
Find $(\mathbf{E}_h, B_h) \in W \times V$ such that

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\boldsymbol{\psi_i}\,d\mathbf{x} - \int_\Omega B\,(\operatorname{curl}\boldsymbol{\psi}_i)\,d\mathbf{x} = -\int_\Omega \mathbf{J}\cdot\boldsymbol{\psi}_i\,d\mathbf{x},\quad \forall i = 1\dots N, \tag{6.49}$$

$$\frac{d}{dt}\int_\Omega B\,\varphi_k\,d\mathbf{x} + \int_\Omega (\operatorname{curl}\mathbf{E})\,\varphi_k\,d\mathbf{x} = 0,\qquad \forall k = 1\dots M, \tag{6.50}$$

which becomes when $\mathbf{E}$ and B are decomposed on the respective bases of W and V

$$M_W\frac{dE}{dt} - KB = J \tag{6.51}$$

$$M_V\frac{dB}{dt} + K^TE = 0 \tag{6.52}$$

where $E = (\sigma_i^W(\mathbf{E}_h))_{1\le i\le N}$ (resp. $B = (\sigma_k^V(B_h))_{1\le k\le M}$) denote vectors of degrees of freedom for the discrete electric and magnetic fields, with

$$(M_W)_{1\le i,j\le N} = \int_\Omega \boldsymbol{\psi}_j\cdot\boldsymbol{\psi}_i\,d\mathbf{x},\ (M_V)_{1\le i,j\le M} = \int_\Omega \varphi_i\varphi_j\,d\mathbf{x},$$

$$(K)_{1\le i\le N,\,1\le j\le M} = \int_\Omega \varphi_j(\operatorname{curl}\boldsymbol{\psi}_i)\,d\mathbf{x}.$$

Remark 18 *Notice that the structure of this linear system is exactly the same as in the 1D case. Hence the same time schemes with the same properties can be used.*

As we already noticed in the 1D case, the exact sequence structure of our Finite Element spaces enables us to express the discrete variational formulation where no Green formula (or integration by parts) was used (6.50) in our case by a strong form. Indeed the exact sequence structure of our discrete spaces implies in particular that if $\mathbf{E}_h \in W$, we have $\operatorname{curl}\mathbf{E}_h \in V$. Hence we can express $\operatorname{curl}\mathbf{E}_h$ on the basis $\{\varphi_k\}_{k=1,\dots,M}$ of V which yields

$$\begin{aligned}\operatorname{curl}\mathbf{E}_h = \sum_{l=1}^M \sigma_l^V(\operatorname{curl}\mathbf{E}_h)\varphi_l &= \sum_{l=1}^M \sigma_l^V\Big(\sum_{j=1}^N \sigma_j^W(\mathbf{E}_h)(\operatorname{curl}\boldsymbol{\psi}_j)\Big)\varphi_l\\ &= \sum_{l=1}^M\sum_{j=1}^N \sigma_j^W(\mathbf{E}_h)\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)\varphi_l.\end{aligned}$$

In particular we get that $\sigma_l^V(\operatorname{curl}\mathbf{E}_h) = \sum_{j=1}^N \sigma_j^W(\mathbf{E}_h)\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)$, and injecting this expression in the discrete Faraday law

$$\begin{aligned}\int_\Omega (\operatorname{curl}\mathbf{E})\varphi_k\,d\mathbf{x} &= \int_\Omega \sum_{l=1}^M\sum_{j=1}^N \sigma_j^W(\mathbf{E})\sigma_l^V(\operatorname{curl}\boldsymbol{\psi_j})\varphi_l\varphi_k\,d\mathbf{x}\\ &= \sum_{l=1}^M\sum_{j=1}^N \int_\Omega \varphi_l\varphi_k\,d\mathbf{x}\,\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)\,\sigma_j^W(\mathbf{E}),\end{aligned}$$

which is the k^{th} line (for k from 1 to M) of the vector $M_V RE$ where R is the matrix defined by

$$(R)_{1\leq i\leq M,\,1\leq j\leq N} = \sigma_i^V(\operatorname{curl}\boldsymbol{\psi}_j),$$

so that, as M_V is non singular, the system (6.51)-(6.52) is algebraically equivalent to the system

$$M_W\frac{dE}{dt} - KB = J, \tag{6.53}$$

$$\frac{dB}{dt} + RE = 0, \tag{6.54}$$

This new formulation yields an explicit expression of B which can then be computed without solving a linear system.

Now having expressed the general structure that we want our Finite Element spaces to have and its consequences, there is still a wide variety of choices of Finite Element spaces that verify these exact sequence property. Let us define some classical spaces on quads and triangles.

On rectangular meshes the cells of which are denoted by K_i, $1 \leq i \leq r$, the three actual subspaces of the exact sequence $X \subset H^1(\Omega)$, $W \subset H(\operatorname{curl},\Omega)$ and $V \subset L^2(\Omega)$ can be defined as follows

$$X = \{\chi \in H^1(\Omega) \;\mid\; \chi_{|K_i} \in \mathbb{Q}_k(K_i)\},$$

$$W = \{\boldsymbol{\psi} \in H(\operatorname{curl},\Omega) \;\mid\; \boldsymbol{\psi}_{|K_i} \in \begin{pmatrix}\mathbb{Q}_{k-1,k}(K_i)\\ \mathbb{Q}_{k,k-1}(K_i)\end{pmatrix},\ \forall i = 1,\ldots,r\},$$

$$V = \{\varphi \in L^2(\Omega) \;\mid\; \varphi_{|K_i} \in \mathbb{Q}_{k-1}(K_i),\ \forall i = 1,\ldots,r\}.$$

where $\mathbb{Q}_{m,n} = \left\{x^i y^j,\ 0 \leq i \leq m, 0 \leq j \leq n\right\}$, with the particular case $\mathbb{Q}_{m,m}$ is simply denoted by $\mathbb{Q}_m$ in the classical way.

All these Finite Element spaces are piecewise polynomials for scalar fields in the case of X and V and for each component of a field for W. In the case of V which is just in $L^2(\Omega)$ there is no additional continuity requirement at the cell interface. In the case of X the inclusion in $H^1(\Omega)$ imposes continuity at the cell interface and in the case of W the inclusion in $H(\operatorname{curl},\Omega)$ imposes continuity of the tangential component of the field at the cell interface.

The space $\mathbb{Q}_k$ is the standard continuous Lagrange Finite Element space on quads. The space W is known as the first family of edge elements $H(\operatorname{curl})$-conforming of Nédélec [87] and the space V is the space of discontinuous functions which restrict to a polynomial of degree $k-1$ with respect to each variable on each cell. This is the kind of approximation used in Discontinuous Galerkin methods.

After having defined the spaces there are still many choices for the degrees of freedom which define the actual basis functions. In the interpretation of Maxwell's equations as differential forms it is natural to take the degrees of

freedom for X which corresponds to 0-forms as point values, the degrees of freedom for W which corresponds to 1-forms as edge integrals, and the degrees of freedom for V which corresponds to 2-forms as cell integrals. But such a choice is not mandatory. The Cohen Monk Finite Elements for W are based on Lagrange degrees of freedom (point values) for Maxwell at Gauss or Gauss-Lobatto points. This has the advantage of leading to a diagonal mass matrix M_W if the mesh is cartesian thanks to the Gauss-Lobatto quadrature formula [38, 37].

Let us now introduce in detail the degrees of freedom that are obtained by an interpretation in terms of differential forms, which have very convenient properties. For this, two types of 1D basis functions are needed in a tensor product construction on quads, first the nodal basis functions which typically are the standard Lagrange basis functions and then the edge basis functions, which are constructed from the nodal basis functions and whose degrees of freedom are edge integrals. Let us consider the 1D reference element $[-1, 1]$ on which we define the Gauss-Lobatto points $(\hat{x}_i)_{0\leq i\leq k}$. Note that uniform interpolation points could also be used for the construction, but they are not stable for higher degrees. We denote by $l_{k,i}$, $0 \leq i \leq k$ the Lagrange basis functions associated to these points. This will define the local basis, on each element, of the discrete space X. This is a standard Lagrange Finite Element for which the degrees of freedom are the point values at the interpolation points and we have $l_{k,i}(\hat{x}_j) = \delta_{i,j}$.

Our aim is now to use this Lagrange basis to construct a basis e_i $0 \leq i \leq k-1$ for W the next space in the sequence. This should be such that the derivatives of linear combinations of the Lagrange basis functions are exactly represented. It will be natural to define the degrees of freedom of this space to be integrals between two successive interpolation points, so that the degrees of freedom of a derivative $u = \frac{d\phi}{dx}$ can be expressed directly with respect to the degrees of freedom of ϕ in X. Indeed

$$\int_{\hat{x}_\nu}^{\hat{x}_{\nu+1}} \frac{d\phi}{d\hat{x}}(\hat{x})\,\mathrm{d}\hat{x} = \phi(\hat{x}_{\nu+1}) - \phi(\hat{x}_\nu).$$

Next we find that the basis associated to these degrees of freedom, *i.e.* verifying

$$\int_{\hat{x}_{j-1}}^{\hat{x}_j} e_i(\hat{x})\,d\hat{x} = \delta_{i,j}, \quad 1 \leq j \leq k.$$

This can be expressed (see [64]) by

$$e_i(\hat{x}) = -\sum_{\nu=0}^{i-1} \frac{dl_{k,\nu}}{d\hat{x}}(\hat{x}), \quad 1 \leq j \leq k. \tag{6.55}$$

Indeed, we have for $1 \leq i, j \leq k$

$$\int_{\hat{x}_{j-1}}^{\hat{x}_j} e_i(\hat{x})\,d\hat{x} = -\sum_{\nu=0}^{i-1}\int_{\hat{x}_{j-1}}^{\hat{x}_j} \frac{dl_{k,\nu}}{d\hat{x}}(\hat{x})\,d\hat{x} = -\sum_{\nu=0}^{i-1}(l_{k,\nu}(\hat{x}_j) - l_{k,\nu}(\hat{x}_{j-1}))$$
$$= -\sum_{\nu=0}^{i-1}(\delta_{\nu,j} - \delta_{\nu+1,j}) = -(\delta_{0,j} - \delta i, j) = \delta_{i,j}$$

as $\delta_{0,j} = 0$ for all $1 \leq j \leq k$.

Now having the local 1D basis functions $(l_{k,i})_{0\leq i\leq k}$ and $(e_i)_{1\leq i\leq k}$, the local 2D basis functions are defined using products of this basis functions.

The local basis functions defining X are the classical $\mathbb{Q}_k$ basis functions $l_k(x,y) = l_{k,i}(x)l_{k,j}(y)$ and the degrees of freedom the values at the points $\hat{x}_i\hat{y}_j$ $0 \leq i,j \leq k$, where the $\hat{x}_i$ as well as the $\hat{y}_j$ are the $k+1$ Gauss-Lobatto points.

Let us denote by $\mathbf{P}$ the set of local basis functions on which W is build

$$\mathbf{P} = \left\{ \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \text{ with } p_1 \in \mathbb{Q}_{k-1,k}, p_2 \in \mathbb{Q}_{k,k-1} \right\}.$$

Hence the basis functions are of two forms

$$\mathbf{e}^1_{i,j} = \begin{pmatrix} e_i(x)l_{k,j}(y) \\ 0 \end{pmatrix},\ 1 \leq i \leq k, 0 \leq j \leq k \quad \mathbf{e}^2_{i,j} = \begin{pmatrix} 0 \\ l_{k,i}(x)e_j(y) \end{pmatrix},\ 0 \leq i \leq k, 1 \leq j \leq k.$$

The degrees of freedom associated to the first set of basis functions are of the form

$$\mathbf{p} \mapsto \int_{\hat{x}_{i-1}}^{\hat{x}_i} p_1(x, y_j)\,dx \quad 1 \leq i \leq k, 0 \leq j \leq k,$$

and the degrees of freedom associated to the second set of basis functions are of the form

$$\mathbf{p} \mapsto \int_{\hat{y}_{j-1}}^{\hat{y}_j} p_2(x_i, y)\,dy \quad 0 \leq i \leq k, 1 \leq j \leq k.$$

Finally for V the local polynomial space is $\mathbf{Q}_{k-1}$ and the degrees of freedom are cell integrals. Hence the local basis functions can be expressed as $s(x,y) = e_i(x)e_j(y)$ $1 \leq i,j \leq k$ and the corresponding degrees of freedom are defined by

$$p \in \mathbf{Q}_{k-1} \mapsto \int_{\hat{x}_{i-1}}^{\hat{x}_i}\int_{\hat{y}_{j-1}}^{\hat{y}_j} p(x,y)\,dx\,dy.$$

These compatible Finite Element spaces enable to express the strong form of one of the Maxwell's equations in a very simple for directly relating the degrees of freedom using only the connectivity of the mesh and no geometric information like distances, lengths or areas. It is not modified by a compatible mapping.

On more general quad meshes the Finite Elements are defined via a bilinear transformation from a rectangular reference element associated to the spaces

defined for one single element in the rectangular case. This is straightforward for the spaces of scalar fields X and V. But care must be taken for the space of vector fields W for which the vector valued basis functions need to be transformed in a covariant way to preserve the inclusion in $H(\mathrm{curl}\,,\Omega)$ of the discrete space W.

On triangular cells the discrete spaces are defined by

$$W = \{\boldsymbol{\psi} \in H(\mathrm{curl},\Omega) \;\mid\; \boldsymbol{\psi}_{|T_i} \in \mathbb{P}^2_{k-1}(T_i) + \overline{\mathbb{P}}_{k-1}(T_i) \begin{pmatrix} y \\ -x \end{pmatrix} \;,\; \forall i = 1,\ldots,r\},$$

$$V = \{\varphi \in L^2(\Omega) \;\mid\; \varphi_{|T_i} \in \mathbb{P}_{k-1}(T_i) \;,\; \forall i = 1,\ldots,r\},$$

where $\overline{\mathbb{P}}_{k-1}$ denotes the set of polynomials of degree exactly $k-1$. The space V is $\mathbb{P}_{k-1}$ on each element and discontinuous across element boundaries (conforming in $L^2(\Omega)$), so is straightforward to construct. For the space W, we have again used the first family of edge elements of Nédélec [87], conforming in $H(\mathrm{curl},\Omega)$, but we have changed the degrees of freedom.

6.5 Discretization of the 3D Maxwell equations

The discretization of the 3D Maxwell is similar to the 2D case it is base on the exact sequence

$$H^1(\Omega) \xrightarrow{grad} H(\mathrm{curl}\,,\Omega) \xrightarrow{\mathrm{curl}} H(\mathrm{div}\,,\Omega) \xrightarrow{\mathrm{div}} L^2(\Omega),$$

and the variational formulations:

Find $(\mathbf{E},\mathbf{B}) \in H(\mathrm{curl}\,,\Omega) \times H(\mathrm{div}\,,\Omega)$ *such that*

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} + c^2 \int_\Omega \mathbf{B}\cdot\mathbf{curl}\,\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F} \quad \forall \mathbf{F} \in H(\mathrm{curl}\,,\Omega). \tag{6.56}$$

$$\frac{d}{dt}\int_\Omega \mathbf{B}\cdot\mathbf{C} + \int_\Omega \mathbf{curl}\,\mathbf{E}\cdot\mathbf{C} = 0 \quad \forall \mathbf{C} \in H(\mathrm{div}\,,\Omega). \tag{6.57}$$

$$-\int_\Omega \mathbf{E}\cdot\nabla q = \frac{1}{\varepsilon_0}\int_\Omega \rho\, q \quad \forall q \in H^1(\Omega). \tag{6.58}$$

6.6 The Discontinuous Galerkin (DG) method

We shall introduce the method for the TE mode of the 2D Maxwell equations. The DG method can be most naturally expressed using hyperbolic conservation laws. For this reason, in view of the coupling with the Vlasov equation, we consider here the generalized form of the Maxwell equations with a hyperbolic correction in order to correct numerical effects linked to the fact that a discrete continuity equation is not exactly verified [83, 82].

In this cas the TE mode of the 2D Maxwell equations becomes

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\ B + \gamma c^2 \nabla p = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{6.59}$$

$$\frac{\partial B}{\partial t} + \text{curl}\ \mathbf{E} = 0, \tag{6.60}$$

$$\frac{\partial p}{\partial t} + \text{div}\ \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{6.61}$$

and for the TM mode we have

$$\frac{\partial E}{\partial t} - c^2 \text{curl}\ \mathbf{B} = -\frac{1}{\varepsilon_0} J, \tag{6.62}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{curl}\ E = 0. \tag{6.63}$$

with $\mathbf{E} = (E_x, E_y)^T$, $E = E_z$, $\mathbf{B} = (B_x, B_y)^T$, $B = B_z$, $\mathbf{curl}\ B_z = (\partial_y B_z, -\partial_x B_z)^T$, curl $\mathbf{E} = \partial_x E_y - \partial_y E_x$, and div $\mathbf{E} = \partial_x E_x + \partial_y E_y$.

The unkonwn p, which can be proved to be 0 at the continuous level if the continuity equation $\frac{\partial \rho}{\partial t} + \text{div}\,\mathbf{J} = 0$ is satisfied, has been added to absorb the numerical errors occurring from the continuity equation only being satisfied approximately at the discrete level. The γ factor in front of $c^2\nabla p$ enables to fix the velocity of the wave transporting p out of the computational domain. In general it is taken slightly larger than 1, so that this non physical wave propagates a little bit faster than the electromagnetic waves.

6.6.1 Principle of the method

In order to define a Discontinuous Galerkin method, one needs a mesh of the computational domain consisting of disjoint cells. Then the unknown is approximated on each cell in a finite dimensional linear space (generally polynomial) which could be different in each cell. We shall consider only approximations based on $(\mathbb{P}_k)^p$ for an unknown with values in $\mathbb{R}^p$. In our cas, for the two-dimensional Maxwell equations we consider polynomials of two variables with $p = 3$, the unknown being $\mathbf{u} = (E_x, E_y, B_z)^T$. Recall that $\mathbb{P}_k$ denotes the space of polynomials of degree less or equal to k. In two space dimensions where we work the dimension of $\mathbb{P}_k$ is $\frac{(k+1)(k+2)}{2}$. On a given cell K we look for an approximation $\mathbf{u}_h = (E_{h,x}, E_{h,y}, B_{h,z})^T \in (\mathbb{P}_k(K))^3$. Using the Galerkin principle we shall derive an equation on $\mathbf{u}_h$ by multiplying the equations by a test function belonging to the same space. However in the Discontinous Galerkin case, the integration shall take place on one unique cell instead of the whole computational domain for the standard Galerkin method. We shall detail the derivation for the TE mode. It can be performed in a similar way for the TM mode.

Remark 19 *The specificity of the Discontinuous Galerkin method with respect to the standard Galerkin method is that the weak form is defined on each*

cell with no continuity constraint between two neighboring cells. A definition of a unique numerical flux at the cell interface is needed to transfer the information from one cell to the other in a consistent manner. This numerical flux and the cell approximation are the building blocks of a DG method.

Let us consider a test function $\mathbf{v}_h = (F_{h,x}, F_{h,y}, C_{h,z})^T \in (\mathbb{P}_k(K))^3$, and take the dot product of the Ampere equation (6.59) with $\mathbf{F}_h = (F_{h,x}, F_{h,y})^T$, multiply the Faraday equation (6.60) by $C_{h,z}$ and the Gauss equation (6.61) by q_h and integrate over K. We then obtain

$$\frac{d}{dt}\int_K \mathbf{E}_h \cdot \mathbf{F}_h - c^2 \int_K \mathbf{curl}\ B_{h,z} \cdot \mathbf{F}_h + \gamma c^2 \int_K \nabla p_h \cdot \mathbf{F}_h = -\frac{1}{\varepsilon_0}\int_K \mathbf{J}_h \cdot \mathbf{F}_h,$$

$$\frac{d}{dt}\int_K B_{h,z} C_{h,z} + \int_K \mathrm{curl}\ \mathbf{E}_h\, C_{h,z} = 0,$$

and on the other hand

$$\frac{d}{dt}\int_K p_h q_h + \int_K \mathrm{div}\, \mathbf{E}_h\, q_h = \frac{1}{\varepsilon_0}\int_K \rho\, q_h.$$

In order to introduce the numerical flux which will link local solutions on neighboring cells, it is necessary to integrate all those equations by parts in order to let a boundary term appear. To this aim we shall need the following Green functions:

$$\int_K \mathbf{F} \cdot \mathbf{curl}\, C - \int_K \mathrm{curl}\, \mathbf{F}\, C = \int_{\partial K} (\mathbf{F} \cdot \boldsymbol{\tau})\, C \quad \forall \mathbf{F} \in H(\mathrm{curl}\,, K),\ C \in H^1(K), \tag{6.64}$$

$$\int_K \mathbf{F} \cdot \nabla q + \int_K \mathrm{div}\, \mathbf{F}\, q = \int_{\partial K} (\mathbf{F} \cdot \mathbf{n})\, q \quad \forall \mathbf{F} \in H(div, K),\ q \in H^1(K), \tag{6.65}$$

where $\boldsymbol{\tau} = (n_y, -n_x)^T$ is the tangent vector on the cell boundary ∂K, $\mathbf{n} = (n_x, n_y)^T$ defining the outbound normal vector. We thus obtain

$$\int_K \mathbf{curl}\ B_{h,z} \cdot \mathbf{F}_h = \int_K B_{h,z} \mathrm{curl}\, \mathbf{F}_h + \int_{\partial K} (\mathbf{F}_h \cdot \boldsymbol{\tau})\, B_{h,z},$$

$$\int_K \mathrm{curl}\ \mathbf{E}_h\, C_{h,z} = \int_K \mathbf{E}_h \cdot \mathbf{curl}\ C_{h,z} - \int_{\partial K} (\mathbf{E}_h \cdot \boldsymbol{\tau})\, C_{h,z},$$

$$\int_K \nabla p_h \cdot \mathbf{F}_h = -\int_K p_h\, \mathrm{div}\, \mathbf{E}_h + \int_{\partial K} p_h\, (\mathbf{F}_h \cdot \mathbf{n}).$$

and

$$\int_K \mathrm{div}\, \mathbf{E}_h\, q_h = -\int_K \mathbf{E}_h \cdot \nabla q_h + \int_{\partial K} (\mathbf{E}_h \cdot \mathbf{n})\, q_h.$$

The unknowns being inherently discontinuous at the cell interfaces in the DG method, the boundary terms are not well defined, at least not consistently on the left hand side and the right hand side of the interface. In order to specify

the DG method a numerical approximation of these fluxes which is identical on both sides of the interface needs to be defined. These are called the numerical fluxes as is the case in Finite Volume methods. Note that the Finite Volume method can be seen as a special case of a DG method for local polynomial approximations of degree 0 in each cell. A simple an natural definition of the numerical flux consists in taking the average value of the left hand side and the right hand side values. This yields a central flux which has some advantages for Maxwell's equations. This flux will lead to an exact energy conservation at the discrete level as is the case for standard elements. In this case as we will see, the matrix form of the DG discretization is identical to what we obtained for Finite Elements. However due to the discontinuities introduced in the DG formulation, it can be good idea to add a dissipation mechanism by considering upwind fluxes. In order to keep the freedom of specifying the numerical flux later, let us simply denote by B_N, $\mathbf{E}_N$ and p_N the values of the approximate electromagnetic flux and of p defined by the specific numerical flux. We then obtain the following discrete formulations:

$$\begin{aligned}&\frac{d}{dt}\int_K \mathbf{E}_h\cdot\mathbf{F}_h - c^2\int_K B_{h,z}\,\mathrm{curl}\,\mathbf{F}_h - \gamma c^2\int_K p_h\,\mathrm{div}\,\mathbf{F}_h\\ &- c^2\int_{\partial K}(\mathbf{F}_h\cdot\boldsymbol{\tau})\,B_N + \gamma c^2\int_{\partial K} p_N\,(\mathbf{F}_h\cdot\mathbf{n}) = -\frac{1}{\varepsilon_0}\int_K \mathbf{J}_h\cdot\mathbf{F}_h,\quad \forall\mathbf{F}_h\in\mathbb{P}_k^2,\end{aligned} \tag{6.66}$$

$$\frac{d}{dt}\int_K B_{h,z}\,C_{h,z} + \int_K \mathbf{E}_h\cdot\mathbf{curl}\,C_{h,z} - \int_{\partial K}(\mathbf{E}_N\cdot\boldsymbol{\tau})\,C_{h,z} = 0\quad \forall C_{h,z}\in\mathbb{P}_k, \tag{6.67}$$

$$\frac{d}{dt}\int_K p_h q_h - \int_K \mathbf{E}_h\cdot\nabla q_h + \int_{\partial K}(\mathbf{E}_N\cdot\mathbf{n})\,q_h = \frac{1}{\varepsilon_0}\int_K \rho\,q_h\quad \forall q_h\in\mathbb{P}_k. \tag{6.68}$$

6.6.2 Matrix formulation of the discrete problem

A classical Finite Element [35, 56] is defined by a triple (K, P, Σ), where K is the cell, P a finite dimensional function space of dimension N_k on K (in our case $\mathbb{P}_k(K)$) and $\Sigma = \{\sigma_1, \ldots, \sigma_{N_k}\}$ is a set of N_k linear forms enabling to identify uniquely an element of P. Thus to the linear forms of Σ is associated a unique basis $(\varphi_j)_{1\leq j\leq N_k}$ of P such that $\sigma_i(\varphi_j) = \delta_{i,j}$ where $\delta_{i,j}$ is the Kronecker symbol. We consider here only Lagrange Finite Elements where the linear forms $(\sigma_i)_{1\leq i\leq N_k}$ are associated to interpolation points $(x_i)_{1\leq i\leq N_k}$ in K. In this case we simply have for a function $G_h\in P$, $\sigma_i(G) = G(x_i)$.

Any element G_h of our finite dimensional function space P is thus defined uniquely by $(\sigma_i(G_h))_{1\leq i\leq N_k}$.

$$G_h(\mathbf{x}) = \sum_{j=1}^{N_k}\sigma_j(G_h)\varphi_j(\mathbf{x}),$$

Let us denote by $\mathbb{G} = (\sigma_1(G_h), \ldots, \sigma_{N_k}(G_h))^T$ the vector whose components are the degrees of freedom. We then obtain a matrix expression of the relations (6.66)-(6.68) by plugging in this expression. We first have

$$\begin{aligned}\int_K \mathbf{E}_h \cdot \mathbf{F}_h &= \int_K (E_{h,x}F_{h,x} + E_{h,y}F_{h,y})\,dxdy \\ &= \sum_{1\le i,j\le N_k} \sigma_j(E_{h,x})\sigma_i(F_{h,x}) \int_K \varphi_i\,\varphi_j\,dxdy \\ &+ \sum_{1\le i,j\le N_k} \sigma_j(E_{h,y})\sigma_i(F_{h,y}) \int_K \varphi_i\,\varphi_j\,dxdy \\ &= \mathbb{F}_x^T M_K \mathbb{E}_x + \mathbb{F}_y^T M_K \mathbb{E}_y,\end{aligned}$$

where the mass matrix on element K is defined by

$$M_K = ((\int_K \varphi_i\,\varphi_j\,dxdy))_{1\le i\le N_k, 1\le j\le N_k}.$$

In the same way

$$\begin{aligned}\int_K B_{h,z}\,C_{h,z} &= \int_K B_{h,z}C_{h,z}\,dxdy \\ &= \sum_{1\le i,j\le N_k} \sigma_j(B_{h,z})\sigma_i(C_{h,z}) \int_K \varphi_i\,\varphi_j\,dxdy \\ &= \mathbb{C}_z^T M_K \mathbb{B}_z.\end{aligned}$$

Then,

$$\begin{aligned}\int_K B_{h,z}\,\mathrm{curl}\,\mathbf{F}_h &= \int_K B_{h,z}(\partial_x F_{h,y} - \partial_y F_{h,x})\,dxdy \\ &= \sum_{1\le i,j\le N_k} \sigma_j(B_{h,z})\sigma_i(F_{h,y}) \int_K \varphi_j \partial_x\varphi_i\,dxdy \\ &\quad - \sum_{1\le i,j\le N_k} \sigma_j(B_{h,z})\sigma_i(F_{h,x}) \int_K \varphi_j \partial_y\varphi_i\,dxdy \\ &= \mathbb{F}_y^T D_K^x \mathbb{B}_z - \mathbb{F}_x^T D_K^y \mathbb{B}_z,\end{aligned}$$

where the derivative matrices D_K^x and D_K^y are defined respectively by

$$D_K^x = ((\int_K \varphi_j\,\partial_x\varphi_i\,dxdy))_{1\le i\le N_k, 1\le j\le N_k},$$

$$D_K^y = ((\int_K \varphi_j\,\partial_y\varphi_i\,dxdy))_{1\le i\le N_k, 1\le j\le N_k}.$$

The last non boundary term of element K is

$$\begin{aligned}\int_K \mathbf{E}_h \cdot \mathbf{curl}\, C_{h,z} &= \int_K (E_{h,x}\partial_y C_{h,z} - E_{h,y}\partial_x C_{h,z})\, dxdy \\ &= \sum_{1\le i,j\le N_k} \sigma_i(C_{h,z})\sigma_j(E_{h,x}) \int_K \varphi_j \partial_y \varphi_i \, dxdy \\ &\quad - \sum_{1\le i,j\le N_k} \sigma_i(C_{h,z})\sigma_j(E_{h,y}) \int_K \varphi_j \partial_x \varphi_i \, dxdy \\ &= \mathbb{C}_z^T D_K^y \mathbb{E}_x - \mathbb{C}_z^T D_K^x \mathbb{E}_y.\end{aligned}$$

Let us now handle the flux term. Let us denote by f an internal or external face of the mesh for which the tangent vector τ is constant and $\tilde{N}_k$, the number of degrees of freedom on the face. Then

$$\begin{aligned}\int_f (\mathbf{F}_h \cdot \boldsymbol{\tau})\, B_N &= \int_f (F_{h,x}\tau_x + F_{h,y}\tau_y) B_N \, d\sigma \\ &= \sum_{1\le i,j\le \tilde{N}_k} \sigma_j(B_N)\sigma_i(F_{h,x})\tau_{f,x} \int_f \varphi_i\varphi_j \, d\sigma \\ &\quad + \sum_{1\le i,j\le \tilde{N}_k} \sigma_j(B_N)\sigma_i(F_{h,y})\tau_{f,y} \int_f \varphi_i\varphi_j \, d\sigma \\ &= \tau_{f,x}\mathbb{F}_x^T M_f \mathbb{B}_N + \tau_{f,y}\mathbb{F}_y^T M_f \mathbb{B}_N,\end{aligned}$$

where $M_f = \int_f \varphi_i\varphi_j \, d\sigma$ is the masse matrix on face f. In the same way

$$\begin{aligned}\int_f (\mathbf{E}_N \cdot \boldsymbol{\tau})\, C_{h,z} &= \int_f (E_{N,x}\tau_x + E_{N,y}\tau_y) C_{h,z} \, d\sigma \\ &= \tau_{f,x}\mathbb{C}^T M_f \mathbb{E}_{\mathbb{N}x} + \tau_{f,y}\mathbb{C}^T M_f \mathbb{E}_{\mathbb{N}y}.\end{aligned}$$

Remark 20 *Let us notice that the matrix relation representing the terms interior to an element only involved the degrees of freedom of the considered element. The coupling between different elements appears only in the flux terms. The notations B_N and $\mathbb{E}_N$ hide a coupling between the values of the two elements (in the case of a conforming mesh) sharing the face. For example in the case of a centered flux, on the face f, $B_N = \frac{1}{2}(B_K + B_L)$ and $\mathbb{E}_N = \frac{1}{2}(\mathbb{E}_K + \mathbb{E}_L)$, where K and L denote the two elements sharing face f.*

We can thus write a matrix relation enabling to compute the degrees of freedom element by element, knowing that there exists a coupling through the terms B_N and $\mathbb{E}_N$. Plugging the above formulations in (6.66)-(6.68) we get

$$\begin{aligned}\frac{d}{dt}(\mathbb{F}_x^T M_K \mathbb{E}_x + \mathbb{F}_y^T M_K \mathbb{E}_y) - c^2(\mathbb{F}_y^T D_K^x - \mathbb{F}_x^T D_K^y)\mathbb{B}_z \\ - c^2 \sum_{f\in\partial K} (\tau_{f,x}\mathbb{F}_x^T + \tau_{y,f}\mathbb{F}_y^T) M_f \mathbb{B}_N \\ = -\frac{1}{\varepsilon_0}(\mathbb{F}_x^T M_K \mathbb{J}_x + \mathbb{F}_y^T M_K \mathbb{J}_y) \quad \forall (\mathbb{F}_x, \mathbb{F}_y) \in \mathbb{R}^{2N_k}, \quad (6.69)\end{aligned}$$

$$\frac{d}{dt}(\mathbb{C}_z^T M_K \mathbb{B}_z)+\mathbb{C}_z^T D_K^y \mathbb{E}_x-\mathbb{C}_z^T D_K^x \mathbb{E}_y-\sum_{f\in\partial K}\mathbb{C}_z^T M_f(\mathbb{E}_{Nx}\tau_{f,x}+\mathbb{E}_{Ny}\tau_{y,f}))=0$$
$$\forall \mathbb{C}_z \in \mathbb{R}^{N_k}, \quad (6.70)$$

which is equivalent to

$$\begin{pmatrix} M_K & 0 \\ 0 & M_K \end{pmatrix} \frac{d}{dt}\begin{pmatrix} \mathbb{E}_x \\ \mathbb{E}_y \end{pmatrix} - c^2 \begin{pmatrix} -D_K^y \\ D_K^x \end{pmatrix} \mathbb{B}_z - c^2 \sum_{f\in\partial K} \begin{pmatrix} \tau_{f,x} M_f \\ \tau_{f,y} M_f \end{pmatrix} \mathbb{B}_N$$
$$= -\frac{1}{\varepsilon_0}\begin{pmatrix} M_K & 0 \\ 0 & M_K \end{pmatrix}\begin{pmatrix} \mathbb{J}_x \\ \mathbb{J}_y \end{pmatrix} \quad (6.71)$$

$$M_K \frac{d}{dt}\mathbb{B}_z + \begin{pmatrix} D_K^y & -D_K^x \end{pmatrix}\begin{pmatrix} \mathbb{E}_x \\ \mathbb{E}_y \end{pmatrix} - \sum_{f\in\partial K}\begin{pmatrix} \tau_{f,x}M_f & \tau_{f,y}M_f \end{pmatrix}\begin{pmatrix} \mathbb{E}_{Nx} \\ \mathbb{E}_{Ny} \end{pmatrix} = 0. \quad (6.72)$$

In view of implementation, it is more convenient to rewrite this expression of the matrices on K using only integrals computed on a reference element $\hat{K}$ and change of variables. We consider here a mesh of triangles and a reference element $\hat{K}$ with vertices $(0,0),(1,0),(0,1)$, an an arbitrary element K with vertices $(x_1,y_1),(x_2,y_2),(x_3,y_3)$, the affine transformation of the plane

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \mapsto \mathcal{A}\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + b,$$

where

$$\mathcal{A} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \text{ and } b = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix},$$

transforms the reference triangle $\hat{K}$ onto the triangle K. Using the corresponding change of variables for computing the integrals on K used in the definition of the matrices, we get only integrals on $\hat{K}$. Thus

$$\int_K \varphi_i\,\varphi_j\,dx\,dy = \det(\mathcal{A})\int_{\hat{K}}\hat{\varphi}_i\,\hat{\varphi}_j\,d\hat{x}\,d\hat{y},$$

and so $M_K = \det(\mathcal{A})M_{\hat{K}}$. For the integrals involving a derivative, we use the chain rule which yields $\hat{\nabla}\hat{\varphi}_j = \mathcal{A}^T\nabla\varphi_j$ and so $\nabla\varphi_j = \mathcal{A}^{-T}\hat{\nabla}\hat{\varphi}_j$ where

$$\mathcal{A}^{-T} = (\mathcal{A}^{-1})^T = \frac{1}{\det\mathcal{A}}\begin{pmatrix} y_3 - y_1 & y_1 - y_2 \\ x_1 - x_3 & x_2 - x_1 \end{pmatrix}.$$

Thus as $\det\mathcal{A}$ of the jacobian cancels with the one from $\mathcal{A}^{-T}$, we have

$$\int_K \varphi_j\,\partial_x\varphi_i\,dxdy = (y_3 - y_1)\int_{\hat{K}}\hat{\varphi}_j\,\partial_{\hat{x}}\hat{\varphi}_i\,d\hat{x}d\hat{y} + (y_1 - y_2)\int_{\hat{K}}\hat{\varphi}_j\,\partial_{\hat{y}}\hat{\varphi}_i\,d\hat{x}d\hat{y},$$

and

$$\int_K \varphi_j \, \partial_y \varphi_i \, dxdy = (x_1 - x_3) \int_{\hat{K}} \hat{\varphi}_j \, \partial_{\hat{x}} \hat{\varphi}_i \, d\hat{x}d\hat{y} + (x_2 - x_1) \int_{\hat{K}} \hat{\varphi}_j \, \partial_{\hat{y}} \hat{\varphi}_i \, d\hat{x}d\hat{y}.$$

It follows that

$$D_K^x = (y_3 - y_1) D_{\hat{K}}^{\hat{x}} + (y_1 - y_2) D_{\hat{K}}^{\hat{y}} \text{ and } D_K^y = (x_1 - x_3) D_{\hat{K}}^{\hat{x}} + (x_2 - x_1) D_{\hat{K}}^{\hat{y}}.$$

Finally the integrals on the faces are computed using the same principle by a change of variables for the mass matrix on a face:

$$\int_f \varphi_i \, \varphi_j \, d\sigma = |f| \int_{\hat{f}} \hat{\varphi}_i \, \hat{\varphi}_j \, d\hat{\sigma}$$

where $|f|$ is the length of face f and $M_{\hat{f}}$ the mass matrix on the reference element of faces.

Plugging these expressions in the matrix formulation (6.71)-(6.72), we obtain a matrix formulation uniquely based on the elementary matrices on the reference element

$$\begin{aligned} \det(\mathcal{A}) \begin{pmatrix} M_{\hat{K}} & 0 \\ 0 & M_{\hat{K}} \end{pmatrix} \frac{d}{dt} \begin{pmatrix} \mathbb{E}_x \\ \mathbb{E}_y \end{pmatrix} - c^2 \begin{pmatrix} -(x_1 - x_3) D_{\hat{K}}^{\hat{x}} + (x_2 - x_1) D_{\hat{K}}^{\hat{y}} \\ (y_3 - y_1) D_{\hat{K}}^{\hat{x}} + (y_1 - y_2) D_{\hat{K}}^{\hat{y}} \end{pmatrix} \mathbb{B}_z \\ - c^2 \sum_{f \in \partial K} |f| \begin{pmatrix} \tau_{f,x} M_{\hat{f}} \\ \tau_{f,y} M_{\hat{f}} \end{pmatrix} \mathbb{B}_N = -\frac{1}{\varepsilon_0} \det(\mathcal{A}) \begin{pmatrix} M_{\hat{K}} & 0 \\ 0 & M_{\hat{K}} \end{pmatrix} \begin{pmatrix} \mathbb{J}_x \\ \mathbb{J}_y \end{pmatrix}, \quad (6.73) \end{aligned}$$

$$\begin{aligned} \det(\mathcal{A}) M_{\hat{K}} \frac{d}{dt} \mathbb{B}_z + \big((x_1 - x_3) D_{\hat{K}}^{\hat{x}} + (x_2 - x_1) D_{\hat{K}}^{\hat{y}} \big) \mathbb{E}_x \\ + \big(-(y_3 - y_1) D_{\hat{K}}^{\hat{x}} + (y_1 - y_2) D_{\hat{K}}^{\hat{y}} \big) \mathbb{E}_y \\ - \sum_{f \in \partial K} |f| \left(\tau_{f,x} M_{\hat{f}} \; \tau_{f,y} M_{\hat{f}} \right) \begin{pmatrix} \mathbb{E}_{\mathbb{N}x} \\ \mathbb{E}_{\mathbb{N}y} \end{pmatrix} = 0. \quad (6.74) \end{aligned}$$

6.6.3 Computation of the fluxes

In a Discontinuous Galerkin method, the inter-element coupling occurs through the numerical fliux which corresponds to the boundary terms in equations (6.66)-(6.68). The discrete fields in the Discontinuous Galerkin method being by definition discontinuous at the inter-element interfaces, these boundary terms could be expressed either using the values on one side of the boundary or on the other. In the case of a continuous solution, these values will obviously differ only by an error corresponding to the order of the method. A natural way to define these fluxes is to use the half sum of the terms on both sides. We shall call *centered flux* the corresponding flux. On the other hand in order to stabilize the method for long time computations it might be useful to use an upwind flux, which will introduce a dissipation mechanism.

This upwinding can be partial using a parameter $\alpha \in [\frac{1}{2}, 1]$ which enables to build a variable amount of upwinding into the scheme.This flux will then be the centered for $\alpha = \frac{1}{2}$ and the fully upwind flux for $\alpha = 1$.

In order to determin the upwind direction, it is necessary to find the direction of propagation of the flow in the vicinity of the interface. To this aim we need to separate the different waves propagating in the direction of the normal vector $\mathbf{n}$. To this aim we write the Maxwell equations (6.59)-(6.60) in the form of a hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + A_x \frac{\partial \mathbf{u}}{\partial x} + A_y \frac{\partial \mathbf{u}}{\partial y} = 0.$$

with

$$\mathbf{u} = \begin{pmatrix} E_x \\ E_y \\ B_z \end{pmatrix}, \quad A_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & c^2 \\ 0 & 1 & 0 \end{pmatrix}, \quad A_y = \begin{pmatrix} 0 & 0 & -c^2 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$

In a classical way for systems of conservation laws, we decompose the flux in the direction $\mathbf{n} = (n_x, n_y)^T$ in order to get the upwind flux using the signs of the eigenvalues of the matrix $A_n = n_x A_x + n_y A_y$. The eigenvalues of A_n are here 0, c and $-c$. Denote by

$$P = \begin{pmatrix} \frac{n_x}{n_y} & -n_y c & nyc \\ 1 & n_x c & -n_x c \\ 0 & 1 & 1 \end{pmatrix}$$

the matrix whose columns are the eigenvectors. This will be the transformation matrix into the diagonalization basis. Its inverse is

$$P^{-1} = \begin{pmatrix} n_y n_x & n_y^2 & 0 \\ -\frac{n_y}{2c} & \frac{n_x}{2c} & \frac{1}{2} \\ \frac{n_y}{2c} & -\frac{n_x}{2c} & \frac{1}{2} \end{pmatrix}.$$

Denote by

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & -c \end{pmatrix}$$

the associated diagonal eigenvalue matrix. We then write $\Lambda = \Lambda^+ + \Lambda^-$ the first matrix containing the positive eigenvalues and the second one the negative. Denoting by $\mathbf{u_L}$ and $\mathbf{u_R}$ the values of the unknowns on the left hand side and the right hand side of the interface, the upwind flux is defined by

$$A_n \mathbf{u}^* = A_n^+ \mathbf{u_L} + A_n^- \mathbf{u_R},$$

with $A_n^+ = P\Lambda^+ P^{-1}$ and $A_n^- = P\Lambda^- P^{-1}$.

The partially upwind flux is then obtained thanks to the formula

$$A_n \mathbf{u}_\alpha^* = A_n^+(\alpha \mathbf{u}_L + (1-\alpha)\mathbf{u}_R) + A_n^-(\alpha \mathbf{u}_R + (1-\alpha)\mathbf{u}_L).$$

Thus denoting for a quantity v, $\bar{v} = \frac{1}{2}(v_R + v_L)$ and $\Delta v = v_R - v_L$, the partially upwind flux for the Maxwell equations writes

$$A_n \mathbf{u}^*_\alpha = \begin{pmatrix} -cn_y(c\bar{B}_z + (\alpha - \frac{1}{2})(n_y \Delta E_x - n_x \Delta E_y) \\ cn_x(c\bar{B}_z + (\alpha - \frac{1}{2})(n_y \Delta E_x - n_x \Delta E_y) \\ n_x \bar{E}_y - n_y \bar{E}_x + (\alpha - \frac{1}{2}) c \Delta B_z \end{pmatrix}.$$

And as we already mentioned we recover the centered flux for $\alpha = \frac{1}{2}$ and the standard upwind flux for $\alpha = 1$.

6.6.4 Semi-discrete energy conservation

Let us consider the homogeneous equations (6.66)-(6.67) with $\gamma = 0$ and indexing the different elements and the discrete fields attached to the elements by the same index

$$\frac{d}{dt} \int_{K_i} \mathbf{E}_{h,i} \cdot \mathbf{F}_h - c^2 \int_{K_i} B_{h,z,i} \operatorname{curl} \mathbf{F}_h - c^2 \int_{\partial K_i} (\mathbf{F}_h \cdot \boldsymbol{\tau})\, B_N = 0, \quad \forall \mathbf{F}_h \in \mathbb{P}^2_k(K_i), \tag{6.75}$$

$$\frac{d}{dt} \int_{K_i} B_{h,z,i}\, C_{h,z} + \int_{K_i} \mathbf{E}_{h,i} \cdot \mathbf{curl}\, C_{h,z} - \int_{\partial K_i} (\mathbf{E}_N \cdot \boldsymbol{\tau})\, C_{h,z} = 0 \quad \forall C_{h,z} \in \mathbb{P}_k(K_i). \tag{6.76}$$

Proposition 13 *Assume periodic boundary conditions or* $\mathbf{E} \cdot \boldsymbol{\tau} = 0$ *on the boundary then the semi discrete energy*

$$\int_\Omega (\|\mathbf{E}_h\|^2 + c^2 B^2_{h,z})$$

is conserved in time.

Proof. Take $\mathbf{F}_h = \mathbf{E}_{h,i}$ and $C_{h,z} = B_{h,z,i}$ and sum the contributions over all the cells considering a centered flux, so that if K_i and K_j share an edge we have $B_N = \frac{1}{2}(B_{h,z,i} + B_{h,z,j}$ and $\mathbb{E}_N = \frac{1}{2}(\mathbf{E}_i + \mathbf{E}_j)$ Note also that, as the tangent vector τ is linked to the outbound normal we have $\tau_i = -\tau_j$ on the face share by K_i and K_j. We then get

$$\sum_i \frac{d}{dt} \Big(\int_{K_i} \|\mathbf{E}_{h,i}\|^2 + c^2 \int_{K_i} B^2_{h,z,i} \Big) - c^2 \int_{K_i} B_{h,z,i} \operatorname{curl} \mathbf{E}_h$$
$$+ c^2 \int_{K_i} \mathbf{E}_{h,i} \cdot \mathbf{curl}\, B_{h,z,i} - c^2 \int_{\partial K_i} (\mathbf{E}_{h,i} \cdot \boldsymbol{\tau}_i)\, B_N - c^2 \int_{\partial K_i} (\mathbf{E}_N \cdot \boldsymbol{\tau}_i)\, B_{h,z,i} = 0.$$

Then using the Green formula (6.64)

$$\int_{K_i} \mathbf{E}_{h,i} \cdot \mathbf{curl}\, B_{h,z,i} = \int_{K_i} B_{h,z,i} \operatorname{curl} \mathbf{E}_{h,i} + \int_{\partial K_i} (\mathbf{E}_{h,i} \cdot \boldsymbol{\tau}_i)\, B_{h,z,i},$$

so that, putting in the centered fluxes, with index j denoting an arbitrary element sharing an interface with K_i

$$\sum_i \frac{d}{dt}\Big(\int_{K_i} \|\mathbf{E}_{h,i}\|^2 + c^2 \int_{K_i} B_{h,z,i}^2\Big) - \frac{c^2}{2}\int_{\partial K_i} (\mathbf{E}_{h,i}\cdot\boldsymbol{\tau}_i)\, B_{k,z,j} - \frac{c^2}{2}\int_{\partial K_i} (\mathbf{E}_{h,j}\cdot\boldsymbol{\tau}_i)\, B_{h,z,i} = 0.$$

Finally the remaining boundary terms vanish as for a shared boundary between K_i and K_j they appear on both sides with an opposite sign as $\tau_i = -\tau_j$ and on the domain boundary they vanish because of the boundary conditions (periodic or $\mathbf{E}\cdot\boldsymbol{\tau} = 0$). We thus get the desired result as the global square of the L^2 norm on Ω is the sum of the square of the L^2 norms on the elements.

6.6.5 The time scheme

For the time discretization, we can use the classical leap-frog scheme which computes the electrical field at integer time steps $t_n = n\Delta t$ and the magnetic field at half integer time steps $t_{n+\frac{1}{2}} = (n+\frac{1}{2})\Delta t$. We notice that on each face $\tau_{f,x} = n_{f,y}$ eand $\tau_{f,y} = -n_{f,x}$. We then get from equations (6.71)-(6.72), the following time scheme:

$$M_K\mathbb{E}_x^{n+1} = M_K\mathbb{E}_x^n - \Delta t\big(c^2 D_K^y \mathbb{B}_z^{n+\frac{1}{2}} - \sum_{f\in\partial K} c n_{f,y} M_f (c\bar{\mathbb{B}}_z^{n+\frac{1}{2}} + (\alpha - \frac{1}{2})(n_{f,y}\Delta E_x^n - n_{f,x}\Delta E_y^n)) + \frac{1}{\epsilon_0} M_K \mathbb{J}_x\big) \quad (6.77)$$

$$M_K\mathbb{E}_y^{n+1} = M_K\mathbb{E}_y^n - \Delta t\big(-c^2 D_K^x \mathbb{B}_z^{n+\frac{1}{2}} + \sum_{f\in\partial K} c n_{f,x} M_f (c\bar{\mathbb{B}}_z^{n+\frac{1}{2}} + (\alpha - \frac{1}{2})(n_{f,y}\Delta\mathbb{E}_x^n - n_{f,x}\Delta\mathbb{E}_y^n)) + \frac{1}{\epsilon_0} M_K \mathbb{J}_y\big) \quad (6.78)$$

$$M_K\mathbb{B}_z^{n+\frac{3}{2}} = M_K\mathbb{B}_z^{n+\frac{1}{2}} - \Delta t\big(D_K^y\mathbb{E}_x^{n+1} - D_K^x\mathbb{E}_y^{n+1} + \sum_{f\in\partial K} M_f(n_{f,x}\mathbb{E}_y^{n+1} - n_{f,y}\mathbb{E}_x^{n+1} + c(\alpha - \frac{1}{2})\Delta B_z^{n+\frac{1}{2}})\big). \quad (6.79)$$

In order to keep a completely explicit time scheme, we have expressed the part of the flux involving the electric field in Ampère's law at time t_n and the part of the flux involving the magnetic field in Faraday's law at time $t_{n+\frac{1}{2}}$. This scheme is thus only first order when upwinding is performed. These parts cancel for the centered flux in which case the scheme is of order 2 in time.

7

Semi-Lagrangian approximation of the Vlasov equation

Semi-Lagrangian methods have become, far behind the Particle-In-Cell (PIC) method a classical choice for the numerical solution of the Vlasov equation, thanks to their good precision and their lack of numerical noise as opposite to PIC methods. They need a phase space mesh and thus are very computationally intensive when going to higher dimensions. Indeed a 3D simulation requires a 6D mesh of phase space. For this reason, semi-Lagrangian methods have become very popular for 1D or 2D problems, but there are still relatively few 3D simulations being performed with this kind of method.

The specificity of semi-Lagrangian methods, compared to classical methods for numerically solving PDEs on a mesh, is that they use the characteristics of the scalar hyperbolic equation, along with an interpolation method, to update the unknown from one time step to the next. These semi-Lagrangian methods exist in different varieties: backward, forward, point based or cell based.

Let us give a non exhaustive overview of semi-Lagrangian schemes for plasma physics applications. The semi-Lagrangian method based on a Strang splitting between space and velocity advection has been initially introduced for the 1D Vlasov-Poisson equation by Cheng and Knorr [33] in 1976. The splitting enables to get constant coefficient advection in each split problem, which could be solved exactly. For the interpolation, they used a trigonometric interpolation in x, which amounts to a spectral scheme, and cubic splines in v. Shoucri and Gagné [94] introduced shortly later a similar method based on a cubic spline interpolation in both directions, which is still typically used in many production semi-Lagrangian codes as it is very robust. However there are many alternatives as well for interpolation as for higher order time splitting as we saw in Chapter 5. In cite [96] the general semi-Lagrangian method that had been very successful for climate simulation [97] has been extended to general Vlasov-type equations, like the guiding-centre or gyrokinetic equations. Nakamura and Yabe introduced the semi-Lagrangian CIP method based on Hermite interpolation with advection of the derivatives. Filbet and co-authors introduced in 2001 the semi-Lagrangian PFC method, which is a finite vol-

ume time semi-Lagrangian method which conserves positivity and mass. In [15] the semi-Lagrangian method was extended to unstructured grids. Grandgirard *et al.* developed in [66] a backward semi-Lagrangian method for the drift-kinetic equations, which was later extended to the full gyrokinetic equations [65] and references therein. The forward semi-Lagrangian method was described in [44] and in [42] general conservative methods were introduced as well as new classes of filters and the equivalence of point based and conservative methods for Vlasov-Poisson with splitting was proved. Qiu and Christlieb [89] developed conservative WENO schemes. and Qiu and Shu [90] developed Discontinuous Galerkin semi-Lagrangian schemes.

Convergence properties of semi-Lagrangian schemes for the Vlasov equations have also been investigated in numerous papers. The convergence of the conservative and positive split PFC method has been proven in [60]. The convergence of the semi-Lagrangian method has been proven for linear interpolation in [12], for higher order spline and Lagrange interpolation in [13] and for Hermite interpolation with propagation of gradients in [14]. On the other hand convergence of an adaptive semi-Lagrangian method was proven in [26].

7.1 The classical semi-Lagrangian method

Let us consider an abstract scalar advection equation of the form

$$\frac{\partial f}{\partial t} + \mathbf{a}(\mathbf{x}, t) \cdot \nabla f = 0. \tag{7.1}$$

The characteristic curves associated to this equation are the solutions of the ordinary differential equations

$$\frac{d\mathbf{X}}{dt} = \mathbf{a}(\mathbf{X}(t), t).$$

We shall denote by $\mathbf{X}(t, \mathbf{x}, s)$ the unique solution of this equation associated to the initial condition $\mathbf{X}(s) = \mathbf{x}$.

The classical semi-Lagrangian method is based on a backtracking of characteristics. Two steps are needed to update the distribution function f^{n+1} at t_{n+1} from its value f^n at time t_n :

1. For each grid point $\mathbf{x}_i$ compute $\mathbf{X}(t_n; \mathbf{x}_i, t_{n+1})$ the value of the characteristic at t_n which takes the value $\mathbf{x}_i$ at t_{n+1}.
2. As the distribution solution of equation (7.1) verifies

$$f^{n+1}(\mathbf{x}_i) = f^n(X(t_n; \mathbf{x}_i, t_{n+1})),$$

we obtain the desired value of $f^{n+1}(x_i)$ by computing $f^n(\mathbf{X}(t_n; \mathbf{x}_i, t_{n+1})$ by interpolation as $\mathbf{X}(t_n; \mathbf{x}_i, t_{n+1})$ is in general not a grid point.

These operations are represented on Figure 7.1.

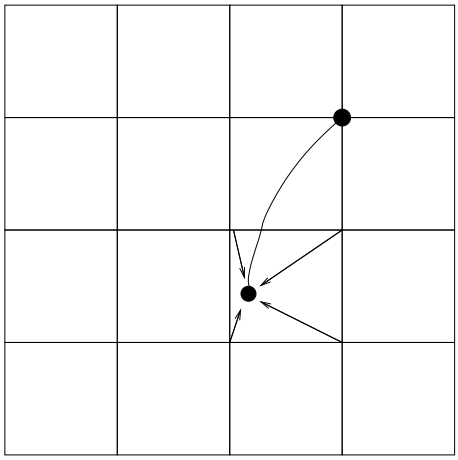

Fig. 7.1. Sketch of the classical semi-Lagrangian method.

Remark 21 *This semi-Lagrangian method is very diffusive if a low order (typically linear) interpolation is used. In practice one often used cubic splines or cubic Hermite interpolation, which offer a good compromise between accuracy and efficiency.*

Let us now specify the algorithm for the 1D Vlasov-Poisson problem where the unknown is the distribution function for the electrons and in presence of motionless neutralizing background ions on a domain $[0, L]$ periodic in x and infinite in v. The equations then read

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - E(x,t)\frac{\partial f}{\partial v} = 0,$$

$$\frac{dE}{dx} = \rho(x,t) = 1 - \int f(x,v,t)\,dv,$$

with the initial condition $f(x, v, 0) = f_0(x, v)$, verifying $\int f_0(x, v)\,dx\,dv = L$.

The infinite velocity space is truncated to a segment $[-A, A]$ sufficiently large so that f stays of the order of the round off errors for velocities less than $-A$ or larger than A during the whole simulation. In practise in the normalized examples we are going to consider, taking A of the order of 10 is very safe for all our test cases. Let us define a uniform grid of phase space $x_i = iL/N$, $i = 0, \ldots, N-1$ (the point x_N which corresponds to x_0 is not used), $v_j = -A + j2A/M$, $j = 0, \ldots, M$.

The full algorithm can in this case be written:

1. **Initialization.** Assume the initial distribution function $f_0(\mathbf{x}, \mathbf{v})$ given. We deduce $\rho(x, 0) = 1 - \int f_0(x, v)\,dv$, and then compute the initial electric field $E(x, 0)$ solving the Poisson equation.

2. **Update from t_n to t_{n+1}.** The function f^n is known at all grid points (x_i, v_j) of phase space and E^n is known at all grid points x_i of the configuration space.
   - We compute f^* by solving
   $$\frac{\partial f}{\partial t} + E^n \frac{\partial f}{\partial v} = 0$$
   on a half time step $\frac{\Delta t}{2}$ using the semi-Lagrangian method.
   - We compute f^{**} by solving on a full time step
   $$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = 0$$
   from the initial condition f^*.
   - We compute $\rho^{n+1}(x) = 1 - \int f^{**}(x, v)\, dv$ and then the corresponding electric field E^{n+1} using the Poisson equation.
   - We compute f^{n+1} by solving on a half time step
   $$\frac{\partial f}{\partial t} + E^{n+1} \frac{\partial f}{\partial v} = 0 \tag{7.2}$$
   from the initial condition f^{**}.

   Note that the actual ρ^{n+1} can be computed using $f^{**}(x, v)$ (instead of $f^{n+1}(x, v)$), as the charge density corresponding to $f^{**}(x, v)$ is identical to that associated to $f^{n+1}(x, v)$. Indeed, we go from $f^{**}(x, v)$ to $f^{n+1}(x, v)$ by solving (7.2), and we notice, integrating this equation in v that it implies that $\frac{d}{dt} \int f(x, v, t)\, dv = 0$ and so that ρ is not modified during this stage.

Advection of derivatives for cubic Hermite interpolation

A good alternative to cubic spline interpolation is cubic Hermite interpolation, which is still C^1 (compared to C^2 for cubic splines and C^0 for cubic Lagrange). This is less diffusive than cubic Lagrange, but needs to compute also the derivates at the grid points. This can be done by also advecting the derivative as proposed in [85] or in [15].

The v advection for a given electric field $E(x)$ reads

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

and its solution on one time step is $f(x, v, t + \Delta t) = f(x, v - E(x)\Delta t, t)$. From this expression we can deduce an explicit formula for the partial derivatives with respect to$x$ and v

$$\frac{\partial f}{\partial v}(x,v,t+\Delta t) = \frac{\partial f}{\partial x}(x,v-E(x)\Delta t,t), \tag{7.3}$$

$$\begin{aligned}\frac{\partial f}{\partial x}(x,v,t+\Delta t) &= \frac{\partial f}{\partial x}(x,v-E(x)\Delta t,t) - \frac{dE}{dx}(x)\Delta t\frac{\partial f}{\partial v}(x,v-E(x)\Delta t,t)\\ &= \frac{\partial f}{\partial x}(x,v-E(x)\Delta t,t) - \rho(x)\Delta t\frac{\partial f}{\partial v}(x,v-E(x)\Delta t,t).\end{aligned} \tag{7.4}$$

We then obtain, in addition to the value of f at the grid points, the values of the derivatives of f at the grid points that are needed for cubic Hermite interpolation.

In the same way the x advection can be written

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

and its explicit solution on one time step is $f(x,v,t+\Delta t) = f(x-v\Delta t,v,t)$, whence

$$\begin{aligned}\frac{\partial f}{\partial x}(x,v,t+\Delta t) &= \frac{\partial f}{\partial x}(x-v\Delta t,v,t),\\ \frac{\partial f}{\partial v}(x,v,t+\Delta t) &= -\Delta t\frac{\partial f}{\partial x}(x-v\Delta t,v,t) + \frac{\partial f}{\partial v}(x-v\Delta t,v,t).\end{aligned}$$

The 1D cubic Hermite interpolating of a function f reads

$$\pi f(x) = \sum_k f(x_k)\varphi(x_k) + f'(x_k)\psi(x_k),$$

where φ_k is the degree 3 polynomial on each cell that verifies $\varphi_k(x_l) = \delta_{k,l}$ et $\varphi_k'(x_l) = 0$ and ψ_k is the degree 3 polynomial on each cell that verifies $\psi_k(x_l) = 0$ and $\psi_k'(x_l) = \delta_{k,l}$ for all grid points x_l.

The scheme in each direction is then determined by using the values of f and its partial derivatives at the grid points. So for the c advection v we have

$$\begin{aligned}f_h(x,v,t+\Delta t) &= \pi(f(x,v-E(x)\Delta t,t))\\ &= \sum_k f(x,v_k,t)\varphi_k(v-E(x)\Delta t) + \frac{\partial f}{\partial v}(x,v_k,t)\psi_k(v-E(x)\Delta t).\end{aligned}$$

So we can compute

$$\frac{\partial f_h}{\partial v}(x,v,t+\Delta t) = \sum_k f(x,v_k,t)\varphi_k'(v-E(x)\Delta t) + \frac{\partial f}{\partial v}(x,v_k,t)\psi_k'(v-E(x)\Delta t),$$

and for the other derivative we use the expression (7.4) computing

$$\frac{\partial f}{\partial x}(x,v_E(x)\Delta t,t)$$

using a quadratic interpolation with respect to the neighbouring grid points.

7.2 Conservation properties of the semi-Lagrangian method with B-splines on a uniform mesh

Let us prove here the exact conservation properties of the classical semi-Lagrangian scheme on a uniform mesh that we just introduced. We shall check that during each split step of the method, which is a constant coefficient advection, total mass and momentum are exactly conserved in the case of periodic boundary conditions in x and infinite domain in v. These exact conservation properties will be violated by the truncation performed in the velocity domain. However if v_{min} and v_{max} are taken large enough this will be of the order of the round-off error and has no influence on the scheme.

7.2.1 Conservation of mass

Proposition 14 *The discrete mass $\Delta x \Delta v \sum_{i,j} f_{i,j}$ is exactly conserved by the numerical scheme.*

Proof. Here we just need to check that for a constant coefficient advection the mass is conserved on each line or each column. So let us consider only the 1D problem. Let us denote by f_i^k the value of the distribution function at the grid points at the beginning of a split step on one given line or column and f_i^{k+1} the value of the distribution function at the grid points at the end of the split step on then same line or column. Then if we prove that $\sum_i f_i^{k+1} = \sum_i f_i^k$, we can conclude that the total mass is conserved.

Starting from grid values f_i^k, we first compute the spline interpolant S on a grid of step h (with grid points of the form $x_i = ih$, $i \in \mathbb{Z}$). The spline S is defined by

$$S(x) = \sum_i c_i N^p(\frac{x}{h} - i), \text{ with } f_j^k = \sum_i c_i N^p(j-i).$$

Note that because $\int N^p(x/h - i)\,dx = h$, we have that $\int S(x)\,dx = h \sum c_i$ and because of the partition of unity property of the B-splines we also have

$$\sum_j f_j^k = \sum_{i,j} c_i N^p(j-i) = \sum_i c_i \sum_j N^p(j-i) = \sum_i c_i.$$

Now using the spline for interpolation at the origin of the characteristics with a constant advection coefficient a, we get

$$\begin{aligned}\sum_j f_j^{k+1} = \sum_j S(x_j - a\Delta t) = \sum_{i,j} c_i N^p(j - i - a\Delta t/h) \\ = \sum_i c_i \sum_j N^p(j - i - a\Delta t/h) = \sum_i c_i,\end{aligned}$$

using again the partition of unity property of the B-splines.

It follows that $\sum_j f_j^k = \sum_i c_i = \sum_j f_j^{k+1}$ from which we get the conservation of discrete mass.

7.2.2 Conservation of total momentum

Proposition 15 *The discrete total momentum*

$$\Delta x \Delta v \sum_{i,j} f_{i,j} v_j$$

is exactly conserved by the numerical scheme provided the Poisson solver verifies

$$\sum_i n_i E_i = 0, \text{ where } n_i = \Delta v \sum_j f_{i,j}.$$

Proof. In this case the advection in x and v need to be treated differently. The advection in x is applied on lines with constant velocities, so that the same computation as the one done for the conservation of mass can be applied. Indeed, for each j we get as previously on the split step $\sum_i f_{i,j}^{k+1} = \sum_i f_{i,j}^k$ and then multiplying by v_j and summing also on j we get the conservation of momentum for this step.

The advection in v is more complex. Performing the 1D advection for each i we have as in the previous paragraph

$$f_{i,j}^{k+1}(v_j) = f^k(v_j + E_i \Delta t) = \sum_l c_l N^p(j - l + E_i \Delta t / \Delta v).$$

So the new total momentum can be expressed as

$$\sum_j v_j f_{i,j}^{k+1}(v_j) = \sum_l c_l \sum_j j \Delta v N^p(j - l + E_i \Delta t / \Delta v).$$

But

$$\sum_j j N^p(j - k + E_i \Delta t / \Delta v) = \sum_j (j - l + E_i \Delta t / \Delta v) N^p(j - l + E_i \Delta t / \Delta v) + (l - E_i \Delta t / \Delta v) \sum_j N^p(j - l + E_i \Delta t / \Delta v),$$

with $\sum_j N^p(j - l + E_i \Delta t / \Delta v) = 1$ due to the partition of unity property of the splines and due to the properties of the cardinal splines proved in Lemma 2 we have that $\sum_j (j - l + E_i \Delta t / \Delta v) N^p(j - l + E_i \Delta t / \Delta v) = M_p$ the first moment of the cardinal spline of degree p. Then

$$\sum_j j f_{i,j}^{k+1}(v_j) = \sum_l c_l (M_p + l - E_i \Delta t / \Delta v).$$

On the other hand the total momentum on the column at the beginning of the time steps can also be expressed with the same spline coefficients simply using the same calculation with $E_i = 0$. Thus

$$\sum_j j f_{i,j}^k(v_j) = \sum_l c_l(M_p + l).$$

It follows that

$$\sum_j v_j f_{i,j}^{k+1}(v_j) = \sum_j v_j f_{i,j}^k(v_j) - \Delta t \sum_l c_l E_i.$$

From the proof of the previous proposition we recall that

$$\Delta v \sum_l c_l = \Delta v \sum_j f_{i,j}^k = n_i.$$

So that we have conservation of total momentum on this split step provided

$$\sum_i E_i n_i = 0$$

which concludes the prove the the proposition.

As E is linked to n via the Poisson solver this property needs to follow from the numerical scheme for the Poisson equation. Note that $\sum_i E_i n_i = 0$ is equivalent to $\sum_i E_i \rho_i = 0$ on a periodic domain provided $\sum_i E_i = 0$ which should be the case because E is a gradient. This is a discrete version of $\int E(x)\rho(x)\,dx$ which we have shown to vanish on a periodic domain.

Proposition 16 *The FFT spectral Poisson solver introduced in Chapter 6 satisfies* $\sum_i E_i = 0$ *and* $\sum_i E_i \rho_i = 0$.

Proof. First by definition of the discrete Fourier Transform $\sum_i E_i = \sqrt{N}\hat{E}_0$ which is set to 0 in the algorithm. Then using the discrete Parseval inequality, noticing that ρ_i and E_i are real while there Fourier transforms are not, we have

$$\sum_j \rho_j E_j = \sum_{k=-N/2}^{N/2-1} \hat{\rho}_k \bar{\hat{E}}_k = \sum_{k=-N/2}^{N/2-1} ik|\hat{E}_k|^2 = \sum_{k=1}^{N/2-1} ik(|\hat{E}_k|^2 - |\hat{E}_{-k}|^2),$$

as the algorithm yields $ik\hat{E}_k = \hat{\rho}_k$ and $\hat{E}_{-N/2} = 0$ by construction. On the other hand

$$\hat{E}_k = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} E_j e^{2i\pi jk/N}$$

from which it easily follows as the E_j are real that $\hat{E}_{-k} = \bar{\hat{E}}_k$ and thus they have the same modulus.

7.3 Numerical integration of the characteristics

Time split semi-Lagrangian methods for the Vlasov-Poisson equations have the great advantage of boiling down at each split step to constant coefficient advections, which enable an exact computation of the origin of the characteristics and thus greatly simplify the algorithm. However the splitting itself is a source of errors giving an even greater importance to the axes directions. In some cases it is interesting to develop a semi-Lagrangian method without splitting. On the other hand, there are case when splitting does not yield characteristics equations, which can be integrated analytically, like for example in the guiding center or gyrokinetic models.

In such cases the origins of the characteristics need to be computed numerically as the solutions for different initial conditions of an ordinary differential equation (ODE) of the form:

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{a}(t, \mathbf{X}),$$

where the advection field $\mathbf{a}$ typically containing at least the self-consistent electric field, depends non linearly on f and thus is not known at time t_{n+1}, when the origins of the characteristics are needed for computing $f(t_{n+1})$.

For the numerical computations, we need to distinguish between separable advection fields, like the Vlasov-Poisson model without splitting and the fully general case, the simplest example of which being the guiding-centre model.

7.3.1 Origin of the characteristics for the non split 1D Vlasov-Poisson equations

The non split characteristics of the electrostatic Vlasov equation read

$$\frac{dV}{dt} = E(X(t), t), \quad \frac{dX}{dt} = V.$$

Note that this ODE needs to be solved backward in time as we are backtracking the characteristics. This would have no influence if $E(x, t)$ was known for all times and we could just use a standard ODE solver backwards in time. However for the Vlasov-Poisson equation the electric field E depends on f, so in particular $E(x, t_{n+1})$ is not known when the origin of the characteristics ending at the grid points at time t_{n+1} need to be computed. So that a predictor-corrector type algorithm is needed.

At time t_n we know f^n and E^n at the grid points and we want to compute the same values at time t_{n+1}. The following order 2 predictor-corrector algorithm can be used to go from time step t_n to time step t_{n+1}:

1. Predict a value $\bar{E}^{n+1}$ for the electric field at time t_{n+1}.
2. For all grid points $x_i = X^{n+1}, v_j = V^{n+1}$ compute successively
   - $V^{n+1/2} = V^{n+1} - \frac{\Delta t}{2}\bar{E}^{n+1}(X^{n+1})$,

- $X^n = X^{n+1} - \Delta t V^{n+1/2}$,
- $V^n = V^{n+1/2} - \frac{\Delta t}{2}\bar{E}^n(X^n)$.
- Interpolate f^n at point (X^n, V^n).

3. We then have a first approximation of $f^{n+1}(x_i, v_j) = f^n(X^n, V^n)$ that can be used to compute a corrected version of E^{n+1} from which a new iteration can be performed if necessary to improve the precision.

In order to initialize the prediction of $\bar{E}^{n+1}$, it is convenient to use the continuity equation that is obtained by integrating the Vlasov equation with respect to the velocity variable:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = 0.$$

with an order 2 centred scheme, we then obtain an approximation of ρ^{n+1} from $\rho^{n-1} = \int f^{n-1}\, dv$ and of $J^n = \int v f^n\, dv$ in the form

$$\rho^{n+1} = \rho^{n-1} - 2\Delta t \nabla \cdot J^n.$$

We then compute $\bar{E}^{n+1}$ solving the Poisson equation with the source term ρ^{n+1}.

More generally for other Vlasov type equations like the guiding-center equation, the drift-kinetic or the gyrokinetic equations, it is possible to proceed in the same manner using the first velocity moments of the Vlasov equation to predict the electric field at time t_{n+1}.

Note that as the origins of the characteristics do not lie on a grid line now a full 2D interpolation in needed. This is typically performed using tensor product spline, Hermite or Lagrange interpolation. Especially in higher dimension, even though tensor product interpolations can also be performed the interpolation involves N^d points in d dimensions when it involves N in 1D. Hence the cost of tensor product interpolation becomes penalising in higher dimensions, which is an argument for using a directional splitting method.

7.3.2 The general case

Let us consider characteristics of the form

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{a}(t, \mathbf{X}).$$

In principle a predictor-corrector method can be avoided by using the following two time steps scheme that was suggested in [96]. This is based on a centered quadrature on two time steps:

$$\mathbf{X}^{n+1} - \mathbf{X}^{n-1} = 2\Delta t\, \mathbf{a}^n(\mathbf{X}^n), \qquad \mathbf{X}^{n+1} + \mathbf{X}^{n-1} = 2\mathbf{X}^n + O(\Delta t^2).$$

Then a fixed point procedure or Newton's method is used to compute $\mathbf{X}^{n-1}$ such that

$$\mathbf{X}^{n+1} - \mathbf{X}^{n-1} = \Delta t\, \mathbf{a}^n\left(\frac{\mathbf{X}^{n+1} + \mathbf{X}^{n-1}}{2}\right).$$

However this procedure appears to be unstable for long time computations because even and odd order time approximations become decoupled. A possible remedy is to recouple them by averaging from time to time.

Our preferred algorithm is to use the following one-step predictor-corrector algorithm. First a second order centered quadrature on one time step yields

$$\mathbf{X}^{n+1} - \mathbf{X}^n = \Delta t\, \mathbf{a}^{n+\frac{1}{2}}(\mathbf{X}^{n+\frac{1}{2}}), \qquad \mathbf{X}^{n+1} + \mathbf{X}^n = 2\mathbf{X}^{n+\frac{1}{2}} + O(\Delta t^2).$$

As $\mathbf{a}^{n+\frac{1}{2}}$ is unknown, it needs to be computed iteratively. A second order in time approximation is

$$\mathbf{a}^{n+\frac{1}{2}} = \frac{\mathbf{a}^n + \mathbf{a}^{n+1}}{2}.$$

As $\mathbf{a}^n$ a predictor corrector procedure is used for $\mathbf{a}^{n+1}$ like in the previous case up to convergence, which is generally obtained in one or two iterations.

Here we also need to use fixed point procedure or Newton iterations to compute $\mathbf{X}^n$ such that

$$\mathbf{X}^{n+1} - \mathbf{X}^n = \Delta t\, \mathbf{a}^{n+\frac{1}{2}} \left(\frac{\mathbf{X}^{n+1} + \mathbf{X}^n}{2} \right).$$

In practice we use linear interpolation for evaluation of $\mathbf{a}^{n+\frac{1}{2}}(X)$ to get an explicit solution for $\mathbf{X}^n$. This has proven sufficient in all of our applications.

7.3.3 Case of 1D characteristics with linear interpolation

A difficulty for the numerical method is that we are backtracking the characteristics and that in our Vlasov type problems the advection field a depends non linearly on f, at least through the electric field, which is generally not known, but can be predicted at time t_{n+1}. A robust and simple second order algorithm for computing the origin of the characteristics is the following. Assuming a is a known function of t and x. We get a second order approximation of the solution $X(t_n) = x_i^*$ at time t_n of $\frac{dX}{dt} = a(t, X)$ with $X(t_{n+1}) = x_i$ using the trapezoidal rule (a midpoint rule would also work and give the same order):

$$x_i - x_i^* = \frac{\Delta t}{2}[a(t_{n+1}, x_i) + a(t_n, x_i^*)]. \tag{7.5}$$

This is in general an implicit equation for x_i^*. However, in our applications a is known only at grid points, so an interpolation procedure is necessary to compute $a(t_n, x_i^*)$ as x_i^* is in general not a grid point. Moreover, in most cases that we have been investigating, no gain is obtained by using more than linear interpolation. And in this case, as described in [42] a completely explicit formula can be obtained: If we denote by x_{i_0} the for now unknown grid point immediately to the left of x_i^* and by $\beta_i = \frac{x_i^* - x_{i_0}}{\Delta x}$, we have $\beta_i \in [0, 1[$. Then we also have $x_i - x_i^* = (i - i_0 - \beta_i)\Delta x$, so that if we can determine i_0, and

β_i, we get x_i^*. Now if we inject this relation, and approximate $a(t_n, x_i^*)$ by a linear interpolation in the cell x_{i_0}, x_{i_0+1}, we get

$$(i - i_0 - \beta_i)\Delta x = \frac{\Delta t}{2}[a(t_{n+1}, x_i) + (1 - \beta_i)a(t_n, x_{i_0}) + \beta_i a(t_n, x_{i_0+1})].$$

Collecting the terms in factor of β_i we get

$$\beta_i\big(\Delta x + \frac{\Delta t}{2}[a(t_n, x_{i_0+1}) - a(t_n, x_{i_0})\big) = (i - i_0)\Delta x - \frac{\Delta t}{2}[a(t_{n+1}, x_i) + a(t_n, x_{i_0}). \tag{7.6}$$

From this we can extract the following formula for β_i:

$$\beta_i = \frac{i - i_0 - \frac{\Delta t}{2\Delta x}[a(t_{n+1}, x_i) + a(t_n, x_{i_0})]}{1 + \frac{\Delta t}{2\Delta x}[a(t_n, x_{i_0+1}) - a(t_n, x_{i_0})]}. \tag{7.7}$$

This formula is valid as long as the denominator does not vanish. This brings us to the question of stability of the semi-Lagrangian algorithm. It relies on the fact that the Lagrangian grid obtained by backtracking all the original grid points along the characteristics remains an acceptable grid. In 1D, this boils down to saying that the order of the grid points needs to be preserved and that those should not get to close to each other. We express this condition by $x_{i+1}^* - x_i^* > tol$, where tol is some small positive tolerance. Now assuming that this condition is verified. Then we can use the fact that $\beta \in [0, 1[$ in (7.6) to determine i_0. First, denoting by

$$M_{i_0} = i_0\Delta x + \frac{\Delta t}{2}[a(t_{n+1}, x_i) + a(t_n, x_{i_0})],$$

(7.6) becomes

$$\beta_i(M_{i_0+1} - M_{i_0}) = i\Delta x - M_{i_0}.$$

Hence $\beta_i \geq 0$ yields $M_{i_0} \leq i\Delta x$. Then $\beta_i < 1$ yields $M_{i_0+1} > i\Delta x$. This relates i_0 to the i it is linked to $(M_{i_0})_{i_0}$ sequence there is always a grid point.

Then i_0 being known formula (7.7) can be used to compute β_i to complete the determination of the origin of the characteristics.

The third and last step consists in computing the average value on the cell at time t_{n+1} by using the relation

$$\begin{aligned}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f^{n+1}(x)\,dx &= \int_{X(t_n; x_{i-\frac{1}{2}}, t_{n+1})}^{X(t_n; x_{i+\frac{1}{2}}, t_{n+1})} p^n(x)\,dx \\ &= \tilde{p}^n(X(t_n; x_{i+\frac{1}{2}}, t_{n+1})) - \tilde{p}^n(X(t_n; x_{i-\frac{1}{2}}, t_{n+1})),\end{aligned}$$

where $p^n(x)$ is the piecewise polynomial function reconstructed on each cell in step 1 and $\tilde{p}^n(x)$ its primitive.

7.4 Importance of conservativity

Consider an abstract Vlasov equation in the form

$$\frac{\partial f}{\partial t} + \mathbf{A}(\mathbf{z}, t) \cdot \nabla_z f = 0,$$

with $\nabla \cdot \mathbf{A} = 0$, where $\mathbf{z}$ represents here all the phase space variables. As we have seen, the property $\nabla \cdot \mathbf{A} = 0$ implies the conservativity of the equation. Consider now a splitting obtained by decompossing $\mathbf{z}$ into two groups of variables $\mathbf{z}_1$ and $\mathbf{z}_2$, we shall call the corresponding components of $\mathbf{A}$, $\mathbf{A}_1$ and $\mathbf{A}_2$. We then need to solve successively

$$\frac{\partial f}{\partial t} + \mathbf{A}_1(\mathbf{z}, t) \cdot \nabla_{x_1} f = 0,$$

and

$$\frac{\partial f}{\partial t} + \mathbf{A}_2(\mathbf{z}, t) \cdot \nabla_{x_2} f = 0.$$

We have $\nabla \cdot \mathbf{A} = \nabla_{z_1} \cdot \mathbf{A}_1 + \nabla_{z_2} \cdot \mathbf{A}_2 = 0$, but in general $\nabla_{z_1} \cdot \mathbf{A}_1$ and $\nabla_{z_2} \cdot \mathbf{A}_2$ are not both vanishing in which case none of the two split equations is conservative. It will then be very challenging to derive a conservative numerical method for the split system.

Examples.

1. The Vlasov-Poisson model. In this case $\mathbf{A} = (\mathbf{v}, \mathbf{E}(\mathbf{x}, t))$. Splitting in the classical manner between $\mathbf{x}$ and $\mathbf{v}$, we obtain $\mathbf{A}_1 = \mathbf{v}$ and $\mathbf{A}_1 = \mathbf{E}(\mathbf{x}, t)$. we have in this case $\nabla_x \cdot \mathbf{A}_1 = 0$ and $\nabla_v \cdot \mathbf{A}_2 = 0$, so that the splitting is conservative.
2. The guiding center model. This is a classical model in magnetized plasma physics which reads

   $$\frac{\partial \rho}{\partial t} + \mathbf{v}_D \cdot \nabla \rho = 0, \qquad -\Delta \phi = \rho,$$

   with

   $$\mathbf{v}_D = \frac{-\nabla \phi \times \mathbf{B}}{B^2} = \begin{pmatrix} -\frac{\partial \phi}{\partial y} \\ \frac{\partial \phi}{\partial x} \end{pmatrix} \text{ if } \mathbf{B} = \mathbf{e}_z \text{ unit vector in the } z\text{-direction.}$$

   We have indeed $\nabla \cdot \mathbf{v}_D = 0$, so that the guiding center model is conservative. However, splitting in the x and y directions, we obtain

   $$\frac{\partial \rho}{\partial t} - \frac{\partial \phi}{\partial y} \frac{\partial \rho}{\partial x} = 0,$$

   and

$$\frac{\partial \rho}{\partial t} + \frac{\partial \phi}{\partial x}\frac{\partial \rho}{\partial x} = 0.$$

The cross derivative $\frac{\partial^2 \phi}{\partial x \partial y}$ does in general not vanish and therefore the splitting is not conservative in this case. When numerical simulations are performed using non conservative split equations, large variations of total particle density (which should be conserved) can be observed, in particle in regions of the simulation where the distribution function is not well resolved. This phenomenon will happen in most problems modelled by the Vlasov equations, as there are filaments appearing and then vortex roll-ups. An illustration of this phenomenon is displayed in Figure 7.2 where the evolution of the L^1 and L^2 norms is compared for methods with and without splitting and also for a conservative method with splitting that shall be introduced in the next section. We observe that the methods without splitting and the conservative split method show a good physical behaviour, whereas for the split non-conservative method, the total number of particles varies by a very large amount which renders the results completely unphysical and this method unacceptable.

In addition to the guiding centre model which is a reduction of the so-called gyrokinetic which is used in strongly magnetised plasmas where the dimensional splitting performed on the advective form is non conservative, the same problem occurs for different versions of the relativistic Vlasov equation [71].

7.5 The conservative semi-Lagrangian method

It is also possible to derive semi-Lagrangian methods using the conservative form of the Vlasov equation. These will then naturally be conservative.

Let us point out that the classical semi-Lagrangian method applied to the Vlasov equation is also exactly conservative. This can be proven by showing that the resulting scheme is algebraically equivalent to a conservative scheme. See [44] for details.

The conservative semi-Lagrangian method has similarities with a Finite Volume method, but the computation of the fluxes is replaced by an integration over the volume occupied at the previous time step t_n by the cell under consideration. The unknown is the average value of f in one cell $\frac{1}{|V|}\int_V f\,dx\,dv$ and, as for finite volumes the numerical algorithm consists of three stages:

1. Reconstruction of a polynomial approximation of the desired degree from the cell averages.
2. Backtrack the cell down the flow of the characteristics (generally only the corner points are backtracked and the origin cell is approximated by a quadrilateral).
3. Compute the cell average of f at t_{n+1} using that $\int_V f\,dx\,dv$ is conserved along characteristics.

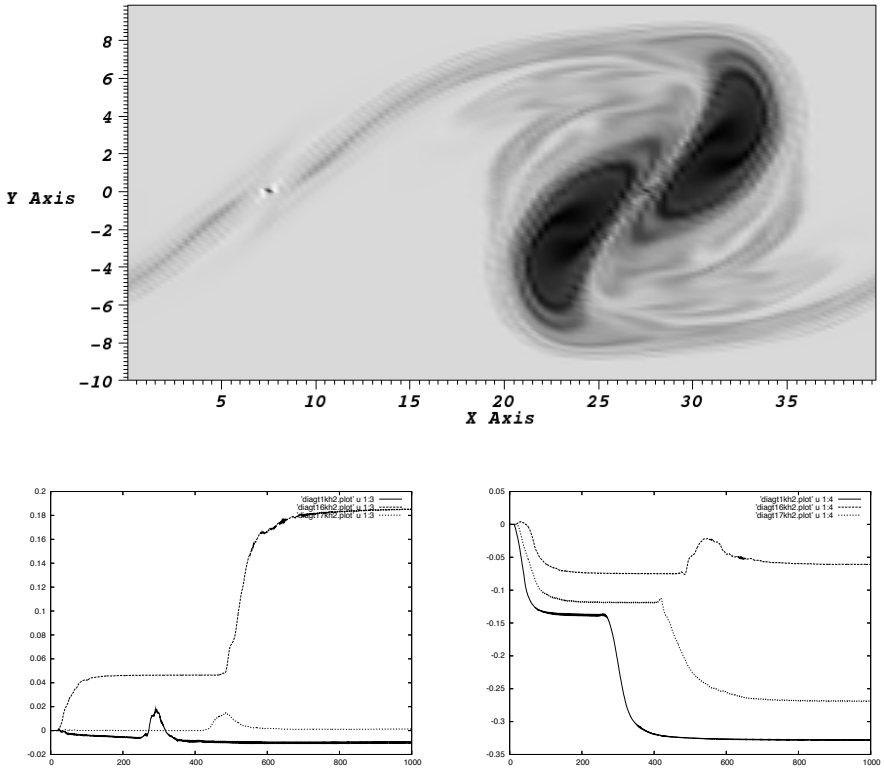

Fig. 7.2. Evolution of a Kelvin-Helmhotz instability for the guiding-center model. The top figure displays a snapshot of the distribution function during the creation of a vortex. The snapshot is taken a the time corresponding to the large increase of the L^1 norm on the bottom left figure. The bottom figures represent the evolution in time of the L^1 (left) and L^2 (right) norms for the non conservative splitting (top curve), conservative splitting (middle curve) and without splitting (bottom curve).

A scheme of principle is given in figure 7.3.

As in this case we work on the conservative form of the equation, the split equations are also in conservative form and thus the splitting does not generate conservativity issues, and as has been shown, the split conservative method performs well, unlike the split method based on the advective form, see Figure 7.2.

For this reason, and also because the conservative method becomes very simple in this case, as the 1D cell is completely determined by its endpoints, we shall only use it for split 1D equations in the conservative form

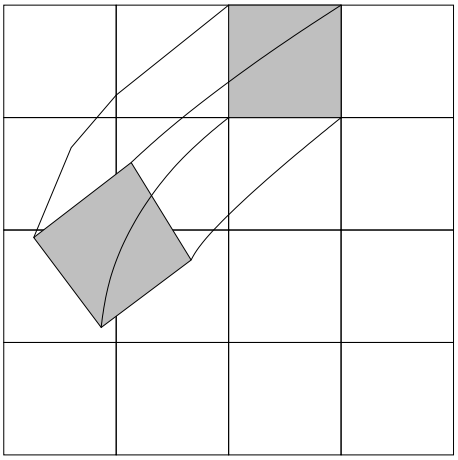

Fig. 7.3. Idea of conservative semi-Lagrangian method.

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x}(a(x,t)f) = 0.$$

Remark 22 *Another important reason for working on split 1D problems is that it avoids the so-called curse of dimensionality when going to higher dimensions. This is particularly important for the Vlasov equation for which realistic problems are posed in 4, 5 or 6 phase space dimensions. Indeed a typical cubic spline evaluation in 1D involves a stencil of 4 points which becomes* 4^d *in* d *dimensions which become large numbers in more that 4 dimensions.*

Let us now detail the 3 steps of the algorithm in the 1D case starting with step 1. This step consists in the reconstruction on each cell of a polynomial of degree m which has a given average value. The classical technique to do this consists in reconstructing the primitive of the polynomial we are looking for as follows:
Let f_j^n be the fixed average value of f^n in the cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ of length $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. We wish to construct a polynomial $p_m(x)$ of degree m such that

$$\frac{1}{h_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_m(x)\,dx = f_j^n.$$

In order to do this we shall define $\tilde{p}_m(x)$ verifying $\frac{d}{dx}\tilde{p}_m(x) = p_m(x)$ so that

$$h_j f_j^n = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_m(x)\,dx = \tilde{p}_m(x_{j+\frac{1}{2}}) - \tilde{p}_m(x_{j-\frac{1}{2}}).$$

Let $W(x) = \int_{x_{\frac{1}{2}}}^{x} \tilde{f}^n(x)\,dx$ be a primitive of the piecewise constant function $\tilde{f}^n$ which takes the value f_j^n on $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. We then have

$$W(x_{j+\frac{1}{2}}) = \sum_{k=1}^{j} h_k f_k^n$$

and

$$W(x_{j+\frac{1}{2}}) - W(x_{j-\frac{1}{2}}) = h_j f_j^n = \tilde{p}_m(x_{j+\frac{1}{2}}) - \tilde{p}_m(x_{j-\frac{1}{2}}).$$

If we take for $\tilde{p}_m$ an interpolation polynomial at points $x_{j+\frac{1}{2}}$ of the function W, we get that

$$\begin{aligned}\frac{1}{h_j}\int_{x_{\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_m(x)\,dx &= \frac{1}{h_j}(\tilde{p}_m(x_{j+\frac{1}{2}}) - \tilde{p}_m(x_{j-\frac{1}{2}})) \\ &= \frac{1}{h_j}(W(x_{j+\frac{1}{2}}) - W(x_{j-\frac{1}{2}})) \\ &= f_j^n,\end{aligned}$$

which is what we wanted.

The procedure we just describe works for any type of interpolation of the primitive. The PFC method uses cubic Lagrange interpolation [59]. Higher order Lagrange interpolation can also be used with benefit. It is also possible to chose a global interpolation of spline type which will enable to have more regularity on the reconstruction. Indeed for a Lagrange interpolation, the primitive will be continuous and the so the reconstructed polynomial p_m will be discontinuous at cell boundaries. On the other hand, if the primitive is for example a cubic spline, it will be on each cell a polynomial of degree 3 and be globally of C^2 regularity. The the reconstructed polynomial p_m will be a quadratic spline, which is a polynomial of degree 2 within each cell and of global regularity C^1. Cubic splines have been used in [42]. Note that in the case of constant coefficient advections it has been shown in [42] that the conservative and advective form with the same interpolation used directly for the function in the classical case and for the primitive in the conservative case are algebraically equivalent.

The second step of the method consists in computing the origin of the characteristics ending at the grid points. This step is the same as for the classical semi-Lagrangian algorithm, which has been described in a previous section of this chapter. For constant coefficient advections an exact solution can be computed. Note that for non constant (in the x variable) coefficient advection, the point based semi-Lagrangian algorithm using the advective form of the equation $\frac{\partial f}{\partial t} + a((t,x)\frac{\partial f}{\partial t} = 0$ is not conservative and not equivalent to the conservative form $\frac{\partial f}{\partial t} + \frac{\partial}{\partial t}(a(t,x)f) = 0$.

7.5.1 Stabilization.

Conservative semi-Lagrangian can benefit during the reconstruction step of a filtering procedure in the same way this is done in traditional finite volume

method, preserving the conservativity of the the method. Note however that most filters used for fluid dynamics problems are too strong and two dissipative for Vlasov type problems where no shocks occur. For many problems it is enough to use a filter just to ensure the positivity of the distribution function. A filter enabling to conserve positivity is described in [59, 42].

As the conservative semi-Lagrangian method is reminiscent of Finite Volume methods. It is natural to look at the reconstruction techniques used in this context in particular in order to handle non smooth solutions. There one finds the so called Essentially Non Oscillatory (ENO) reconstruction and its more efficient Weighted ENO (WENO) successor. See the review paper of Shu [95] for a detailed description to the WENO scheme with different applications and references.

The idea of ENO and WENO is to combine Lagrange interpolation on different stencils in order to reduce oscillations around numerical discontinuities. Let us explain how they can be use to reconstruct a polynomial of degree 2 in the interval $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. For the reconstruction of this polynomial three values are needed. Instead of using the intervals $S_2 = \{I_{i-1}, I_i, I_{i+1}\}$ for providing these values, one can also consider using the uncentered stencils $S_1 = \{I_{i-2}, I_{i-1}, I_i\}$ or $S_3 = \{I_i, I_{i+1}, I_{i+2}\}$. Let us denote by $p_j = \tilde{p}'_j$ the polynomial reconstructed using the stencil S_j.

In order to decide which of the polynomials should be used, the idea is to compute smoothness indicators which measure the magnitude of the derivatives in the cell $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. For this β_j is usually chosen to be

$$\beta_j = \sum_{l=1}^{m} h_j^{2l-1} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \big(\frac{d^l}{dx^l} p_j(x)\big)^2 \, dx. \tag{7.8}$$

This formula yields explicit formulas depending on the average values f_j^n being used for the reconstruction. In our case we get

$$\begin{aligned}
\beta_1 &= \frac{13}{12}(f_{i-2}^n - 2f_{i-1}^n + f_i^n)^2 + \frac{1}{4}(f_{i-2}^n - 4f_{i-1}^n + 3f_i^n)^2, \\
\beta_2 &= \frac{13}{12}(f_{i-1}^n - 2f_i^n + f_{i+1}^n)^2 + \frac{1}{4}(f_{i-1}^n - f_{i+1}^n)^2, \\
\beta_3 &= \frac{13}{12}(f_i^n - 2f_{i+1}^n + f_{i+2}^n)^2 + \frac{1}{4}(3f_i^n - 4f_{i+1}^n + f_{i+2}^n)^2.
\end{aligned}$$

The higher the value of β_j, the less smooth the polynomial p_j is on the cell I_i.

The ENO method then simply consists in using the reconstruction polynomial provided by the polynomial p_j for which β_j is the smallest. In practice ENO type stencil have not proved efficient for the Vlasov equation, as they increase the diffusivity [58].

Once three polynomials have been computed which all give an accuracy of Δx^4 if the underlying function is smooth, a natural idea is to use them all to get a better accuracy in regions where the function is smooth. As the

reconstruction procedure is linear with respect the the f_j^n a well chose linear combination of the three polynomials will give the interpolation polynomial using the 5 different know values on the three stencil which will be of order 6. This can be written as

$$p(x) = \gamma_1 p_1(x) + \gamma_2 p_2(x) + \gamma_3 p_3(x),$$

where $(\gamma_j)_j$ are computed so that $p(x)$ is the reconstructed polynomial using the 5 intervals stencil (or equivalently its primitive $\tilde{p}(x)$ is the degree 5 interpolation polynomial at the 6 interval extremities of the stencil.

Now in order to handle non smooth regions the weights γ_j are modified as follows using the smoothness indicators (7.8). The WENO reconstructed polynomial is defined by

$$p(x) = w_1 p_1(x) + w_2 p_2(x) + w_3 p_3(x),$$

where

$$w_j = \frac{\tilde{w}_j}{\tilde{w}_1 + \tilde{w}_2 + \tilde{w}_3}, \quad \text{with } \tilde{w}_j = \frac{\gamma_j}{(\varepsilon + \beta_j)^2}.$$

The number ε is a small positive real number add to avoid a division by 0. Typically it would be 10^{-6}.

Well designed WENO methods have had some success [30, 89, 90]. Note that, the Vlasov equation generates subcell oscillations but no shocks, the issue for designing good limiters are therefore different than in traditional conservation laws arising in fluid dynamics. Limiters are also needed for enforcing positivity, in the PFC algorithm [59] only such very weak limiting is performed.

7.5.2 Equivalence of conservative and classical semi-Lagrangian methods

In cases where the algorithm can be reduced by splitting to constant coefficient advections it has been proved in [42] that the conservative and classical semi-Lagrangian methods are equivalent.

7.6 The forward semi-Lagrangian method

As we saw previously in this chapter Backward semi-Lagrangian methods (BSL) need a predictor-corrector algorithm when the advection field is dependent on the distribution function and the problem cannot be simplified by operator splitting like for the Vlasov-Poisson problem. Moreover, the origins of the characteristics have to be computed iteratively, with Newton fixed point methods. This is due to an implicit way of solving the characteristics. This strategy makes high order resolution quite difficult and expensive. Making the

problem explicit enables to get rid of the two drawbacks and to use for example high order Runge Kutta methods more easily. This is one of the main advantages of the Forward semi-Lagrangian (FSL) method that was introduced for the Vlasov-Poisson equations in [44]. Once the new position of the particles computed, a remapping (or a deposition) step has to be performed. The idea is sketched in Figure 7.4. This issue is achieved using cubic spline polynomials which deposit the contribution of the Lagrangian particles on the uniform Eulerian mesh. This step is similar to the deposition step which occurs in PIC codes but in our case, the deposition is performed in all the phase space grid. The algorithm can be seen as a PIC method with remapping on a grid as was already introduced by Denavit in 1972 [51]. Similar algorithms have also been developed in [39, 84, 91] for meteorology applications and also in the book [41] for classical fluid dynamics problems. The method has also been extended to higher order by considering shaped particles that are linearly deformed by flow [27, 25]

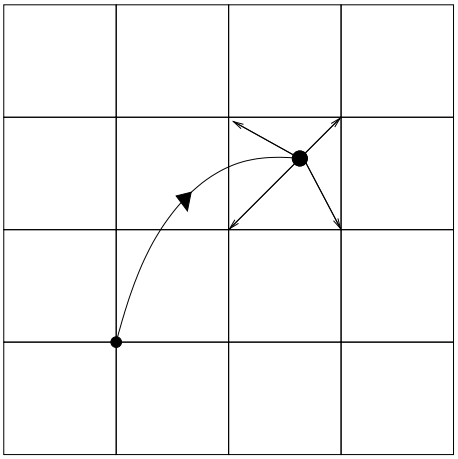

Fig. 7.4. Idea of conservative semi-Lagrangian method.

7.6.1 The general algorithm

Let us now introduce the different stages of the forward semi-Lagrangian method (FSL) in 2D, but generalisation to any dimension is straightforward, and try emphasize the differences with the traditional backward semi-Lagrangian method (BSL).

We consider here a generic Vlasov-type equation of the form

$$\frac{\partial f}{\partial t} + \mathbf{a}(t, \mathbf{x}) \nabla f = 0, \tag{7.9}$$

associated to the characteristic equation

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{A}(t, \mathbf{X}). \tag{7.10}$$

Let us consider a grid of the computational domain, possibly in phase-space with N_x and N_y cells in the x direction $[0, L_x]$ and in the y direction $[0, L_y]$. We then define

$$\Delta x = L_x/N_x, \;\; \Delta y = L_y/N_y, \;\; x_i = i\Delta x, \;\; y_j = j\Delta y,$$

for $i = 0, .., N_x$ and $j = 0, .., N_y$. One important point of the present method is the definition of the approximate distribution function, which is expressed for examples in a cubic B-splines basis. Other B-splines good be used and many other options can be taken from [41].

$$f(t, x, y) = \sum_{k,l} \omega_{k,l}^n S(x - X_1(t; x_k, y_l, t^n)) S(y - X_2(t; x_k, y_l, t^n)), \tag{7.11}$$

where $\mathbf{X}(t; x_k, y_l, t^n) = (X, Y)(t; x_k, y_l, t^n)$ corresponds to the solution of the characteristics at time t of the two dimensional system (7.10) whose value at time t^n was the grid point (x_k, y_l). The cubic B-spline S is defined as follows

$$6S(x) = \begin{cases} (2 - |x|)^3 & \text{if } 1 \le |x| \le 2, \\ 4 - 6x^2 + 3|x|^3 & \text{if } 0 \le |x| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

In the expression (7.11), the weight $w_{k,l}^n$ is associated to the particle located at the grid point (x_k, y_l) at time t^n; it corresponds to the coefficient of the cubic spline determined by the following interpolation conditions

$$\begin{aligned} f(t^n, x_i, y_j) &= \sum_{k,l} \omega_{k,l}^n S\left(x_i - X_1(t^n; x_k, y_l, t^n)\right) S\left(y_j - X_2(t^n; x_k, y_l, t^n)\right) \\ &= \sum_{k,l} \omega_{k,l}^n S(x_i - x_k) S(y_j - y_l). \end{aligned}$$

Adding boundary conditions (for example the value of the normal derivative of f at the boundaries, we obtain a set of linear systems in each direction from which the weights $\omega_{k,l}^n$ can be computed as in [96, 66]).

We can now express the full algorithm for the forward semi-Lagrangian method

- Initialisation:
  - Evaluate initial condition on grid $f_{i,j}^0 = f_0(x_i, y_j)$.
  - Compute the cubic splines coefficients $\omega_{k,l}^0$ such that

$$f_{i,j}^0 = \sum_{k,l} \omega_{k,l}^0 S(x_i - x_k) S(y_j - y_l).$$

- Time loop:
  - Integrate (7.10) from t^n to t^{n+1}, given as initial data the grid points $\mathbf{X}(t^n) = (x_k, y_l)$ to get $\mathbf{X}(t; x_k, y_l, t^n)$ for $t \in [t^n, t^{n+1}]$, assuming the advection field $\mathbf{A}$ is known.
  - Project on the phase space grid using (7.11) with $t = t^{n+1}$ to get $f_{i,j}^{n+1} = f^{n+1}(x_i, y_j)$
  - Compute the cubic spline coefficients $\omega_{k,l}^{n+1}$ such that

$$f_{i,j}^{n+1} = \sum_{k,l} \omega_{k,l}^{n+1} S(x_i - x_k) S(y_j - y_l).$$

7.6.2 An explicit computation of the characteristics

To compute the origin of the characteristics in the BSL method an iterative method is needed when the problem cannot be reduced to a case where there is an analytical solution, like the case of constant coefficient advections. The big advantage of the FSL algorithm is that this is no longer needed. As the time advance is in the forward direction the starting point of the characteristics is known so that classical ODE solvers, like Runge-Kutta methods can be used to achieve high order accuracy in time. Let us explicit the details of this explicit solution of the characteristics, in Vlasov-Poisson and Guiding-Center models. In our case we will describe the second-order Verlet algorithm, and Runge-Kutta of order 4 for Vlasov-Poisson, and, as Verlet cannot be applied, a Runge-Kutta method of order 2 for the Guiding-Center model. Higher order Runge-Kutta methods would work in the same way.

For the 1D Vlasov-Poisson system, we denote by $\mathbf{X}(t^n) = (x^n, v^n)$ the phase space grid, and the advection field is $\mathbf{A}(X(t^n), t^n) = (v^n, E(x^n, t^n))$ the advection velocity.

At time t_n the initial positions of the characteristics are the grid points and the electric field $E(x^n, t^n))$ is known. The Verlet algorithm to advance the characteristics to t_{n+1} for all grid points reads

- $v_{k,l}^{n+\frac{1}{2}} - v_l^n = \frac{\Delta t}{2} E(x_k^n, t^n)$,
- $x_{k,l}^{n+1} - x_k^n = \Delta t\, v_{k,l}^{n+1/2}$,
- Knowing the final characteristics positions at t_{n+1}, we can compute the electric field at time t^{n+1} as follows:
  - deposit the particles $x_{k,l}^{n+1}$ on the spatial grid x_i for the density ρ: $\rho(x_i, t^{n+1}) = \sum_{k,l} \omega_{k,l}^n S(x_i - x_{k,l}^{n+1})$, like in a PIC method.
  - Solve the Poisson equation on the grid which yields $E(x_i, t^{n+1})$ for all grid points.
- $v_{k,l}^{n+1} - v_{k,l}^{n+\frac{1}{2}} = \frac{\Delta t}{2} E(x_{k,l}^{n+1}, t^{n+1})$.

Let us remark that the particles (which start in our case from the grid points) move in the two-dimensional phase space; hence a double index (k, l) is necessary to denote the position and the velocity of the particles.

Another option for Vlasov-Poisson are Runge-Kutta algorithms. The fourth order Runge-Kutta algorithm needs to compute intermediate values in time of the density and the electric field. Let us detail the algorithm omitting the indices k, l for the sake of simplicity

- $k_1 = (v^n, E(x^n, t^n)) = (k_1(1), k_1(2))$,
- Compute the electric field at intermediate time t_1:
  - Compute the charge density ρ on the grid using $\rho(x_i, t_1) = \sum_{k,l} \omega_{k,l}^n S[x_i - (x_k^n + \Delta t/2\ k_1(1))]$.
  - Solve the Poisson equation on the grid to get $E(x_i, t_1)$ at each grid point
- Compute $k_2 = (v^n + \frac{\Delta t}{2} k_1(2), E(x^n + \frac{\Delta t}{2} k_1(1), t_1)$
- Compute the electric field at intermediate time t_2 with the same procedure as at t_1.
- Compute $k_3 = (v^n + \frac{\Delta t}{2} k_2(2), E(x^n + \frac{\Delta t}{2} k_2(1), t_2)$
- Compute the electric field at intermediate time t_3 with the same procedure as at t_1.
- Compute $k_4 = (v^n + \Delta t\ k_3(2), E(x^n + \Delta t\ k_3(1), t_3)$
- Accumulate previously computed values to get

$$X^{n+1} = X^n + \frac{\Delta t}{6} \left[k_1 + 2k_2 + 2k_3 + k_4\right].$$

In both Verlet and Runge-Kutta algorithms, the value of the electric field E at intermediate time steps is needed. This is computed like in PIC algorithms by advancing the particles (which coincide at time t^n with the mesh in this method) up to the required intermediate time. Using a deposition step, the density is computed thanks to cubic splines of coefficients w_i^n on the mesh at the right time, and thus the electric field can also be computed at the same time thanks to the Poisson equation. Using an interpolation operator, the electric field is then evaluated at the required location. Let us remark that this step involves a high order interpolation operator (cubic spline for example) which has been proved to be more accurate than a linear interpolation [44].

For the Guiding-Center model, the advection field is not separable and so the Verlet scheme can not be used. We thus restrict ourselves to Runge-Kutta type methods. Moreover, compared to the Vlasov-Poisson model in the deposition step, which is needed to compute the electric field needed in the different stages of the Runge-Kutta method is now two dimensional instead of one dimensional.

As an illustration, let us explicit the second order Runge-Kutta method applied to the guiding center model. At time t_n the characteristics $\mathbf{X}^n = (x^n, y^n)$ are at the grid points and the advection field $\mathbf{A}(X^n, t^n) = \mathbf{E}^\perp(\mathbf{X}^n, t^n)$ is known. In order to compute X^{n+1} the following steps are implemented:

- $\tilde{\mathbf{X}}^{n+1} - \mathbf{X}^n = \Delta t \mathbf{E}^\perp(\mathbf{X}^n, t^n)$
- Compute the electric field at time t^{n+1}

  - Compute the two dimensional charge density ρ using

$$\rho(x_j, y_i, t^{n+1}) = \sum_k \omega_k^n S[x_j - x_{k,l}^{n+1}] S[y_i - y_{k,l}^{n+1}].$$

  - Solve the two-dimensional Poisson equation on the grid yielding $E(x_j, y_i, t_{n+1})$ at all grid points.
- Accumulate stages to get

$$\mathbf{X}^{n+1} = \mathbf{X}^n + \frac{\Delta t}{2} \left[\mathbf{E}^\perp(\mathbf{X}^n, t^n) + \mathbf{E}^\perp(\tilde{\mathbf{X}}^{n+1}, t^{n+1}) \right].$$

8

Particle approximation of the Vlasov equation

8.1 Introduction

The method which is still by far the most used method for the simulation of the Vlasov-Maxwell equations is the Particle In Cell (PIC) method which consists in the coupling of a particle method for the Vlasov equation and a mesh based method for the computation of the self-consistent field using Maxwell's equations or some reduced model. The principle of the method is to discretize the distribution function by a collection of macro-particles representing the initial distribution function $f_0(\mathbf{x}, \mathbf{v})$ which, when normalized such that its integral is 1, represents a probability density. The macro-particles are then advanced in time by solving the equations of motion of the particles in the global electromagnetic field. Coupling the field solver with the particles is done by computing the sources of Maxwell's equations ρ and $\mathbf{J}$ from the particles using some regularization method. Any classical solver for Maxwell's equations can then be used on the mesh. In order to continue the time loop the fields need then to be computed at the particle positions, which can be done in a natural way using some solvers (Finite Elements for example), where the discrete fields are defined at any place. In order case some interpolation procedure needs to be defined. A huge literature on these methods exists, including two books that are rather physics oriented, by Birdsall and Langdon [17] and Hockney and Eastwood [70]. Mathematical convergence proofs of the algorithms have also been performed in some special cases, see Neunzert and Wick [88], Cottet and Raviart [40], Victory and Allen [103] and Wollman [105].

There also exists a variant of the PIC method which is often used when the physics that is being investigated remains close to some equilibrium configuration, examples are PIC simulations of tokamak plasmas or of particle accelerators. This method is called δf. It consists in expanding the distribution function in the neighborhood of a known equilibrium f^0 in $f = f^0 + \delta f$ and to approximate only the δf part with a PIC method. Another particle method, linked to SPH (smooth particle hydrodynamics) used in fluid dynam-

ics has been introduced by Bateson and Hewett [9], but seems not to have been used very much since. It consists in pushing a relatively small number of macro-particles in the form of a Gaussian whose size can vary and that interact directly with each other.

8.2 The PIC method

The principle of a particle method is to approximate the distribution function f solution of the Vlasov equation by a sum of Dirac masses centered at the particle positions in phase space $(\mathbf{x}_k(t), \mathbf{v}_k(t))_{1\leq k\leq N}$ of a number N of macro-particles each having a weight w_k. The approximated distribution function that we denote by f_N then writes

$$f_N(\mathbf{x}, \mathbf{v}, t) = \sum_{k=1}^{N} w_k \delta(\mathbf{x} - \mathbf{x}_k(t))\ \delta(\mathbf{v} - \mathbf{v}_k(t)).$$

Positions $\mathbf{x}_k^0$, velocities $\mathbf{v}_k^0$ and weights w_k are initialised such that $f_N(\mathbf{x}, \mathbf{v}, 0)$ is an approximation, in some sense that remains to be precised, of the initial distribution function $f_0(\mathbf{x}, \mathbf{v})$. The time evolution of the approximation is done by advancing the macro-particles along the characteristics of the Vlasov equation, *i.e.* by solving the system of differential equations

$$\begin{aligned} \frac{d\mathbf{x}_k}{dt} &= \mathbf{v}_k \\ \frac{d\mathbf{v}_k}{dt} &= \frac{q}{m}(\mathbf{E}(\mathbf{x}_k, t) + \mathbf{v}_k \times \mathbf{B}(\mathbf{x}_k, t)) \\ \mathbf{x}_k(0) = \mathbf{x}_k^0, &\quad \mathbf{v}_k(0) = \mathbf{v}_k^0. \end{aligned}$$

Proposition 17 *The function f_N is a solution in the sense of distributions of the Vlasov equation associated to the initial condition $f_N^0(\mathbf{x}, \mathbf{v}) = \sum_{k=1}^{N} w_k \delta(\mathbf{x} - \mathbf{x}_k^0)\ \delta(\mathbf{v} - \mathbf{v}_k^0)$.*

Proof. Let $\varphi \in C_c^\infty(\mathbb{R}^3 \times \mathbb{R}^3 \times]0, +\infty[)$. Then f_N defines a distribution of $\mathbb{R}^3 \times \mathbb{R}^3 \times]0, +\infty[$ in the following way:

$$\langle f_N, \varphi \rangle = \sum_{k=1}^{N} \int_0^T w_k \varphi(\mathbf{x}_k(t), \mathbf{v}_k(t), t)\, dt.$$

We then have

$$\langle \frac{\partial f_N}{\partial t}, \varphi \rangle = -\langle f_N, \frac{\partial \varphi}{\partial t} \rangle = -\sum_{k=1}^{N} w_k \int_0^T \frac{\partial \varphi}{\partial t}(\mathbf{x}_k(t), \mathbf{v}_k(t), t)\, dt,$$

but

$$\frac{d}{dt}(\varphi(\mathbf{x}_k(t),\mathbf{v}_k(t),t)) = \frac{d\mathbf{x}_k}{dt}\cdot\nabla_x\varphi + \frac{d\mathbf{v}_k}{dt}\cdot\nabla_v\varphi + \frac{\partial\varphi}{\partial t}(\mathbf{x}_k(t),\mathbf{v}_k(t),t),$$

and as φ has compact support in $\mathbb{R}^3\times\mathbb{R}^3\times]0,+\infty[$, it vanishes for $t=0$ and $t=T$. So

$$\int_0^T \frac{d}{dt}(\varphi(\mathbf{x}_k(t),\mathbf{v}_k(t),t))\,dt = 0.$$

It follows that

$$\begin{aligned}\langle\frac{\partial f_N}{\partial t},\varphi\rangle &= \sum_{k=1}^N w_k\int_0^T(\mathbf{v}_k\cdot\nabla_x\varphi + \frac{q}{m}(\mathbf{E}(\mathbf{x}_k,t)+\mathbf{v}_k\times\mathbf{B}(\mathbf{x}_k,t))\cdot\nabla_v\varphi)\,dt\\ &= -\langle\mathbf{v}\cdot\nabla_x f_N + \frac{q}{m}(\mathbf{E}(\mathbf{x}_k,t)+\mathbf{v}\times\mathbf{B}(\mathbf{x}_k,t))\cdot\nabla_v f_N,\varphi\rangle.\end{aligned}$$

Which means that f_N verifies exactly the Vlasov equation in the sense of distributions.

8.2.1 Consequence

If it is possible to solve exactly the equations of motion, which is sometimes the case for a sufficiently simple applied field, the particle method gives the exact solution for an initial distribution function which is a sum of Dirac masses.

The self-consistent electromagnetic field is computed on a mesh of physical space using a classical method (e.g. Finite Elements, Finite Differences, ...) to solve the Maxwell or the Poisson equations.

In order to determine completely a particle method, it is necessary to precise how the initial condition f_N^0 is chosen and what is numerical method chosen for the solution of the characteristics equations and also to define the particle-mesh interaction.

Let us detail the main steps of the PIC algorithm:

8.2.2 Choice of the initial condition.

- *Deterministic method:* Define a phase space mesh (uniform or not) and pick as the initial position of the particles $(\mathbf{x}_k^0,\mathbf{v}_k^0)$ the barycentres of the cells and for weights w_k associated to the integral of f_0 on the corresponding cell: $w_k = \int_{V_k} f_0(\mathbf{x},\mathbf{v})\,d\mathbf{x}d\mathbf{v}$ so that $\sum_k w_k = \int f_0(\mathbf{x},\mathbf{v})\,d\mathbf{x}d\mathbf{v}$.
- *Monte-Carlo method:* Pick the initial positions in a random or pseudo-random way using the probability density associated to f_0.

Remark 23 *Note that randomization occurs through the non-linear processes, which are generally such that holes appear in the phase space distribution of particles when they are started from a grid. Moreover the alignment of the particles on a uniform grid can also trigger some small physical, e.g. two stream, instabilities. For this reason a pseudo-random initialization is usually the best choice and is mostly used in practice.*

8.2.3 Particle-Mesh coupling.

The particle approximation f_N of the distribution function does not naturally give an expression for this function at all points of phase space. Thus for the coupling with the field solver which is defined on the mesh a regularizing step is necessary. To this aim we need to define convolution kernels which can be used the this regularization procedure. On cartesian meshes B-splines are mostly used as this convolution kernel. B-splines can be defined recursively: The degree 0 B-spline that we shall denote by S^0 is defined by

$$S^0(x) = \begin{cases} \frac{1}{\Delta x} & \text{if } -\frac{\Delta x}{2} \leq x < \frac{\Delta x}{2}, \\ 0 & \text{else.} \end{cases}$$

Higher order B-splines are then defined by:
For all $m \in \mathbb{N}^*$,

$$\begin{aligned} S^m(x) &= (S^0)^{*m}(x), \\ &= S^0 * S^{m-1}(x), \\ &= \frac{1}{\Delta x} \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} S^{m-1}(u)\, du. \end{aligned}$$

In particular the degree 1 spline is

$$S^1(x) = \begin{cases} \frac{1}{\Delta x}(1 - \frac{|x|}{\Delta x}) & \text{si } |x| < \Delta x, \\ 0 & \text{sinon,} \end{cases}$$

the degree 2 spline is

$$S^2(x) = \frac{1}{\Delta x} \begin{cases} \frac{1}{2}(\frac{3}{2} - \frac{|x|}{\Delta x})^2 & \text{si } \frac{1}{2}\Delta x < |x| < \frac{3}{2}\Delta x, \\ \frac{3}{4} - (\frac{x}{\Delta x})^2 & \text{si } |x| < \frac{1}{2}\Delta x, \\ 0 & \text{sinon,} \end{cases}$$

the degree 3 spline is

$$S^3(x) = \frac{1}{6\Delta x} \begin{cases} (2 - \frac{|x|}{\Delta x})^3 & \text{si } \Delta x \leq |x| < 2\Delta x, \\ 4 - 6\left(\frac{x}{\Delta x}\right)^2 + 3\left(\frac{|x|}{\Delta x}\right)^3 & \text{si } 0 \leq |x| < \Delta x, \\ 0 & \text{sinon.} \end{cases}$$

B-splines verify the following important properties

Proposition 18 • *Unit mean*

$$\int S^m(x)\, dx = 1.$$

- *Partition of unit. For $x_j = j\Delta x$,*

$$\Delta x \sum_j S^m(x - x_j) = 1.$$

- *Parity*
$$S^m(-x) = S^m(x).$$

The sources for Maxwell's equations ρ and $\mathbf{J}$ are defined from the numerical distribution function f_N, for a particle species of charge q by

$$\rho_N = q\sum_k w_k \delta(\mathbf{x}-\mathbf{x}_k), \quad \mathbf{J}_N = q\sum_k w_k \delta(\mathbf{x}-\mathbf{x}_k)\mathbf{v}_k.$$

We then apply the convolution kernel S to defined ρ and $\mathbf{J}$ at any point of space and in particular at the grid points:

$$\rho_h(\mathbf{x},t) = \int S(\mathbf{x}-\mathbf{x}')\rho_N(\mathbf{x}')\,d\mathbf{x}' = q\sum_k w_k S(\mathbf{x}-\mathbf{x}_k), \tag{8.1}$$

$$\mathbf{J}_h(\mathbf{x},t) = \int S(\mathbf{x}-\mathbf{x}')\mathbf{J}_N(\mathbf{x}')\,d\mathbf{x}' = q\sum_k w_k S(\mathbf{x}-\mathbf{x}_k)\mathbf{v}_k. \tag{8.2}$$

In order to get conservation of total momentum, when a regularization kernel is applied to the particles, the same kernel needs to be applied to the field seen as Dirac masses at the grid points in order to compute the field at the particle positions. We then obtain

$$\mathbf{E}_h(\mathbf{x},t) = \sum_j \mathbf{E}_j(t)S(\mathbf{x}-\mathbf{x}_j), \quad \mathbf{B}_h(\mathbf{x},t) = \sum_j \mathbf{B}_j(t)S(\mathbf{x}-\mathbf{x}_j). \tag{8.3}$$

Note that in the classical case where $S = S^1$ this regularization is equivalent to a linear interpolation of the fields defined at the grid points to the positions of the particles, but for higher order splines this is not an interpolation anymore and the regularized field at the grid points is not equal to its original value $\mathbf{E}_j$ anymore, but for example in the case of S^3, to $\frac{1}{6}\mathbf{E}_{j-1} + \frac{2}{3}\mathbf{E}_j + \frac{1}{6}\mathbf{E}_{j+1}$.

8.2.4 Conservation properties at the semi-discrete level.

- Conservation of mass. The discrete mass is defined as $\int f_N(\mathbf{x},\mathbf{v},t)\,d\mathbf{x}d\mathbf{v} = \sum_k w_k$. This is obviously conserved if no particle gets in or out of the domain, as w_k is conserved for each particle when the particles move.
- Conservation of momentum. The total momentum of the system is defined as
$$\mathcal{P} = m\int \mathbf{v} f_N(\mathbf{x},\mathbf{v},t)\,d\mathbf{x}d\mathbf{v} = \sum_k m_k w_k \mathbf{v}_k(t).$$
So
$$\frac{d\mathcal{P}}{dt} = \sum_k m_k w_k \frac{d\mathbf{v}_k}{dt} = \sum_k w_k q_k \mathbf{E}_h(\mathbf{x}_k,t).$$
In the case $\mathbf{E}_h$ is computed using a Finite Difference approximation, its value at the particle position should be computed using the same convolution kernel as is used for computing the charge and current densities from the particle positions. Then $\mathbf{E}_h(\mathbf{x}_k,t) = \sum_j \mathbf{E}_j(t)S(\mathbf{x}_k-\mathbf{x}_j)$ and so

$$\frac{d\mathcal{P}}{dt} = \sum_k w_k q_k \sum_j \mathbf{E}_j(t) S(\mathbf{x}_k - \mathbf{x}_j).$$

Then exchanging the sum on the grid points i and the sum on the particles k we get

$$\frac{d\mathcal{P}}{dt} = \sum_j \mathbf{E}_j(t) \sum_k w_k q_k S(\mathbf{x}_k - \mathbf{x}_j) = \sum_j \mathbf{E}_j(t)\rho_j(t),$$

so that the total momentum is conserved provided the field solver is such that $\sum_j \mathbf{E}_j(t)\rho_j(t)$.

In the case of a Finite Element PIC solver the Finite Element interpolant naturally provides and expression of the fields everywhere in the computational domain and the weak form of the right-hand side provides a natural definition of the source term for the finite element formulation. Let us also check the conservation of momentum in this case. Denoting by φ_i the Finite Element basis functions, we have $\mathbf{E}_h(\mathbf{x}_k, t) = \sum_j \mathbf{E}_j(t)\varphi_j(\mathbf{x}_k)$ and so

$$\frac{d\mathcal{P}}{dt} = \sum_k w_k q_k \sum_j \mathbf{E}_j(t)\varphi_j(\mathbf{x}_k) = \sum_j \mathbf{E}_j(t)\rho_j(t)$$

where $\rho_j = \int q f_N(\mathbf{x}_k, \mathbf{v}_k, t)\varphi_j(x)\, d\mathbf{x}d\mathbf{v}$.

Remark 24 *Note the conservation of momentum is linked to the self-force problem that is often mentioned in the PIC literature. Indeed if the system is reduced to one particle. The conservation of momentum is equivalent to the fact that a particle does not apply a force on itself.*

8.2.5 Time scheme for the particles.

Let us consider first only the case when the magnetic field vanishes (Vlasov-Poisson). Then the macro-particles obey the following equations of motion:

$$\frac{d\mathbf{x}_k}{dt} = \mathbf{v}_k, \quad \frac{d\mathbf{v}_k}{dt} = \frac{q}{m}\mathbf{E}(\mathbf{x}_k, t).$$

This system being hamiltonian, it should be solved using a symplectic time scheme in order to enjoy long time conservation properties. The scheme which is used most of the time is the Verlet scheme, which is defined as follows. We assume $\mathbf{x}_k^n$, $\mathbf{v}_k^n$ and $\mathbf{E}_k^n$ known.

$$\mathbf{v}_k^{n+\frac{1}{2}} = \mathbf{v}_k^n + \frac{q\Delta t}{2m}\mathbf{E}_k^n(\mathbf{x}_k^n), \tag{8.4}$$

$$\mathbf{x}_k^{n+1} = \mathbf{x}_k^n + \Delta t \mathbf{v}_k^{n+\frac{1}{2}}, \tag{8.5}$$

$$\mathbf{v}_k^{n+1} = \mathbf{v}_k^{n+\frac{1}{2}} + \frac{q\Delta t}{2m}\mathbf{E}_k^{n+1}(\mathbf{x}_k^{n+1}). \tag{8.6}$$

We notice that step (8.6) needs the electric field at time t_{n+1}. It can be computed after step (8.5) by solving the Poisson equation which uses as input ρ_h^{n+1} that needs only $\mathbf{x}_k^{n+1}$ and not $\mathbf{v}_k^{n+1}$.

8.2.6 Time loop.

Let us now summarize the main stages to go from time t_n to time t_{n+1}:

1. We compute the charge density ρ_h and current density $\mathbf{J}_h$ on the grid using relations (8.1)-(8.2).
2. We update the electromagnetic field using a classical mesh based solver (finite differences, finite elements, spectral, ...).
3. We compute the fields at the particle positions using relations (8.3).
4. Particles are advanced using a numerical scheme for the characteristics for example Verlet (8.4)-(8.6).

8.3 Monte Carlo Simulation

8.3.1 Principle

We want to define a Monte Carlo algorithm to approximate some real number a which represents for example the value of an integral. To this aim, we need to construct a real valued random variable X such that

$$\mathbb{E}(X) = a.$$

Then we define an approximation by considering a sequence of independent random variables $(X_i)_i$ distributed like X and approximate $\mathbb{E}(X)$ by the sample mean

$$M_N = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{8.7}$$

In order for this procedure to be useful, we need first to be able to recast our problem in the form of the computation of an expected value of an adequate random variable X that we need to define. Then we need to be able to draw independent variables distributed like X and finally we need to check that the approximation we defined converges in some sense to the exact value and possibly estimate the speed of convergence.

Here the sample mean is an example of what is called an *estimator* in statistics, which is a rule for computing some statistical quantity, which is a function of the random variable, here the expected value, from sample data.

Definition 7 *The difference between the expected value of the estimator and the statistical quantity it approximates is called bias. If this difference is zero, the estimator is said to be unbiased.*

Let us compute the bias of the sample mean given by (8.7), we easily get as the X_i are all distributed like X and thus have the same expected value that

$$\mathbb{E}(M_N) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i) = \mathbb{E}(X)$$

so that the bias is zero and our sample mean is unbiased.

Assuming the sample number $N \geq 2$ an unbiased estimator of the variance is given by the following sample variance

$$V_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - M_N)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \frac{1}{N}\sum_{i=1}^{N}X_i\right)^2. \qquad (8.8)$$

Indeed, let us compute the expected value of V_N. Denoting by $a = \mathbb{E}(X_i)$ for $i = 1, \ldots, N$, we have

$$V_N = \frac{1}{N-1}\sum_{i=1}^{N}((X_i - a) + (a - M_N))^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - a)^2 - \frac{N}{N-1}(M_N - a)^2,$$

as $2\sum_{i=1}^{N}(X_i - a)(a - M_N) = -2N(M_N - a)^2$. Hence

$$\mathbb{E}(V_N) = \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}((X_i - a)^2) - \frac{N}{N-1}\mathbb{E}((M_N - a)^2) = \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{V}(X_i) - \frac{N}{N-1}\mathbb{V}(M_N).$$

And because of Bienaymé's theorem

$$N^2\mathbb{V}(M_N) = \mathbb{V}(\sum_{i=1}^{N}X_i) = \sum_{i=1}^{N}\mathbb{V}(X_i) = N\mathbb{V}(X).$$

Hence

$$\mathbb{E}(V_N) = \frac{N}{N-1}\mathbb{V}(X) - \frac{1}{N-1}\mathbb{V}(X) = \mathbb{V}(X).$$

Remark 25 *Note the $1/(N-1)$ factor in the variance estimator instead of the $1/N$ that one would expect at the first glance. Using $1/N$ instead would also yield an estimator of the variance, but this one would be biased, i.e. it would not have the right expected value.*

8.3.2 Estimation of the error in a Monte Carlo simulation

Assume $\hat{\theta}$ is an estimator for the statistical quantity θ which is a real number that can be computed as a function of a random variable X.

Let us first compute in a general way the mean square error (MSE) of an estimator. The MSE is defined by

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = \int(\hat{\theta} - \theta)^2 dP.$$

Note that the root mean square error or RMS error, which is the square root of the MSE, is the classical L^2 error.

Lemma 7 *Assume the random variable $\hat{\theta}$ is an estimator for θ and $\mathbb{E}(\hat{\theta}^2) < +\infty$. Then*

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{V}(\hat{\theta}) + Bias(\hat{\theta})^2. \tag{8.9}$$

Proof. A straightforward calculation yields

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbb{E}(\hat{\theta}) \\ &= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2 + \mathbb{E}(\hat{\theta})^2 + \theta^2 - 2\theta\mathbb{E}(\hat{\theta}) \\ &= (\mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2) + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\ &= \mathbb{V}(\hat{\theta}) + (Bias(\hat{\theta}))^2. \end{aligned}$$

Assume that the random variable X defining our Monte Carlo simulation verifies $\mathbb{E}(X^2) < +\infty$. Then we can apply the previous lemma to M_N as an estimator of $\mathbb{E}(X)$, which yields

$$MSE(M_N) = \mathbb{V}(M_N) + (\mathbb{E}(M_N) - \mathbb{E}(X))^2.$$

So the RMS error is composed of two parts, the error coming from the variance of the sample and the possible bias on the sample occurring when the expected value of M_N is not exactly equal to the expected value of the random variable X being approximated.

In many cases the bias can be made to be zero, but in some cases it can be useful to introduce some bias in order to decrease the variance of the sample and the total error.

Lemma 8 *Assume $\mathbb{E}(X^2) < +\infty$. Then the RMS error for an unbiased simulation based on the random variable X is*

$$e_{rms} = \sigma(M_N) = \frac{\sigma(X)}{\sqrt{N}}.$$

Proof. The formula (8.9) giving the mean squared error of an estimator shows that if the simulation is unbiased $\mathbb{E}(M_N) = \mathbb{E}(X)$ and

$$e_{rms} = \sqrt{\mathbb{V}(M_N)} = \sigma(M_N).$$

Now using Bienaymé's theorem we also have

$$N^2\mathbb{V}(M_N) = \mathbb{V}(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} \mathbb{V}(X_i) = N\mathbb{V}(X).$$

And thus $\mathbb{V}(M_N) = \mathbb{V}(X)/N$, which gives the result.

On the other hand, Chebyshev's inequality gives us, assuming $\mathbb{E}(X^2) < +\infty$ that for any $\epsilon > 0$ we have, as $\mathbb{E}(X) = \mathbb{E}(M_N)$

$$P\big(|M_N - \mathbb{E}(X)| \geq \epsilon\big) \leq \frac{\mathbb{V}(M_N)}{\epsilon^2} = \frac{\sigma^2(X)}{N\epsilon^2}.$$

Hence when $N \to +\infty$, we have that

$$P\big(|M_N - \mathbb{E}(X)| \geq \epsilon\big) \to 0.$$

This means that M_N converges to $\mathbb{E}(X)$ in probability. This is called the weak law of large numbers. The corresponding strong law of large numbers, the proof of which is more involved, states that M_N converges to $\mathbb{E}(X)$ almost surely, which means that

$$P\left[\omega | \lim_{N\to+\infty} M_N(\omega) = \mathbb{E}(X)\right] = 1.$$

The law of large numbers, strong or weak, implies that the sample mean converges towards the desired expected value, which justifies the Monte Carlo method.

Another major theorem of probability theory, the central limit theorem, gives a precise estimation of the error committed by an approximation.

Theorem 3 (Central Limit Theorem) *Assume $(X_1, X_2, \ldots, X_N)$ is a sequence of independent identically distributed random variables such that $\mathbb{V}(X) = \sigma^2(X) < \infty$. Then*

$$\lim_{N\to+\infty} P\left[\frac{|M_N - \mathbb{E}(X)|}{\sigma(X)/\sqrt{N}} \leq \lambda\right] = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-u^2/2}\, du. \tag{8.10}$$

This tells us that the asymptotic distribution of $\frac{M_N - \mathbb{E}(X)}{\sigma(X)/\sqrt{N}}$ is a unit normal distribution, or equivalently that M_N is a normal distribution with mean $\mathbb{E}(X)$ and standard deviation $\sigma(X)/\sqrt{N}$.

This tells us that the asymptotic distribution of $\frac{M_N - \mathbb{E}(X)}{\sigma(X)/\sqrt{N}}$ is a unit normal distribution, or equivalently that M_N is a normal distribution with mean $\mathbb{E}(X)$ and standard deviation $\sigma(X)/\sqrt{N}$.

The right hand side of (8.10) is a number that can be computed explicitly, and that is called *confidence coefficient*. For $\lambda = 3$ the confidence coefficient is 0.9973 and for $\lambda = 4$ the confidence coefficient is 0.9999 (see e.g. [53] for other values). This is the probability that the true mean lies in the so-called *confidence interval* $[M_N - \lambda\sigma(X)/\sqrt{N}, M_N + \lambda\sigma(X)/\sqrt{N}]$. Note that as opposite to deterministic error estimates, which are generally of the form h^p or $1/N^p$, where h is a cell size and N a number of discretisation points, and lie on a deterministic curve. The error estimate in a Monte Carlo method is random, but it is alway a normal distribution with variance which tends to 0 when the number of sample points tends to $+\infty$. In practice a good estimate of the error is given by $\sigma(X)/\sqrt{N}$, which is all the more interesting that the variance (or standard deviation) can be well estimated by the sample variance (or sample standard deviation), which is an a posteriori estimate that can be directly used in actual computations to measure the error.

8.3.3 Error monitoring in PIC codes

In order to check the validity of simulation, it is important to monitor the evolution of some key quantities. In particular quantities that are conserved in the continuous model should be computed and the accuracy with which they are conserved will give a good indicator of the accuracy of the code: For the Vlasov Poisson system, key conserved quantities are total number of particles $\mathcal{N} = 1 = \int f\,dx\,dv$, total momentum $\mathcal{P} = \int fv\,dx\,dv$ and total energy $\mathcal{E} = \frac{1}{2}\int fv^2\,dx\,dv + \frac{1}{2}\int \rho\phi\,dx$.

In our Monte Carlo approximation, assuming the particles are distributed according to the PDF f, we have

$$\mathcal{N} = \mathbb{E}(1), \quad \mathcal{P} = \mathbb{E}(V), \quad \mathcal{E} = \mathbb{E}(\frac{1}{2}(V^2 + \phi(X))).$$

$\mathbb{E}(1) = \frac{1}{N}N = 1$ is conserved by construction, so there is nothing to monitor. For the others we can compute for a given initial condition the error due to sampling and for subsequent time steps the sample mean and sample standard deviation divided by $\sqrt{N}$ can be monitored to give a measure of the error. This can of coarse be compared to the error given by the actual sample with respect to the conserved value known from the initial condition.

Example. Consider a Landau damping initial condition, on a 1-periodic interval in x:

$$f_0 = (1 + \alpha\cos(kx))\frac{1}{\sqrt{2\pi}}e^{\frac{-v^2}{2}}, \qquad (\alpha < 1).$$

The random variable (X, V) is then randomly drawn according to this distribution. We can then compute

$$\mathbb{E}(V) = \int v f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v = 0, \qquad \mathbb{V}(V) = \int v^2 f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v - 0 = 1.$$

So the RMS error committed by approximating $\mathcal{P}$ by the unbiased estimator $\mathcal{P}_N = \frac{1}{N}\sum_{i=1}^{N} V_i$ will be $1/\sqrt{N}$.

Let us now consider the total energy. First for the kinetic energy, we need to compute

$$\mathbb{E}(V^2) = \int v^2 f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v = 1,$$

$$\mathbb{V}(V^2) = \int v^4 f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v - \mathbb{E}(V^2)^2 = 3 - 1 = 2.$$

On the other hand he potential ϕ associated to the initial condition is solution of $\phi'' = \alpha\cos(kx)$. Assuming 0 average, we get $\phi(x) = -\frac{\alpha}{k^2}\cos(kx)$. We then can compute

$$\mathbb{E}(\phi(X)) = -\frac{\alpha}{k^2}\int \cos(kx) f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v = -\frac{\alpha^2}{2k^2},$$

$$\mathbb{V}(\phi(X)) = \frac{\alpha^2}{k^4} \int \cos^2(kx) f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v - \frac{\alpha^4}{4k^4} = \frac{\alpha^2(2-\alpha^2)}{4k^4}.$$

It follows that the total energy of the initial condition is

$$\mathcal{E} = \mathbb{E}(\frac{1}{2}(V^2 + \phi(X))) = \frac{1}{2} - \frac{\alpha^2}{4k^2},$$

and as the random variable V and $\phi(X)$ are independent, the variance of the total energy is the sum of the variances of the kinetic energy and the potential energy.

A natural estimator for the energy based on the sample $(X_i, V_i)_{1\le i\le N}$, distributed like (X, V), used for the Monte Carlo simulation is here

$$\mathcal{E}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(V_i^2 + \phi(X_i)),$$

from which it easily follows that $\mathbb{E}(\mathcal{E}_N) = \mathbb{E}(\mathcal{E})$ so that the estimator is unbiased. Moreover we can compute the variance of the estimator using Bienaymé's equality

$$\begin{aligned}\mathbb{V}(\mathcal{E}_N) = \mathbb{V}(\frac{1}{2}(V^2 + \phi(X))/N = \frac{1}{4}(\mathbb{V}(V^2) + \mathbb{V}(\phi(X))))/N \\ = \frac{1}{4N}\left(2 + \frac{\alpha^2(2-\alpha^2)}{4k^4}\right).\end{aligned}$$

which is also the MSE error of the estimator as the simulation is unbiased.

After the initial time step, the exact distribution is not known, so that only empirical estimations can be computed. In order to monitor the noise (or error) on each computed quantity $\theta(X)$, we define the relative error

$$R = \frac{\sigma(\theta_N)}{\theta_N},$$

which is the inverse ratio of the estimated value and its standard deviation. We have

$$R = \frac{1}{N-1}\frac{\sqrt{\overline{\theta^2} - \bar{\theta}^2}}{\bar{\theta}},$$

where $\theta(\bar{X}) = \frac{1}{N}\sum_{i=1}^{N}\theta(X_i)$

8.3.4 Error on the probability density

A standard way of estimating a probability density in $\mathbb{R}^d$ from a sample is the kernel density estimator. It relies on a kernel which we shall call S_d, which is a real function in $\mathbb{R}^d$, that we shall assume to be the product of d identical functions: $S_d(x_1, x_2, \ldots, x_d) = S(x_1)S(x_2)\ldots S(x_d)$ verifying $\int S(x)\,\mathrm{d}x = 1$ and $S(x) = S(-x)$ which implies $\int xS(x)\,\mathrm{d}x = 0$.

We then define for a given density $f : \mathbb{R}^d \to \mathbb{R}$

$$\begin{aligned} f_h(x_1, \dots, x_d) &= \frac{1}{h^d} \int S\left(\frac{x_1 - y_1}{h}\right) \dots S\left(\frac{x_d - y_d}{h}\right) f(y_1, y_2, \dots, y_d) \, \mathrm{d}y_1 \dots \, \mathrm{d}y_d, \\ &= \frac{1}{h^d} \mathbb{E}\left(S\left(\frac{x_1 - Y_1}{h}\right) \dots S\left(\frac{x_d - Y_d}{h}\right)\right) \end{aligned}$$

where $(Y_1, \dots, Y_d)$ are distributed according to the density f. From this we can define the following estimator for $f(x_1, \dots, x_d)$, we also use the fact that S is even:

$$f_{h,N}(x_1, \dots, x_d) = \frac{1}{Nh^d} \sum_{i=1}^{N} S\left(\frac{Y_{1,i} - x_1}{h}\right) \dots S\left(\frac{Y_{d,i} - x_d}{h}\right).$$

As usual to estimate the mean squared error committed by this estimator, we compute its bias and variance

$$\begin{aligned} &\mathrm{Bias}(f_{h,N}(x_1, \dots, x_d)) = \mathbb{E}(f_{h,N}(x_1, \dots, x_d)) - f(x_1, \dots, x_d) \\ &\quad = \frac{1}{h^d} \mathbb{E}\left(S\left(\frac{Y_1 - x_1}{h}\right) \dots S\left(\frac{Y_d - x_d}{h}\right)\right) - f(x_1, \dots, x_d) \\ &\quad = \frac{1}{h^d} \int \left(S\left(\frac{y_1 - x_1}{h}\right) \dots S\left(\frac{y_d - x_d}{h}\right)\right) f(y_1, \dots, y_d) \, \mathrm{d}y_1 \dots \, \mathrm{d}y_d \\ &\qquad\qquad - f(x_1, \dots, x_d) \\ &\quad = \int S(z_1) \dots S(z_d) f(x_1 + hz_1, \dots, x_d + hz_d) \, \mathrm{d}z_1 \dots \, \mathrm{d}z_d - f(x_1, \dots, x_d), \end{aligned}$$

making the change of variables $z_1 = \frac{y_1 - x_1}{h}, \dots, z_d = \frac{y_d - x_d}{h}$. Finally as $\int S(z) \, \mathrm{d}z = 1$, and Taylor expanding f assuming enough smoothness we get

$$\begin{aligned} \mathrm{Bias}(f_{h,N}(x_1, \dots, x_d)) &= \int S(z_1) \dots S(z_d)(f(x_1 + hz_1, \dots, x_d + hz_d) \\ &\qquad - f(x_1, \dots, x_d)) \, \mathrm{d}z_1 \dots \, \mathrm{d}z_d \\ &= \int S(z_1) \dots S(z_d) h(z_1 \frac{\partial f}{\partial x_1}(x_1, \dots, x_d) + \dots z_d \frac{\partial f}{\partial x_d}(x_1, \dots, x_d) \\ &\qquad + \frac{h^2}{2} \mathbf{z}^T H(f) \mathbf{z} + O(h^3)) \, \mathrm{d}z_1 \dots \, \mathrm{d}z_d, \end{aligned}$$

where $H(f) = (\frac{\partial^2 f}{\partial x_i \partial x_j})_{1 \le i,j \le d}$ is the Hessian matrix of f and $\mathbf{z} = (z_1, \dots, z_d)^T$. Because of the symmetry of S, the terms in h as well as the off-diagonal second order terms and the third order terms vanish. Hence the bias can be written

$$\mathrm{Bias}(f_{h,N}(x_1, \dots, x_d)) = \frac{h^2}{2} \kappa_2(S) \Delta f(x_1, \dots, x_d) + O(h^4), \tag{8.11}$$

where $\kappa_2(S) = \int x^2 S(x) \, \mathrm{d}x$ is the second order moment of the kernel S and $\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \dots + \frac{\partial^2 f}{\partial x_d^2}$ the Laplace operator. We note that the bias depends

only on h the width of the kernel, and not on the number of particles N. It goes to zero when h goes to 0.

Let us now compute the variance of the estimator. With Bienaymé's equality we get

$$\mathbb{V}(f_{h,N}(x_1,\dots,x_d)^2) = \frac{1}{N}\mathbb{V}\left(\frac{1}{h^d}S\left(\frac{Y_1-x_1}{h}\right)\dots S\left(\frac{Y_d-x_d}{h}\right)\right)$$

Then with the same change of variables as for the bias,

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{h^{2d}}S^2\left(\frac{Y_1-x_1}{h}\right)\dots S^2\left(\frac{Y_d-x_d}{h}\right)\right)\\
&\quad= \frac{1}{h^d}\int S^2(z_1)\dots S^2(z_d)f(x_1+hz_1,\dots,x_d+hz_d)\,\mathrm{d}z_1\dots\,\mathrm{d}z_d\\
&\quad= \frac{1}{h^d}\int S^2(z_1)\dots S^2(z_d)\,\mathrm{d}z_1\dots\,\mathrm{d}z_d(f(x_1,\dots,x_d)+O(h))\\
&\quad= \frac{1}{h^d}R(S)^d(f(x_1,\dots,x_d)+O(h))
\end{aligned}$$

where the $R(S) = \int S^2(x)\,\mathrm{d}x$ is called the roughness of the kernel S. On the other hand, using the previous computation of the bias and the fact that $\int S(x)\,\mathrm{d}x = 1$, we have

$$\mathbb{E}(\frac{1}{h^d}S\left(\frac{Y_1-x_1}{h}\right)\dots S\left(\frac{Y_d-x_d}{h}\right))^2 = (f(x_1,\dots,x_d)+O(h))^2.$$

When $h\to 0$ this term can be neglected compared to the other contribution to the variance. Hence

$$\mathbb{V}(f_{h,N}(x_1,\dots,x_d)) = \frac{R(S)^d}{Nh^d}f(x_1,\dots,x_d)+O(\frac{1}{N}). \tag{8.12}$$

And finally, the mean squared error of the estimator is the sum of its variance and squared bias, which yields

$$\begin{aligned}
MSE(f_{h,N}(x_1,\dots,x_d)) = \frac{R(S)^d}{Nh^d}f(x_1,\dots,x_d)+\frac{h^4}{4}\kappa_2^2(S)(\Delta f)^2(x_1,\dots,x_d)\\
+O(\frac{1}{N})+O(h^6).
\end{aligned} \tag{8.13}$$

Note that for the MSE to converge, one needs obviously the number of samples $N\to+\infty$, $h\to 0$, but also $Nh^d\to+\infty$ for the first term to tend to 0. As h^d is a measure of the cell size in a d-dimensional space, this means that the number of particles per cell needs to converge to $+\infty$. In general in PIC methods, one is not really interested in the convergence of the distribution function, but it is essential to have a good convergence of the density in physical space. For this reason, one generally imposes the number of particles

per cell in physical space to be large enough, and all the larger that the cells become smaller. Keeping the number of particles per cell constant when h decreases does not yield convergence of the method.

To get a unique parameter yielding an order of convergence, one can minimise the dominating terms of $MSE(f_{h,N}(x_1,\ldots,x_d))$ with respect to h, yielding an expression of h in function of N.

Standard kernels in statistics beyond the top hat kernel, are the Gaussian kernel and Epanechnikov type kernels of the form $S(x) = c_s(1-x^2)^s$ for $|x| < 1$ and 0 else, where c_s is a normalisation constant insuring that $\int S(x)\,\mathrm{d}x = 1$. s is a small integer, typically 1,2 or 3 giving the smoothness of the kernel.

In PIC codes S is generally chosen to be a spline function. A spline function of degree m is a piecewise polynomial of degree m and which is in C^{m-1}. It can be defined by recurrence: The degree 0 B-spline that we shall denote by S^0 is defined by

$$S^0(x) = \begin{cases} 1 & \text{if } -\frac{1}{2} \le x < \frac{1}{2}, \\ 0 & \text{else.} \end{cases}$$

Higher order B-splines are then defined by:
For all $m \in \mathbb{N}^*$,

$$\begin{aligned} S^m(x) &= (S^0)^{*m}(x), \\ &= S^0 * S^{m-1}(x), \\ &= \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} S^{m-1}(y)\,dy. \end{aligned}$$

In particular the degree 1 spline is

$$S^1(x) = \begin{cases} (1-|x|) & \text{if } |x| < 1, \\ 0 & \text{else,} \end{cases}$$

the degree 2 spline is

$$S^2(x) = \begin{cases} \frac{1}{2}(\frac{3}{2}-|x|)^2 & \text{if } \frac{1}{2} < |x| < \frac{3}{2}, \\ \frac{3}{4} - x^2 & \text{if } |x| < \frac{1}{2}, \\ 0 & \text{else,} \end{cases}$$

the degree 3 spline is

$$S^3(x) = \frac{1}{6}\begin{cases} (2-|x|)^3 & \text{if } 1 \le |x| < 2, \\ 4 - 6x^2 + 3|x|^3 & \text{if } 0 \le |x| < 1, \\ 0 & \text{else.} \end{cases}$$

8.3.5 Aliasing

In PIC codes, where the evolution of the density function given by the Vlasov equation, needs to be coupled with the computation of the electric field on a grid, aliasing which is inherent to sampling on a grid plays an important role in the choice of the kernel.

Theorem 4 (Shannon) *If support of $\hat{f}$ is included in $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$, then*

$$f(t) = \sum_{k=-\infty}^{+\infty} f(kh)\,\mathrm{sinc}\,(\frac{\pi(t-kh)}{h}),$$

where $\mathrm{sinc}\,t = \frac{\sin t}{t}$ is called the sinus cardinal function.

This means that f is completely determined by sampling with uniform step h if it has bounded support in $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$. However the support of an arbitrary function is generally not in $[-\frac{\pi}{h}, \frac{\pi}{h}]$. If the support is bounded, it is enough to take h small enough. If the support is not bounded but f tends to 0 fast enough at infinity one also gets a good approximation if h is small enough. The question is what happens when h is not small enough to get a good approximation of $\hat{f}$ in $[-\frac{\pi}{h}, \frac{\pi}{h}]$.

In the case when $supp(f) \not\subset [-\frac{\pi}{h}, \frac{\pi}{h}]$, in the formula giving the Fourier transform of a sampled function

$$\hat{f}_h(\omega) = \frac{1}{h} \sum_{n=-\infty}^{+\infty} \hat{f}(\omega - \frac{2n\pi}{h}).$$

the supports of $\hat{f}(\omega - \frac{2n\pi}{h})$ of different n will have a non empty intersection. In particular $\hat{f}(\omega - \frac{2n\pi}{h})$ intersects $[-\frac{\pi}{h}, \frac{\pi}{h}]$ for $|n| \geq 1$. Which means that high frequencies will appear in a low frequency interval. This is called '*aliasing*.

In this case with the reconstruction formula of Shannon's theorem

$$\tilde{f}(t) = (g_h \star f_h)(t) = \sum_{k=-\infty}^{+\infty} f(kh) g_h(t-kh),$$

whose Fourier is

$$\hat{\tilde{f}}(\omega) = \hat{f}_h(\omega)\hat{g}_h(\omega) = h\hat{f}_h(\omega)\chi_{[-\frac{\pi}{h},\frac{\pi}{h}]} = \chi_{[-\frac{\pi}{h},\frac{\pi}{h}]} \sum_{k=-\infty}^{+\infty} \hat{f}(\omega - \frac{2n\pi}{h})$$

which can be very different of $\hat{f}(\omega)$ because of the high frequency contributions.

To suppress aliasing, f needs to be approximated by $\tilde{f}$ which is the closest function in L^2 whose Fourier transform is in $[-\frac{\pi}{h}, \frac{\pi}{h}]$.

Due to Plancherel's formula

$$\begin{aligned} \|f - \tilde{f}\|_2^2 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega) - \hat{\tilde{f}}|^2\, d\omega \\ &= \frac{1}{2\pi} \int_{|\omega| > \frac{\pi}{h}} |\hat{f}(\omega)|^2\, d\omega + \frac{1}{2\pi} \int_{|\omega| < \frac{\pi}{h}} |\hat{f}(\omega) - \hat{\tilde{f}}|^2\, d\omega. \end{aligned}$$

The distance between the two functions is minimal when the second integral vanishes. We hence take $\hat{\tilde{f}}$ to be the restriction of $\hat{f}$ to $[-\frac{\pi}{h}, \frac{\pi}{h}]$, which writes

$$\hat{\tilde{f}}(\omega) = \hat{f}(\omega)\chi_{[-\frac{\pi}{h},\frac{\pi}{h}]}(\omega) = \frac{1}{h}\hat{f}(\omega)\hat{g}_h(\omega),$$

with $g_h(t) = \text{sinc}\,\frac{\pi t}{h}$. It then follows $\tilde{f} = \frac{1}{h} f \star g_h$.

Using the *sinc* would thus suppress all aliasing problems. However it has the problem that its support in physical space is unbounded, which means that all particles would contribute to all grid points, which is very time consuming in practice. Such an algorithm can only be used in practice, when working directly in Fourier space and only a few Fourier modes are needed.

On the other hand, the Fourier transforms of the B-splines are

$$S^m(k) = \text{sinc}^{\,m+1}(\frac{k}{2}),$$

which means that S^m decays like $1/k^{m+1}$ in Fourier space, which is quite fast for quadratic or cubic splines, thus limiting the aliasing problems.

8.4 Initialisation of given PDF

A Monte Carlo simulation relies on a random sequence following some given probability law. Such a sequence can be generated from a uniform random sequence on $[0, 1]$. Obtaining a good approximations of uniform random sequence is a complex task, but some good solutions are given by libraries included with standard compilers or numerical software. We will rely on those. Let us just mention that a computer cannot generate a truly random sequence, but generates two kind of random sequences: the first one called pseudo-random has the objective to provide good approximations of truly random sequences and the other one called quasi-random is designed to fill in the interval as uniformly as possible, yielding a smaller variance.

Then having a good random generator for a uniform random sequence in $[0, 1]$, there are different ways to draw values for any other probability density function. Some are specific to a given form of PDF like normal distributions, other are limited to some class of PDF like products of 1D functions and others are very general. A large number of techniques is described in the book [53]. We will describe the techniques that are the most useful for PIC simulations.

8.4.1 Inversion of the CDF

Let F be the cumulative distribution function (CDF) of the random variable we wish to simulate.

Proposition 19 *Assume $F : [a, b] \to [0, 1]$ is a strictly increasing function. Let U be a uniformly distributed random variable on $[0, 1]$, then $X = F^{-1}(U)$ is a real value random variable with distribution function F.*

Proof. Let $x \in [a, b]$. Then $F^{-1}(U) \leq x \Leftrightarrow U \leq F(x)$.

The distribution function of X is defined by

$$F_X(x) = P(X \leq x) = P(U \leq F(x)) = F(x)$$

as U has a uniform distribution.

In many cases F can be inverted analytically and when $F(x)$ can be computed, it can be inverted numerically using a fine grid and the assumption that F grows linearly between two grid points.

Examples:

1. Uniform distribution on $[a, b]$
   The uniform distribution on $[a, b]$ has the distribution function $F(x) = \frac{x-a}{b-a}$, and to get its inverse we solve the equation $y = F(x) = \frac{x-a}{b-a}$. The solution is
   $$x = a + (b - a)y = F^{-1}(y).$$
2. Numerical inversion of an analytically know distribution function F.
   This amounts for a given point y which is obtained from a uniform distribution in $[0, 1]$ to compute x such that $F(x) = y$, which means solving $y - F(x) = 0$. The most efficient way, in general, to do this numerically is Newton's method which computes x as the limit of the iterations
   $$x_{n+1} = x_n - \frac{y - F(x)}{-F'(x)}.$$
3. Numerical inversion of a function known at discrete grid points.
   We assume the values of F are known on a grid $a = x_0 < x_1 < \cdots < x_{N_x} = b$. Because an approximation is involved in interpolating the values between the grid points, rather than computing directly the inverse for each value y given by the uniform random generator, we start by computing $F^{-1}(y_j)$ where $0 = y_0 < y_1 < \cdots < y_{N_y} = 1$ is a uniform grid of $[0, 1]$. This can be done very easily as F is an increasing function using the following algorithm:

   - $F^{-1}(0) = a$, $F^{-1}(1) = a$, i=0
   - For $j = 1, \ldots, N_y - 1$
     - Find i such that $F(x_{i-1}) < y_i \leq F(x_i)$ (while $(F(x_i) < y_j)$ do $i = i + 1$)
     - Interpolate $F^{-1}(y_j)$ linearly (in order to maintain that F^{-1} is non decreasing between $F(x_{i-1})$ and $F(x_i)$).

Once $F^{-1}(y_j)$ is known on the grid $0 = y_0 < y_1 < \cdots < y_{N_y} = 1$, for any y drawn uniformly on $[0,1]$, find j such that $y_j \leq y < y_{j+1}$ and interpolate linearly $F^{-1}(y)$.

Remark 26 *This method can also be used when F is analytically known by first computing its values on a fine grid. This is generally more efficient that Newton's method and most of the time accurate enough.*

8.4.2 Acceptance-rejection method

This also sometimes simply called the *rejection* method. Assume, we want to draw according to the PDF f and we know how to draw from the PDF g with $f(\mathbf{x}) \leq cg(\mathbf{x})$ for some given constant c. If the support of f vanishes outside of a compact set F we can take for example g uniform in K and $c = \max(f/g)$.

The the rejection algorithm is the following

1. Draw $\mathbf{x}$ from g
2. Draw a uniform random number on $[0,1]$ u
3. If $u \leq f(\mathbf{x})/(cg(\mathbf{x}))$, accept $\mathbf{x}$,
4. else reject $\mathbf{x}$ and start again from (1).

The rejection method is very general, but in order to be efficient the number of rejections should be held as small as possible and cg chosen as close as possible to f, with the constraint of course that one needs to be able to draw from g.

8.4.3 Composition method

This method is also known as the *probability mixing* method and can be used when the PDF that ones wants to sample from is the sum of two simpler PDF. Given two PDF f_1, f_2 that we know how to sample from, and

$$f(x) = \alpha f_1(x) + (1-\alpha) f_2(x), \qquad \text{with } \alpha < 1.$$

A value x can be sampled from f by the following procedure

1. Select a random number r_i from a uniform distribution on $[0,1]$,
2. If $r_i < A_1$ draw x_i according to the PDF f_1,
3. Else draw x_i according to the PDF f_2.

This can be extended to the weighted sum of an arbitrary number of probability density functions. If

$$f(x) = \alpha_1 f_1(x) + \cdots + \alpha_n f_n(x), \qquad \text{with } \alpha_1 + \cdots + \alpha_n = 1.$$

One can then draw from f by drawing a random number r from a uniform distribution on $[0,1]$ and then drawing from f_i if $\alpha_0 + \cdots + \alpha_{i-1} < r < \alpha_0 + \cdots + \alpha_i$, $1 \leq i \leq n$, denoting by $\alpha_0 = 0$.

8.5 Variance reduction techniques

As we saw the Monte Carlo error for the approximation of the expected value of a random variable X is in $\sigma(X)/\sqrt{N}$. Apart from increasing the number of realisations N, the most efficient method to reduce the error is to use available information to replace X by an other random variable with the same expected value but a lower variance. We shall describe a few techniques to do that in the context of Particle in Cell methods.

8.5.1 Control variates

Consider the standard Monte Carlo problem of approximating $a = \mathbb{E}(X)$, for a given random variable X, by a sample mean.

Assume now that there exists a random variable Y the expected value of which is known, that is somehow correlated to X. For a given $\alpha \in \mathbb{R}$, let us define the new random variable

$$Z_\alpha = X - \alpha(Y - \mathbb{E}(Y)).$$

Obviously, we have for any α that $\mathbb{E}(Z_\alpha) = \mathbb{E}(X) = a$, which means that the sample mean of Z_α

$$M_{N,\alpha} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \alpha(Y_i - \mathbb{E}(Y))) = \alpha\mathbb{E}(Y) + \frac{1}{N}\sum_{i=1}^{N}(X_i - \alpha Y_i)$$

could be used instead of the sample mean of X to approximate a. The random variable αY is called a control variate for X.

Let us know look under what conditions the variance of Z_α is lower than the variance of X. We assume that both $\mathbb{V}(X) > 0$ and $\mathbb{V}(Y) > 0$.

Lemma 9 *If the random variables X and Y are not independent, there exists a value of α for which the variance of Z_α is smaller than the variance of X. More precisely*

$$\min_{\alpha\in\mathbb{R}} \mathbb{V}(Z_\alpha) = \mathbb{V}(X)(1-\rho^2(X,Y)) = \mathbb{V}(Z_{\alpha^*}), \quad \textit{with } \alpha^* = \frac{Cov(X,Y)}{\mathbb{V}(Y)}.$$

Moreover

$$\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \Leftrightarrow \left|\begin{array}{ll} \alpha < 2\alpha^* & \textit{if } \alpha > 0, \\ \alpha > 2\alpha^* & \textit{if } \alpha < 0. \end{array}\right.$$

Proof. As $Z_\alpha = X - \alpha Y + \alpha\mathbb{E}(Y)$, and $\mathbb{E}(Z_\alpha) = \mathbb{E}(X)$ we have

$$\begin{aligned}
\mathbb{V}(Z_\alpha) &= \mathbb{E}(Z_\alpha^2) - \mathbb{E}(X)^2, \\
&= \mathbb{E}((X-\alpha Y)^2) + 2\alpha\mathbb{E}(Y)\mathbb{E}(X-\alpha Y) + \alpha^2\mathbb{E}(Y)^2 - \mathbb{E}(X)^2, \\
&= \mathbb{E}(X^2) - 2\alpha\mathbb{E}(XY) + \alpha^2\mathbb{E}(Y^2) + 2\alpha\mathbb{E}(Y)\mathbb{E}(X) - 2\alpha^2\mathbb{E}(Y)^2 \\
&\qquad + \alpha^2\mathbb{E}(Y)^2 - \mathbb{E}(X)^2, \\
&= \mathbb{V}(X) - 2\alpha\mathrm{Cov}(X,Y) + \alpha^2\mathbb{V}(Y), \\
&= \sigma^2(X) - 2\alpha\sigma(X)\sigma(Y)\rho(X,Y) + \alpha^2\sigma^2(Y),
\end{aligned}$$

introducing the standard deviation of a random variable $\sigma^2(X) = \mathbb{V}(X)$ and the correlation coefficient of two random variables

$$\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma(X)\sigma(Y)).$$

So the variance of Z_α is a second order polynomial in α the minimum of which is reached for

$$\alpha^* = \frac{\sigma(X)}{\sigma(Y)}\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma^2(Y)},$$

and plugging this into the expression of $\mathbb{V}(Z_\alpha)$, we get

$$\mathbb{V}(Z_{\alpha^*}) = \sigma^2(X) - 2\sigma(X)^2\rho(X,Y)^2 + \sigma^2(X)\rho(X,Y)^2 = \mathbb{V}(X)(1-\rho^2(X,Y)).$$

On the other hand

$$\mathbb{V}(Z_\alpha) - \mathbb{V}(X) = \alpha\sigma(Y)(\alpha\sigma(Y) - 2\sigma(X)\rho(X,Y)).$$

Hence for $\alpha > 0$,

$$\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \Leftrightarrow \alpha < 2\frac{\sigma(X)}{\sigma(Y)}\rho(X,Y) = 2\alpha^*,$$

and for $\alpha < 0$, $\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \Leftrightarrow \alpha > 2\alpha^*$.

Remark 27 *This results means that provided* $\mathrm{Cov}(X,Y) \neq 0$*, i.e. X and Y are not independent, there is always an interval around the optimal value α^* for which Z_α has a lower variance than X. The more correlated X and Y are, the larger this interval is. So the most important is to find a random variable Y the expectation of which is known, that is as correlated with X as possible. Then if a good approximation of* $\mathrm{Cov}(X,Y)$ *can be computed, one can use this to get closer to α^* and minimise the variance as much as possible with the random variable Y.*

A typical example is when $X = Y + \epsilon\tilde{Y}$, where ϵ is small and $\mathbb{E}(Y)$ is known and for simplicity Y and $\tilde{Y}$ are independent. Plugging this in the expression of $\mathbb{V}(Z_\alpha)$ in the above proof yields

$$\mathbb{V}(Z_\alpha) = \mathbb{V}(Y) + \epsilon^2\mathbb{V}(\tilde{Y}) - 2\alpha\mathbb{V}(Y) + \alpha^2\mathbb{V}(Y) = (1-\alpha)^2\mathbb{V}(Y) + \epsilon^2\mathbb{V}(\tilde{Y}).$$

So that taking $\alpha = 1$ yields that $\mathbb{V}(Z_\alpha)$ is of order ϵ^2 assuming $\mathbb{V}(\tilde{Y})$ of order 1. This is typically the form that is used in PIC simulations.

8.5.2 Importance sampling

We are interested in computing, for some given probability density f, quantities of the form

$$\int \psi(\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

The standard Monte Carlo method for doing this is to define our integral as an expected value using a random variable $\mathbf{Z}$ of density f. Then

$$\int \psi(\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \mathbb{E}(\psi(\mathbf{Z})).$$

Depending on the function ψ it might not be the best approach to use directly the density f for drawing the random variable used in the simulation. Indeed if g is any other probability density that does not vanish in the support of f one can express our integral as an expectation using a random variable $\tilde{\mathbf{Z}}$ of density g:

$$\int \psi(\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \int \psi(\mathbf{z}) \frac{f(\mathbf{z})}{g(\mathbf{z})} g(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \mathbb{E}(W(\tilde{\mathbf{Z}})\psi(\tilde{\mathbf{Z}})),$$

where the random variable $W(\tilde{\mathbf{Z}}) = f(\tilde{\mathbf{Z}})/g(\tilde{\mathbf{Z}})$ is called weight.

The Monte Carlo approximation using independent random variables distributed identically with density g can be expressed as

$$\tilde{M}_N = \frac{1}{N} \sum_{i=1}^{N} W(\tilde{\mathbf{Z}}_i)\psi(\tilde{\mathbf{Z}}_i),$$

from which we get

$$\mathbb{E}(\tilde{M}_N) = \mathbb{E}(W(\tilde{\mathbf{Z}})\psi(\tilde{\mathbf{Z}})) = \int \psi(\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

So $\tilde{M}_N$ is another unbiased estimator of the integral we wish to compute and the approximation error for a given number of samples N is determined by its variance.

Let us now investigate how g can be chosen to get a smaller variance. For this we need to compare the variance of $W(\tilde{\mathbf{Z}})\psi(\tilde{\mathbf{Z}})$ and the variance of $\psi(\mathbf{Z})$ knowing that both have the same expected value.

$$\mathbb{E}(W(\tilde{\mathbf{Z}})^2\psi(\tilde{\mathbf{Z}})^2) = \int \psi(\mathbf{z})^2 W(\mathbf{z})^2 g(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \int \psi(\mathbf{z})^2 W(\mathbf{z}) f(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

On the other hand

$$\mathbb{E}(\psi(\mathbf{Z})^2) = \int \psi(\mathbf{z})^2 f(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

So we see that there is a factor W difference between the two expressions and obviously if $W < 1$ in regions where ψ is larger, the procedure will lead to a smaller variance. Note that because f and g both have an integral one, we cannot have $W < 1$ everywhere.

We also remark that, assuming that $\psi(\mathbf{z})$ does not vanish, if we take $W(\mathbf{z}) = \mathbb{E}(\psi(\mathbf{Z}))/\psi(\mathbf{z})$ which corresponds to $g(\mathbf{z}) = f(\mathbf{z})\psi(\mathbf{z})/\mathbb{E}(\psi(\mathbf{Z}))$, we get

$$\mathbb{E}(W(\tilde{\mathbf{Z}})^2\psi(\tilde{\mathbf{Z}})^2) = \mathbb{E}(\psi(\mathbf{Z}))\int \psi(\mathbf{z})f(\mathbf{z})\,\mathrm{d}\mathbf{z} = \mathbb{E}(\psi(\mathbf{Z}))^2 = \mathbb{E}(W(\tilde{\mathbf{Z}})\psi(\tilde{\mathbf{Z}}))^2$$

so that $\mathbb{V}(W(\tilde{\mathbf{Z}})\psi(\tilde{\mathbf{Z}})) = 0$. This of course cannot be done in practice as $\mathbb{E}(\psi(\mathbf{Z}))$ is the unknown quantity we wish to approximate, but it can be used as a guideline to find a density g that reduces the variance as much as possible and tells us that the density g should be proportional to the integrand $f\psi$, *i.e.* that markers should be distributed according to the integrand.

8.5.3 Application to the PIC method.

For the PIC method, we can combine the importance sampling method and the control variates method.

Importance sampling

The choice of a density for importance sampling depends on the expected value that we are interested in. There are many of those in a PIC code, but arguably the accurate computation of the electric field, which determines the self-consistent dynamics is the most important. Depending on the physical problem we want to deal with more particles will be needed in some specific phase space areas, like for example in some region of the tail for a bump-on-tail instability. For this reason, it is interesting in a PIC code to have the flexibility of drawing the particles according to any density, but one needs to be careful with the choice of this density as the results can become better or worse.

Initialisation. Assume we know the density g_0 according to which we want to draw the markers. Then we initialise the marker's phase space positions $\mathbf{z}_i^0 = (\mathbf{x}_i^0, \mathbf{v}_i^0)$ as realisations of a random variable $\mathbf{Z}^0$ with density g_0.

Time stepping. The markers evolve along the characteristics of the Vlasov equation so that at time t the random variable $\mathbf{Z_t} = (\mathbf{X}_t, \mathbf{V}_t)$ is distributed according to the density $g(t, \mathbf{z})$, that is the solution of the Vlasov-Poisson equation with initial condition g_0.

Then as we saw, the different quantities we need to compute using the Monte Carlo approximation are of the form

$$\int \psi(\mathbf{z})f(t,\mathbf{z})\,\mathrm{d}\mathbf{z} = \int \psi(\mathbf{z})\frac{f(t,\mathbf{z})}{g(t,\mathbf{z})}g(t,\mathbf{z})\,\mathrm{d}\mathbf{z} = \mathbb{E}\left(\psi(\mathbf{Z})\frac{f(t,\mathbf{Z})}{g(t,\mathbf{Z})}\right) \tag{8.14}$$

for some analytically known function $\psi(\mathbf{z})$. This means that we need to simulate the random variable $Y_t = \psi(\mathbf{Z}_t)\frac{f(t,\mathbf{Z}_t)}{g(t,\mathbf{Z}_t)} = \psi(\mathbf{Z}_t)W$, where the random

variable W is defined by $W = f(t, \mathbf{Z}_t)/g(t, \mathbf{Z}_t)$. Because f and g are conserved along the same characteristics we have

$$W = \frac{f(t, \mathbf{Z}_t)}{g(t, \mathbf{Z}_t)} = \frac{f_0(\mathbf{Z}^0)}{g_0(\mathbf{Z}^0)},$$

so that the random variable W does not depend on time and is set once for all at the initialisation.

Using importance sampling, we obtain the so-called weighted PIC method, in which the particles or markers are advance like in the standard PIC method, but have in addition an importance weight which does not evolve in time. The drawback of this method is that the variance can increase when large importance weights and small importance weights are mixed close together in phase space which often happens in long nonlinear simulations.

Control variates

We combine here control variates with importance sampling for most generality, but it can also be used without importance sampling by taking $g_0 = f_0$.

In the PIC method expected values of the form (8.14) cannot be exactly computed because the particle density in phase space $f(t, \mathbf{z})$ is not analytically known except at the initial time. However in many problems, *e.g.* Landau damping, bump-on-tail instability the distribution function stays close to an analytically known distribution function $\tilde{f}(t, \mathbf{z})$. Next to the random variable Y_t associated to $f(t, \mathbf{z})$, this can be used to build the control variate $\tilde{Y}_t$ associated to $\tilde{f}(t, \mathbf{z})$ such that

$$Y_t = \psi(\mathbf{Z})\frac{f(t, \mathbf{Z})}{g(t, \mathbf{Z})}, \qquad \tilde{Y}_t = \psi(\mathbf{Z})\frac{\tilde{f}(t, \mathbf{Z})}{g(t, \mathbf{Z})}.$$

Indeed we have

$$\mathbb{E}(\tilde{Y}_t) = \int \psi(\mathbf{z})\frac{\tilde{f}(t, \mathbf{z})}{g(t, \mathbf{z})} g(t, \mathbf{z})\, \mathrm{d}\mathbf{z} = \int \psi(\mathbf{z})\tilde{f}(t, \mathbf{z})\, \mathrm{d}\mathbf{z}$$

which can be computed analytically for simple enough functions ψ and $\tilde{f}$. Moreover if $\tilde{f}$ is close enough to f then $\tilde{Y}_t$ will be close to Y_t and from the previous discussion a variance reduction of the order of the squared distance between the two random variables can be expected.

Let us now explain how this can be implemented in a PIC simulation.

Initialisation. As for importance sampling, the initial phase space positions of the markers are sampled as realisations $(\mathbf{z}_i^0)_{1 \le i \le N}$ of the random variable $\mathbf{Z}^0$ of density g_0. The importance weights are then defined by the corresponding realisations of the random variable $W = f_0(\mathbf{Z}^0)/g_0(\mathbf{Z}^0)$, *i.e.* $w_i = f_0(\mathbf{z}_i^0)/g_0(\mathbf{z}_i^0)$.

We also initialise the importance weights for $\delta f = f - \tilde{f}$, which are defined by the random variable

$$W_\alpha^0 = \frac{f_0(\mathbf{Z}^0) - \alpha\tilde{f}(t_n, \mathbf{Z}^n)}{g_0(\mathbf{Z}^0)} = W - \alpha\frac{\tilde{f}(0, \mathbf{Z}^0)}{g_0(\mathbf{Z}^0)}.$$

Time stepping. The markers $\mathbf{Z}$ are advanced by numerically solving the characteristics of the Vlasov equation. This means that given their positions $\mathbf{Z}^n$ at time t_n, an ODE solver is used to compute an approximation of their position $\mathbf{Z}^{n+1}$ at time t^{n+1}. Because f and g satisfy the same Vlasov-Poisson equation, they are conserved along the same characteristics so that, as for importance sampling

$$W = \frac{f(t_n, \mathbf{Z}^n)}{g(t_n, \mathbf{Z}^n)} = \frac{f_0(\mathbf{Z}^0)}{g_0(\mathbf{Z}^0)}$$

is a random variable which does not depend on time. On the other hand, we know $\tilde{f}$ analytically and know that f and g are conserved along the characteristics, so that we can compute the importance weight for δf at time t_n from the phase space positions of the markers at the same time:

$$W_\alpha^n = \frac{f(t_n, \mathbf{Z}^n) - \alpha\tilde{f}(t_n, \mathbf{Z}^n)}{g(t_n, \mathbf{Z}^n)} = \frac{f_0(\mathbf{Z}^0) - \alpha\tilde{f}(t_n, \mathbf{Z}^n)}{g_0(\mathbf{Z}^0)} = W - \alpha\frac{\tilde{f}(t_n, \mathbf{Z}^n)}{g_0(\mathbf{Z}^0)}.$$

So W_α^n is a time dependent random variable which can be computed explicitly using the analytical functions $\tilde{f}$, f_0 and g_0. These values can be used to express the sample mean for the new simulated random variable $\tilde{Y}_\alpha = Y - \alpha(\tilde{Y} - \mathbb{E}(\tilde{Y}))$. This is defined by

$$M_{\alpha,N}^n = \frac{1}{N}\sum_{i=1}^{N}(Y_i^n - \alpha\tilde{Y}_i^n) + \alpha\mathbb{E}(\tilde{Y}).$$

Plugging in the values for Y_i^n and $\tilde{Y}_i^n$ we get

$$\begin{aligned} M_{\alpha,N}^n &= \frac{1}{N}\sum_{i=1}^{N}\left(\psi(\mathbf{Z}_i^N)\frac{f(t_n, \mathbf{Z}_i^n) - \alpha\tilde{f}(t_n, \mathbf{Z}_i^n)}{g(t_n, \mathbf{Z}_i^n)}\right) + \alpha\mathbb{E}(\tilde{Y}) \\ &= \frac{1}{N}\sum_{i=1}^{N} W_{\alpha,i}^n\psi(\mathbf{Z}_i^N) + \alpha\mathbb{E}(\tilde{Y}). \end{aligned}$$

This yields an estimator for $\psi(\mathbf{Z})$ based on the weights W_α^n and the expected value that can be computed analytically $\mathbb{E}(\tilde{Y})$. If no estimation of the optimal α^* is available this method is used with $\alpha = 1$.

This is classically known as the δf method in the PIC literature [6, 2], as its interest lies in the expression $f = \tilde{f} + \delta f$ with $\tilde{f}$ known. A large variance reduction for $\alpha = 1$ is obtained as long as $\delta f \ll \tilde{f}$, else one can also achieve some variance reduction by optimising for α [73].

8.6 Coupling the Monte Carlo Vlasov solver with a grid based Poisson solver

The steps of the PIC algorithm are the following

1. Initialisation:
   a) Draw markers $(\mathbf{x}_i, \mathbf{v}_i)$ according to the probability density $g_0(\mathbf{x}, \mathbf{v})$, if g_0 is not the initial particle distribution f_0 compute the importance weights $w_i = f_0(\mathbf{x}_i, \mathbf{v}_i)/g_0(\mathbf{x}_i, \mathbf{v}_i)$.
   b) Compute the initial electric field corresponding to the particles positons by solving the Poisson equation on a grid of physical space. For this a discrete value, depending on the Poisson solver being used, of the charge density $\rho(t, \mathbf{x}) = 1 - \int f(t, \mathbf{x}, \mathbf{v})\,\mathrm{d}\mathbf{v}$ is needed.
2. Time stepping to go from t_n to t_{n+1}:
   a) Push the particles from t_n to t_{n+1} using the known discrete electric field. For this the electric field needs to be evaluated at the particle postions.
   b) Compute the electric field corresponding to the new particle positions.

Next to the Monte Carlo solver for the Vlasov equation an important building block is the grid based Poisson solver and the interaction between the two. We shall distinguish here Poisson solvers needing values at discrete points like Finite Difference or spectral collocation methods and solvers using finite dimensional function spaces like Finite Elements which are coupled with markers in a different manner.

The two steps linked to the coupling, are on the one hand the computation of the discrete charge density needed by the Poisson solver from the particle positions and on the other hand the computation of the electric field at the particle positions.

8.6.1 Finite Difference PIC methods

We consider the 1D Poisson equation on the interval $[0, L]$

$$-\Delta\phi = \rho = 1 - \int f(x, v)\,\mathrm{d}v,$$

with periodic boundary conditions. This is well posed provided the average of ϕ, $\int_0^L \phi(x)\,\mathrm{d}x = 0$. We consider a uniform N_x points discretisation of the periodic interval $[0, L]$, $x_j = j\Delta x = jL/N_x$ for $0 \le j \le N_x - 1$. Because of the periodicity we have for any discrete function $(g_j)_{0 \le j \le N_x - 1}$ that $g_{j+kN_x} = g_j$ for any $k \in \mathbb{Z}$, where we denote by g_j an approximation of $g(x_j)$. The standard second order centred Finite Difference for solving this equation reads

$$\frac{-\phi_{j+1} + 2\phi_j - \phi_{j-1}}{\Delta x^2} = \rho_j \quad \text{for } 0 \le j \le N_x - 1. \tag{8.15}$$

This yields a system of N equation with N unknowns. However all constant vectors are in the kernel of the associated matrix. Hence we need to set the constant to get a unique solution. This can be done thanks to the vanishing average hypothesis on ϕ, which implies on the discrete function that $\sum_{j=0}^{N_x-1} \phi_j = 0$. A second order Finite Difference formula for computing the electric field then writes

$$E_j = -\frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} \quad \text{for } 0 \leq j \leq N_x - 1. \tag{8.16}$$

Proposition 20 *Assume the electrostatic potential is computed from* $(\rho_0, \ldots, \rho_{N_x-1})$ *using* (8.15) *and the electric field using* (8.16) *with periodic boundary conditions. Then we have the following relations:*

$$\sum_{j=0}^{N_x-1} E_j \rho_j = 0, \qquad \sum_{j=0}^{N_x-1} \phi_j \rho_j = \sum_{j=0}^{N_x-1} \frac{(\phi_{j+1} - \phi_j)^2}{\Delta x^2}.$$

Proof. Using (8.15) and (8.16) we first compute

$$\begin{aligned}
\sum_{j=0}^{N_x-1} E_j \rho_j &= -\sum_{j=0}^{N_x-1} \frac{(-\phi_{j+1} + 2\phi_j - \phi_{j-1})}{\Delta x^2} \frac{(\phi_{j+1} - \phi_{j-1})}{2\Delta x}, \\
&= \frac{1}{2\Delta x^3} \sum_{j=0}^{N_x-1} (\phi_{j+1}^2 - \phi_{j-1}^2 - 2\phi_j\phi_{j+1} + 2\phi_j\phi_{j-1}), \\
&= 0.
\end{aligned}$$

Indeed, by change of index, using the periodicity, we have

$$\sum_{j=0}^{N_x-1} \phi_{j+1}^2 = \sum_{j=0}^{N_x-1} \phi_{j-1}^2 \text{ and } \sum_{j=0}^{N_x-1} \phi_j\phi_{j+1} = \sum_{j=0}^{N_x-1} \phi_j\phi_{j-1}.$$

Now multiplying (8.15) by ϕ_j we get, using again periodicity and change of index

$$\begin{aligned}
\sum_{j=0}^{N_x-1} \phi_j \rho_j &= -\sum_{j=0}^{N_x-1} \phi_j \frac{(\phi_{j+1} - \phi_j) - (\phi_j - \phi_{j-1})}{\Delta x^2}, \\
&= \sum_{j=0}^{N_x-1} \frac{(\phi_{j+1} - \phi_j)^2}{\Delta x^2}.
\end{aligned}$$

Remark 28 *The properties in this proposition are discrete versions of the following properties verified by the continuous equations with periodic boundary conditions:*

$$-\int_0^L \rho\nabla\phi \,\mathrm{d}x = \int_0^L \Delta\phi\nabla\phi \,\mathrm{d}x = 0, \quad \textit{and} \int_0^L \rho\phi \,\mathrm{d}x = -\int_0^L \Delta\phi\phi \,\mathrm{d}x = \int_0^L (\nabla\phi)^2 \,\mathrm{d}x.$$

These are necessary conditions for the conservation laws and to have them satisfied at the discrete level, one needs a discrete version of these. As we verified a standard centred second order scheme provides them, but there are many others, like higher order centred schemes or classical spectral Fourier schemes.

8.6.2 Finite Element PIC methods

Still for the 1D Poisson equation on the interval $[0, L]$

$$-\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = \rho = 1 - \int f(x,v)\,\mathrm{d}v,$$

with periodic boundary conditions.

A variational formulation of this equation is obtained by multiplied by a smooth test function and integrating by parts the left hand side. Then the variational formulation reads:

Find $\phi \in H^1_\sharp(0,L)$ *such that*

$$\int_0^L \phi'(x)\psi'(x)\,\mathrm{d}x = \int_0^L \rho(x)\psi(x)\,\mathrm{d}x, \quad \forall \psi \in H^1_\sharp(0,L), \tag{8.17}$$

where we denote $H^1_\sharp(0,L)$ the space of L-periodic functions with vanishing mean.

A Finite Element approximation, is a Galerkin approximation of (8.17), which means that we are looking for a function $\phi_h \in V_h$, with V_h a finite dimensional subspace of $H^1_\sharp(0,L)$, the test functions ψ_h also being in V_h. Expressing the unknown functions ϕ_h and the test functions ψ_h in the same finite dimensional basis of size N_x, the variational formulation in the finite dimensional space is algebraically equivalent to a non singular linear system of size N_x.

We consider now a Finite Element discretisation using the finite dimensional subspace of periodic spline functions of degree p on the uniform grid $x_j = j\Delta x = jL/N_x$:

$$\mathcal{S}_h^p = \{\phi_h \in C^{p-1}_\sharp(0,L) \mid \phi_{h|[x_j,x_{j+1}]} \in \mathbb{P}^p([x_j,x_{j+1}]),\},$$

where $C^{p-1}_\sharp(0,L)$ is the space of L-periodic $p-1$ time continuously derivable functions and the space of polynomials of degree p on the interval $[x_j, x_{j+1}]$ is denoted by $\mathbb{P}^p([x_j,x_{j+1}])$. Then a finite dimensional subspace of $H^1_\sharp(0,L)$ is

$$V_h = \{\phi_h \in \mathcal{S}_h^p \mid \int_0^L \phi_h(x)\,\mathrm{d}x\} = 0.$$

A basis of $\mathcal{S}_h^p$ can be defined using the B-splines of degree p on a uniform periodic grid of step Δx. Those are defined by induction by the de Boor

recursion formula $S_j^0 = 1$ if $j\Delta x \leq x < (j+1)\Delta x$ and 0 else. And for all $p \in \mathbb{N}^*$,

$$S_j^p(x) = \frac{x/\Delta x - j}{p} S_j^{p-1}(x) + \frac{(j+p+1) - x/\Delta x}{p} S_{j+1}^{p-1}(x). \tag{8.18}$$

From this definition, it also easily follows the formula for the derivative of a uniform B-spline

$$\frac{\mathrm{d}S_j^p(x)}{\mathrm{d}x} = \frac{S_j^{p-1}(x) - S_{j+1}^{p-1}(x)}{\Delta x}. \tag{8.19}$$

Using this B-spline basis a function $\phi_h \in V_h$ writes $\phi_h = \sum_{j=0}^{N_x-1} \phi_j S_j^p(x)$, with $\sum_{j=0}^{N_x-1} \phi_j = 0$ so that the average of ϕ vanishes. Plugging this into the variational formulation (8.17) with test function $\psi_h = S_k^p$ for $k = 0, \ldots, N_x - 1$ we get the Galerkin approximation of the Poisson equation

$$\begin{aligned}\sum_{j=0}^{N_x-1} \phi_j \int_0^L S_j'(x) S_k'(x)\,\mathrm{d}x &= \int_0^L \rho(t,x) S_k(x)\,\mathrm{d}x \\ &= 1 - \int_0^L \int_{-\infty}^{+\infty} f(t,x,v) S_k(x)\,\mathrm{d}x\,\mathrm{d}v.\end{aligned}$$

Now for a random variable (X_t, V_t) having density $g(t,x,v)$ and importance weight W with respect to the density $f(t,x,v)$, we have

$$\int_0^L \int_{-\infty}^{+\infty} f(t,x,v) S_k(x)\,\mathrm{d}x\,\mathrm{d}v = \mathbb{E}(W S_k(X)) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i)$$

with our Monte Carlo approximation. Hence the Poisson equation we need to solve, with a source term coming from the Monte Carlo approximation for Vlasov becomes

$$\sum_{j=0}^{N_x-1} \phi_j \int_0^L S_j'(x) S_k'(x)\,\mathrm{d}x = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i). \tag{8.20}$$

This yields the linear system $K\tilde{\phi} = b$, the coefficients of K being

$$\int_0^L S_j'(x) S_k'(x)\,\mathrm{d}x,$$

the components of the column vector $\tilde{\phi}$ being ϕ_j and the components of the column vector b being $\frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i)$. The matrix K is called the *stiffness matrix* in Finite Element terminology. In our case because of the periodic boundary conditions, K is singular of rank $N_x - 1$ as all constant vectors are in its kernel. To get a unique solution of the system we need the additional condition $\sum_{j=0}^{N_x-1} \phi_j = 0$ (for translation invariant basis functions).

As opposed to the Finite Difference discretisation where we need an additional smoothing kernel, the Finite Element basis functions provide the needed regularisation naturally.

Remark 29 *The Galerkin procedure also provides a kernel density estimate by projecting orthogonally in L^2 the density on the span of S^j, $0 \leq j \leq N_x - 1$. This reads*

$$\sum_{j=0}^{N_x-1} \rho_j \int_0^L S_j(x) S_k(x) \, \mathrm{d}x = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i),$$

or as a linear system $M\tilde{\rho} = b$, where M is the matrix with coefficients $\int_0^L S_j(x) S_k(x) \, \mathrm{d}x$ $\tilde{\rho}$ is the column vector of components ρ_j and b is defined as above. M is called mass matrix *in the Finite Element terminology.*

9

Coupling the Vlasov and Maxwell equations

9.1 Introduction

The numerical coupling of the Vlasov and the Maxwell equations introduces new challenges on which we are going to focus in this chapter. This chapter is based on the articles [43, 24, 23].

The model we are going to consider throughout the chapter is the relativistic Vlasov-Maxwell system. The relativistic Vlasov equation for a particle species s reads

$$\frac{\partial f_s}{\partial t} + \mathbf{v}_s(\mathbf{p}) \cdot \boldsymbol{\nabla}_{\mathbf{x}} f_s + q_s(\mathbf{E} + \mathbf{v}(\mathbf{p}) \times \mathbf{B}) \cdot \boldsymbol{\nabla}_{\mathbf{p}} f_s = 0, \tag{9.1}$$

where $\mathbf{v}_s(\mathbf{p}) = \mathbf{p}/(m_s\gamma_s)$, the Lorentz factor being defined by

$$\gamma_s = \sqrt{1 + p^2/(m_s c^2)}$$

with c the velocity of light. Macroscopic quantities relevant to the plasma are obtained as moments in $\mathbf{p}$ of the distribution function f_s for each particle species. In particular the total charge and current densities are defined as

$$\rho = \sum_s q_s f_s(t, \mathbf{x}, \mathbf{p})\, \mathrm{d}\mathbf{p}, \qquad \mathbf{J} = \sum_s q_s \mathbf{v}_s(\mathbf{p}) f_s(t, \mathbf{x}, \mathbf{p})\, \mathrm{d}\mathbf{p}. \tag{9.2}$$

Integrating the Vlasov equation (9.1) over $\mathbf{p}$ and summing over the species yields the following continuity equation that will play a major role in this chapter:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0. \tag{9.3}$$

The self-consistent electromagnetic fields $\mathbf{E}$ and $\mathbf{B}$ appearing in the Vlasov equation satisfy the following Maxwell equations

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \nabla \times \mathbf{B} = -\mathbf{J}/\varepsilon_0 \tag{9.4}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0 \tag{9.5}$$

$$\nabla \cdot \mathbf{E} = \rho/\varepsilon_0 \tag{9.6}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{9.7}$$

in addition to initial and boundary conditions. Taking the divergence of the Ampere equation (9.4) and using the Gauss law (9.6), we obtain

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0,$$

which means that the continuity equation (9.3) is a compatibility condition for Maxwell's equations, those being ill-posed when the continuity equation is not satisfied. Moreover it can be shown that provided the divergence constraints (9.6)-(9.7) are satisfied at the initial time, they remain satisfied for all times by the solution of Ampere (9.4) and Faraday (9.5), which have a unique solution by themselves provided adequate initial and boundary conditions.

As we saw the continuity equation is a consequence of the Vlasov equation, so that at the continuous level all is fine for the Vlasov-Maxwell equations. However at the discrete level, there is no reason, for a given discretization of the Vlasov and the Maxwell equations that a discrete continuity equation compatible with the discrete Maxwell equations holds. Even though this is generally the most acute problem in electromagnetic PIC simulations, the way in which the Gauss laws (or divergence contraints) are satisfied needs to be compatible with the discrete Ampere and Faraday equations. Handling these compatibility issues is one way to solve the problem. Another is to modify the Maxwell equations, so that they are well posed independently of the sources, by introducing two additional scalar unknowns that can be seen as Lagrange multipliers for the divergence constraints. These should become arbitrarily small when the continuity condition is close to being satisfied.

The aim of this chapter is to review the different methods that have been proposed in the literature and classify them in one of the two above categories: using a structure preserving discretization with compatible discrete Gauss laws and continuity equation or using a generalised set of Maxwell equations with additional unknowns that are easier to discretise. Indeed the infinite dimensional kernel of the curl operator and the lack of compactness of the inverse Maxwell operator has made it particularly hard to find good discretization for Maxwell's equations, especially for the eigenvalue problem [18, 19, 22, 29, 68].

9.2 Generalised Maxwell's equations

Even though, provided the divergence constraints are satisfied at the initial condition, they are satisfied at all times for the continuous Maxwell's equations when only Ampere (9.4) and Faraday's (9.5) are solved, this is not true when the sources are computed numerically from a Particle In Cell (PIC) method or from a grid based Vlasov solver. This has been recognised early in the PIC literature and the first solution proposed by Boris [21], the so-called Boris correction, consists in correcting a posteriori, after each field solve, the electric field $\mathbf{E}$ into $\tilde{\mathbf{E}} = \mathbf{E} + \nabla\varphi$ such that $\nabla \cdot \tilde{\mathbf{E}} = \rho$. This yields the Poisson equation

$$-\Delta\varphi = \nabla \cdot \mathbf{E} - \rho.$$

In order to avoid a costly Poisson solved, Marder [80] proposed the following correction of the electric field

$$\tilde{\mathbf{E}}^{n+1} = \mathbf{E}^{n+1} + \Delta t\, \mathbf{grad}(\nu(\nabla \cdot \mathbf{E}^n - \rho^n))$$

ν is a diffusion parameter chosen small enough for stability. This method has been improved by Langdon [76]

$$\tilde{\mathbf{E}}^{n+1} = \mathbf{E}^{n+1} + \Delta t\, \mathbf{grad}(\nu(\nabla \cdot \mathbf{E}^{n+1} - \rho^{n+1}))$$

This can also be seen as one Jacobi iteration for solving the Poisson equation proposed by Boris. A comparison of these methods is performed in [79].

These classical method can all be interpreted as imposing the divergence constraint on the electric field by using a Lagrange multiplier using the following generalised formulation of Maxwell's equations introduced in [83]

$$\begin{aligned} \partial_t \mathbf{E} - c^2 \nabla \times \mathbf{B} + c^2 \nabla p &= -\frac{\mathbf{J}}{\epsilon_0}, \\ \partial_t \mathbf{B} + \nabla \times \mathbf{E} &= 0, \\ g(p) + \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}, \\ \nabla \cdot \mathbf{B} &= 0. \end{aligned}$$

This implies $\frac{\partial g(p)}{\partial t} - c^2 \Delta p = \frac{1}{\epsilon_0}(\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J})$.

A mathematical study of this system has been performed in [7].

The Boris correction is equivalent to the case $g = 0$. Then the Lagrange multiplier p satisfies a Poisson equation with source $\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J}$. The Marder and Langdon corrections are equivalent to two different discretizations of these equations with $g(p) = p$ in which case p satisfies a heat equation transporting the continuity error out of the domain. It then becomes natural to also consider the case $g(p) = \partial_t p$ in which case the Lagrange multiplier satisfies a wave equation and the generalised Maxwell's equations become hyperbolic and even strictly hyperbolic if a Lagrange multiplier is also used for the $\nabla \cdot \mathbf{B}$ constraint. This set of generalised Maxwell's equations reads

$$\begin{aligned}\partial_t \mathbf{E} - c^2 \nabla \times \mathbf{B} + c^2 \nabla p &= -\frac{\mathbf{J}}{\epsilon_0},\\ \partial_t \mathbf{B} + \nabla \times \mathbf{E} + \nabla q &= 0,\\ \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0},\\ \frac{\partial q}{\partial t} + \nabla \cdot \mathbf{B} &= 0.\end{aligned}$$

This is called the hyperbolic correction and was introduced in [82]. The error is transported out of the domain fast enough to avoid accumulation. For this absorbing boundary conditions are needed.

The idea has also been adapted for imposing the $\nabla \cdot \mathbf{B} = 0$ constraint in MHD by Dedner, Kemm, Kröner, Munz, Schnitzer, Wesenberg with considerable success [47].

The hyperbolic generalised Maxwell operator has a compact inverse which can be used to prove the existence and uniqueness of solutions [77], which is not the case for the standard Maxwell's equations. In this case the compactness of the evolution operator is guaranteed for divergence free functions. Hence the Gauss laws enforce that the solution remains in the correct domain. This is what will need to be reproduced at the discrete level. For a generalised formulation such a problem does not exists which makes it robust for any kind of consistent discretizations. However this comes at some expenses: First two new scalar unknowns are introduced and thus the system to be solved becomes larger. Then these unknowns need boundary conditions which needed to be found according to the physics problem which might not always be easy, finding good boundary conditions that dissipate the error is particularly challenging for the hyperbolic correction. Moreover all the generalised Maxwell's equations introduce new non physical propagation speeds in the equations which need to be tuned so as not to distort the physics, which is not always easy. In particular for elliptic and parabolic corrections these wave speeds are infinite. A situation where this is particularly problematic is laser plasma interactions where instabilities can be triggered before the laser hits the plasma due to these spurious wave speeds. This motivates the need of structure preserving algorithms, which might be more complicated.

9.3 Structure preserving discretizations

At the continuous level, we have seen that the Gauss law was preserved thanks to the fact that taking the divergence of the Ampere equation (9.4) and using the continuity equation (9.3) yields $\frac{\partial}{\partial t}(\nabla \cdot \mathbf{E} - \rho/\varepsilon_0) = 0$. Here we have used in addition to the continuity equation that the divergence of a curl always vanishes, i.e., for all $\mathbf{F}$

$$\nabla \cdot \nabla \times \mathbf{F} = 0. \tag{9.8}$$

The idea of structure preserving discretizations is to get a discrete version of these relations. Thus, we shall look for (semi-) discrete approximations of the Maxwell equations of the form

$$\begin{cases} \dfrac{\partial \mathbf{E}_h}{\partial t} - \mathbf{curl}_h\, B_h = -\dfrac{1}{\varepsilon_0}\mathbf{J}_h \\ \dfrac{\partial B_h}{\partial t} + \mathrm{curl}_h\, \mathbf{E}_h = 0 \end{cases} \tag{9.9}$$

with the following properties:

a) the approximate sources must satisfy a discrete continuity equation

$$\frac{\partial \rho_h}{\partial t} + \mathrm{div}_h\, \mathbf{J}_h = 0\ ; \tag{9.10}$$

b) the underlying discrete operators must satisfy a property analogous to that of the continuous ones, namely

$$\mathrm{div}_h \mathrm{curl}_h = 0. \tag{9.11}$$

Clearly, the resulting field will then preserve the corresponding Gauss law,

$$\mathrm{div}_h\, \mathbf{E}_h = \frac{1}{\varepsilon_0}\rho_h \tag{9.12}$$

and a similar procedure can be applied for the magnetic field. Numerical methods satisfying the above properties are often said to be *charge conserving* because no spurious charges appear in the longitudinal (i.e., curl-free) part of $\mathbf{E}_h$.

9.3.1 Enforcing a discrete continuity equation

The typical cases where the program outlined above gives satisfactory results is provided by Finite Differences (Yee) schemes and curl-conforming Finite Elements which can be seen as an extension of the former method to higher orders and unstructured meshes. In all these methods, the charge density is computed in the classical way described previously but the current density is computed differently. Consistent with the physical interpretation, the current is deposited on all cell faces through which a particle passes.

In the scope of Finite Differences the core idea has been introduced by Villasenor and Buneman [104] for the classical cloud in cell method, where particles with hat function shapes (piecewise $\mathbb{Q}_1$ basis functions) are coupled with the Yee scheme [106], and generalised to arbitrary B-spline shape functions by Barthelmé and Parzani [8]. Using a splitting technique, Esirkepov could simplify and accelerate the procedure forcing particle displacements along the axes [57]. In the same spirit Umeda and co-workers introduced a fast procedure for the lowest order scheme. As in this case only the end points of the trajectory are involved in the definition of the charge density, they modify the trajectory between the end points so that the particles cross cell boundaries only through the grid points [101].

9.3.2 Conforming mixed Finite Elements

Structure preserving Finite Elements are provided by the discrete exact sequence property that was introduced in section 6.4.3. This provides naturally that $\mathrm{div}_h \, \mathrm{curl}_h \, = 0$ as the continuous divergence and curl operators are applied directly in the discrete spaces. We need to see now how the current and the charge density need to be approximated so as to satisfy a discrete continuity equation.

In the framework of the Finite Element method a conservative current deposition scheme has been introduced by Eastwood [54, 55] and generalised in [24] to curl-conforming Finite Elements of arbitrary orders on unstructured meshes.

Following the notations of section 6.4.3. We consider conforming Finite Element spaces $X \subset H^1(\Omega)$, $W \subset H(\mathrm{curl}, \Omega)$ and $V \subset L^2(\Omega)$ that can be defined on quadrangles or on triangles.

Based on this spaces, the standard Finite Element approximation of the time-dependent Maxwell equations consists of finding $\mathbf{E}_h(t) \in W$ and $B_h(t) \in V$ such that

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \displaystyle\int \mathbf{E}_h \cdot \mathbf{F} \, \mathrm{d}x - \int B_h \, \mathrm{curl} \, \mathbf{F} \, \mathrm{d}x = -\frac{1}{\varepsilon_0} \int \mathbf{J}_h \cdot \boldsymbol{\varphi} \, \mathrm{d}x \qquad \forall \mathbf{F} \in W \\ \dfrac{\mathrm{d}}{\mathrm{d}t} \displaystyle\int B_h \cdot C \, \mathrm{d}x + \int \mathrm{curl} \, \mathbf{E}_h \, C \, \mathrm{d}x = 0 \qquad \forall C \in V \end{cases} \tag{9.13}$$

holds for all t. We note that this method corresponds to defining the discrete curl operators in (9.9) by

$$\mathrm{curl}_h \, : W \ni \mathbf{w} \mapsto \mathrm{curl}_h \, \mathbf{v} \in V$$

and

$$\mathbf{curl}_h \, := (\mathrm{curl}_h \,)^* :\to W,$$

where we recall that the latter amounts to setting

$$\int \mathbf{curl}_h \, u \cdot \mathbf{v} \, \mathrm{d}x := \int u \, \mathrm{curl}_h \, \mathbf{v} \, \mathrm{d}x \quad \forall \mathbf{v} \in W.$$

Given these operators, charge-conserving PIC schemes are based on computing the current density $\mathbf{J}_h$ from the particles in such a way that a discrete continuity equation in Finite Element form, that is a discrete counterpart of (6.45) holds. This reads

$$-\int \mathbf{E}_h \cdot \nabla q \, \mathrm{d}\mathbf{x} = \frac{1}{\varepsilon_0} \int_\Omega \rho_h \, q \, \mathrm{d}\mathbf{x} \quad \forall q \in X. \tag{9.14}$$

As in the continuous case, this equation is a consequence of (9.13). Indeed as for any $q \in X$ we have $\nabla q \in W$, one can use $\mathbf{F} = \nabla q$ in (9.13), from which it follows that

$$\frac{d}{dt}\int \mathbf{E}_h \cdot \nabla q = -\int_\Omega \mathbf{J}_h \cdot \nabla q \quad \forall q \in X.$$

Then if ρ_h and $\mathbf{J}_h$ are defined such that they verify the following variational discrete continuity equation

$$\frac{d}{dt}\int \rho_h\, q\,\mathrm{d}\mathbf{x} = \int \mathbf{J}_h \cdot \nabla q\,\mathrm{d}\mathbf{x} \quad \forall q \in X, \tag{9.15}$$

we find that

$$\frac{d}{dt}\big(\int \mathbf{E}\cdot\nabla q + \int \rho\, q\big) = 0 \quad \forall q \in X,$$

so that if the discrete variational Gauss law (9.14) is satisfied at the initial time it remains satisfied for all times.

In practice the particle current must be deposited in such a way that a fully discrete version of (9.14) is satisfied, which is essentially done by averaging in time the standard PIC current carried by the particles. In a leap-frog time scheme for instance, defining

$$\mathbf{J}_S^{n+\frac{1}{2}}(\mathbf{x}) := \int_{t_n}^{t_{n+1}} \mathbf{J}_S(\mathbf{x},t)\frac{\mathrm{d}t}{\Delta t} = \sum_{k=1}^{N} q w_k \int_{t_n}^{t_{n+1}} S(\mathbf{x}-\mathbf{x}_k(t))\mathbf{v}(\mathbf{p}_k(t))\frac{\mathrm{d}t}{\Delta t} \tag{9.16}$$

yields [24, Lemma 3.3]

$$\int (\rho_S^{n+1}-\rho_S^n)\, q\,\mathrm{d}x - \Delta t \int \mathbf{J}_S^{n+\frac{1}{2}}\cdot\nabla\varphi\,\mathrm{d}x = 0 \qquad q \in X,$$

where the time-discrete charge density is just $\rho_S^n(\mathbf{x}) := \rho_S(t_n,\mathbf{x})$. Thus a fully discrete version of (9.14) holds with ρ_h^n and $\mathbf{J}_h^{n+1/2}$ defined as the orthogonal projections of ρ_S^n and $\mathbf{J}_S^{n+1/2}$ on the continuous and curl-conforming finite element spaces, respectively. As for the source vector involved in the matrix form of (a fully discrete version of) the Finite Element Method (9.13), its entries are the moments of the discrete current $\mathbf{J}_h^{n+1/2}$ against the basis functions $\boldsymbol{\varphi}_i$ of W. In the case of point particles ($S=\delta$) their value is

$$\mathbf{J}_i^{n+\frac{1}{2}} := \int \mathbf{J}_h^{n+\frac{1}{2}}\cdot\boldsymbol{\varphi}_i\,\mathrm{d}x = \int \mathbf{J}_S^{n+\frac{1}{2}}\cdot\boldsymbol{\varphi}_i\,\mathrm{d}x = \sum_{k=1}^{N} q w_k \int_{t_n}^{t_{n+1}} \mathbf{v}(\mathbf{p}_k(t))\boldsymbol{\varphi}_i(\mathbf{x}_k(t))\frac{\mathrm{d}t}{\Delta t} \tag{9.17}$$

and for piecewise affine trajectories the function $t \mapsto \mathbf{v}_k^{n+\frac{1}{2}}\cdot\boldsymbol{\varphi}_i(\mathbf{x}_k(t))$ is itself polynomial on every time interval $[\tau,\tau+\Delta\tau]\subset[t_n,t_{n+1}]$ that a particle spends inside a cell. In particular a Gauss formula with enough quadrature points is exact, *i.e.*

$$\int_\tau^{\tau+\Delta\tau} \boldsymbol{\varphi}_i(\mathbf{x}_k(t))\frac{\mathrm{d}t}{\Delta t} = \frac{\Delta\tau}{2\Delta t}\sum_{j=1}^{q}\lambda_j\,\boldsymbol{\varphi}_i\big(\mathbf{x}_k(\tau_j)\big)$$

where q needs to be chosen in compliance with the degree of the Finite Element functional space (for instance, with the above choice of Nedelec elements one must take $q \geq \frac{p+1}{2}$ if a Gauss-Legendre quadrature is used).

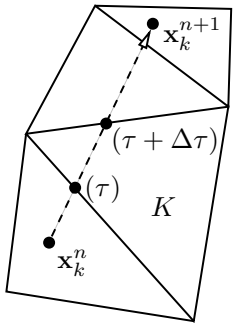

Fig. 9.1. Charge-conserving current deposition on an unstructured grid

9.3.3 The finite difference Yee scheme

The case of Finite Difference schemes can be described with the same arguments, since at lowest order ($p = 1$) the above Finite Element method applied on a cartesian mesh with the mass lumping procedure of Cohen and Monk [36] is equivalent to the Yee scheme [81]. The above deposition method then coincides with that of Villasenor and Buneman [104].

Let us now derive the explicit formulation of our first order method on quadrangles. We shall notice that it is completely equivalent to the charge conserving method of Villasenor and Buneman coupled to Yee's Maxwell solver.

Let us first recall the definition of the B-splines. They are defined inductively by $S^0(x) = 1$ if $x \in [0, 1]$ and 0 elsewhere, and $S^m(x) = S^{m-1} \star S^0$. In particular, S^1 is the hat function, $S^1(x) = 1 + x$ if $x \in [-1, 0]$, $S^1(x) = 1 - x$ if $x \in [0, 1]$ and 0 elsewhere.

The B-splines are rescaled to the cell size h by $S_h(x) = \frac{1}{h} S(\frac{x}{h})$. Note that the definition of S_h implies that

$$\int_{-\infty}^{+\infty} S_h(x)\, dx = \int_{-\infty}^{+\infty} S(x)\, dx = 1.$$

A straightforward computation shows that the basis functions of our first order space W and V on squares of side h are of the following form

$$\varphi^1_{i+\frac{1}{2},j} = \begin{pmatrix} S^0_h(x - x_i) S^1_h(y - y_j) \\ 0 \end{pmatrix}, \qquad \varphi^1_{i,j+\frac{1}{2}} = \begin{pmatrix} 0 \\ S^1_h(x - x_i) S^0_h(y - y_j) \end{pmatrix},$$

$$\varphi^2_{i+\frac{1}{2},j+\frac{1}{2}} = S^0_h(x - x_i) S^0_h(y - y_j).$$

The indices characterizing the basis functions correspond to the point at which their associated degrees of freedom are located.

We have $(S^1_h)'(x) = \frac{1}{h}(S^0_h(x - h) - S^0_h(x)$. Hence

$$\begin{aligned}\operatorname{curl}\varphi^1_{i+\frac12,j} &= \frac{1}{h}S^0_h(x-x_i)(S^0_h(y-y_j)-S^0_h(y-y_{j-1})),\\ &= \frac{1}{h}(\varphi^2_{i+\frac12,j+\frac12}-\varphi^2_{i+\frac12,j-\frac12}),\end{aligned}$$

$$\begin{aligned}\operatorname{curl}\varphi^1_{i,j+\frac12} &= \frac{1}{h}(S^0_h(x-x_i)-S^0_h(x-x_{i-1}))S^0_h(y-y_j),\\ &= \frac{1}{h}(\varphi^2_{i+\frac12,j+\frac12}-\varphi^2_{i-\frac12,j+\frac12}).\end{aligned}$$

This enables us to write the matrix K

On the other hand, the lumped mass matrix M_1 is diagonal and its diagonal terms are

$$\int_\Omega(\varphi^1_{i+\frac12,j})^2\,d\mathbf{x}=\int_\Omega(\varphi^1_{i,j+\frac12})^2\,d\mathbf{x}\approx\frac{1}{h}^2,$$

the lumping being done using the trapezoidal rule. Finally the matrix M_2 is diagonal, its diagonal terms being

$$\int_\Omega(\varphi^2_{i+\frac12,j+\frac12})^2\,d\mathbf{x}=\frac{1}{h}^2.$$

Thus equations (9.13) become in this case for each integer couple (i,j)

$$\begin{aligned}\frac{d}{dt}E_{x,i+\frac12,j}-\frac{c^2}{h}(B_{z,i+\frac12,j+\frac12}-B_{z,i+\frac12,j-\frac12})&=J_{x,i+\frac12,j},\\ \frac{d}{dt}E_{y,i,j+\frac12}+\frac{c^2}{h}(B_{z,i+\frac12,j+\frac12}-B_{z,i-\frac12,j+\frac12})&=J_{y,i,j+\frac12},\\ \frac{d}{dt}B_{z,i+\frac12,j+\frac12}+\frac{1}{h}\big(E_{y,i,j+\frac12}-E_{y,i,j-\frac12}-E_{x,i+\frac12,j}+E_{x,i-\frac12,j}\big)&.\end{aligned}$$

Using a leap-frog scheme for time integration, we recognize the classical Yee scheme on staggered meshes.

Let us now consider the expression yield by (9.17) in our case. Noticing that $J_{x,i+\frac12,j}$ corresponds to the basis function $\varphi^1_{i+\frac12,j}$ and that $J_{y,i,j+\frac12}$ corresponds to the basis function $\varphi^1_{i,j+\frac12}$ we get

$$J_{x,i+\frac12,j}=\frac{v_x^{n+\frac12}}{\Delta t}\int_{t_n}^{t_{n+1}}S^0_h(x_i-x_k(t))S^1_h(y_j-y_k(t))\,dt,$$

$$J_{y,i,j+\frac12}=\frac{v_y^{n+\frac12}}{\Delta t}\int_{t_n}^{t_{n+1}}S^1_h(x_i-x_k(t))S^0_h(y_j-y_k(t))\,dt,$$

which is exactly the charge conserving expression given by Villasenor and Buneman.

A

Complex analysis and Laplace transform

We recall here the necessary mathematical background that is needed for the solution of the linearised Vlasov-Poisson system and the computation of approximate solutions of the dispersion relations that come out. For a more detailed introduction to complex analysis we refer to the textbook by Ahlfors [1] .

A.1 Analytic functions

Definition 8 *Let u be an open subset of $\mathbb{C}$. A function from $U \subset \mathbb{C} \to \mathbb{C}$ is called holomorphic on U if it has a derivative defined by*

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

for all $z_0 \in U$.

Definition 9 *A function $f : U \subset \mathbb{C} \to \mathbb{C}$ is analytic on U if it writes for all $x \in U$*

$$f(x) = \sum_n c_i (x - x_0)^n.$$

Proposition 21 *Every function that is analytic on an open set U is also holomorphic on U and conversely.*

Proposition 22 (Properties of analytic functions)
The sum and product of two analytic functions is analytic.
The ratio f/g of two analytic functions f and g is analytic at all point where g does not vanish.
The composition of two analytic functions is analytic.

Proposition 23 (Characterisation of analytic functions) *Let $f : \Omega \to \mathbb{C}$ such that $f(z) = u(z) + iv(z)$ with $z = x + iy$, $u(z) = \Re(f(z))$ et $v(z) = \Im(f(z))$. Then f is analytic if and only if $(u, v) \in C^1(\mathbb{C})^2$ with*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{A.1}$$

The equations (A.1) are called Cauchy-Riemann equations.

A.2 Path integration

Let γ be an arc of the complexe plane parametrised by $t \mapsto z(t)$ for $t \in [a, b]$. The integral along the contour γ is defined by

$$\int_\gamma f(z)\, dz = \int_a^b f(z(t))z'(t)\, dt.$$

This definition is independent of the chosen parametrisation.

Example.

Let us compute

$$\int_\gamma \frac{dz}{z-a}$$

where γ is the circle of centre a and radius R. A parametrisation of γ is then given by $z(\theta) = a + Re^{i\theta}$, $\theta \in [0, 2\pi]$. We then have $z'(\theta) = iRe^{i\theta}$. Hence

$$\int_\gamma \frac{dz}{z-a} = \int_0^{2\pi} \frac{iRe^{i\theta}}{Re^{i\theta}} = 2\pi.$$

Theorem 5 *If $f : \Omega \to \mathbb{C}$ is analytic in the domain Ω and γ is a contour of Ω, then $\int_\gamma f(z)\, dz$ depends only on the end points of γ.*

Theorem 6 (Cauchy) *Let γ be a closed curve and f an analytic function on a domain containing γ and its interior. Then $\int_\gamma f(z)\, dz = 0$.*

Definition 10 *Let f of function of the form*

$$f(z) = \frac{c_{-k}}{(z-a)^k} + \cdots + \frac{c_{-1}}{z-a} + \varphi(z),$$

with φ analytic in the neighbourhood of a. We then call residue *of f at a, and we denote by* $\operatorname{Res}_{z=a} f(z)$*, le terme c_{-1}.*

In the frequent case when f admits a single pole at a, we have

$$f(z) = \frac{c_{-1}}{z-a} + \varphi(z), \quad \text{et } \operatorname{Res}_{z=a} f(z) = f(z)(z-a)\big|_{z=a}.$$

Definition 11 *We definie the* index *of point a with respect to the closed curve γ by*

$$n(\gamma, a) = \frac{1}{2\pi i} \int_\gamma \frac{dz}{z-a}$$

Remark 30 *The index of γ with respect to a, denoted by $n(\gamma, a)$ corresponds to the winding number of the curve γ around the point a.*

Theorem 7 (Residue theorem) *Let Ω be a simply connected set of the complex plane and let f be an analytic function of Ω minus p isolated singularities denoted by a_j, $j = 1, \ldots, p$. Then*

$$\int_\gamma f(z)\,dz = 2i\pi \sum_{j=1}^{p} n(\gamma, a_j)\, Res_{z=a_j} f(z) \tag{A.2}$$

for any close curve γ which does not go through a singularity.

Remark 31 *In practice, we shall build contours that winds either once or not at all around a given pole, so that $n(\gamma, a_j) = 1$ if a_j lies inside γ and 0 else. The residue formula can then be written in a simpler form*

$$\int_\gamma f(z)\,dz = 2i\pi \sum_{j|\ a_j inside} \mathrm{Res}_{z=a_j} f(z).$$

The sum is taken only on the j such that a_j is inside the curve γ.

A.3 Laplace transform

For a detailed discussion of the Laplace transform, the reader can refer to the textbooks of Schwartz [93] or of Bellman-Roth [11].

Let $f \in L^1(\mathbb{R}_+)$ such that there exist real valued constants a and b such that $|f(t)| \leq ae^{bt}$. Then

$$\int_0^{+\infty} |e^{-st} f(t)|\,dt \leq a \int_0^{+\infty} |e^{(b-s)t}|\,dt < +\infty$$

for $\Re(s) > b$. We define in this case for $s \in \mathbb{C}$ such that $\Re(s) > b$, the Laplace transform $\tilde{f}(s)$ of f by

$$\tilde{f}(s) = \int_0^{+\infty} f(t) e^{-st}\,dt. \tag{A.3}$$

Remark 32 *The Laplace transform of a real valued function takes its values in $\mathbb{C}$ and is defined in the half plane $\Re(s) > b$ which can also be the whole $\mathbb{C}$ or the empty set depending on the behaviour of f for large times.*

The following theorem gives a practical framework in which the Laplace transform can be inverted.

Theorem 8 *We assume that there exist two real valued constants M and R such that*

i) $\tilde{f}$ *is analytic in the half plane* $\Re(s) > R$,
ii) $|s\tilde{f}(s)| \leq M$ *for all* s *such that* $|s| > R$.

Then, if we define

$$f(t) = \frac{1}{2i\pi} \int_{u-i\infty}^{u+i\infty} \tilde{f}(s) e^{st}\, ds \quad \forall t > 0,\ u > R, \tag{A.4}$$

$\tilde{f}(s)$ *is the Laplace transform of* f.

Example.

Let us show how to solve a differential equation using the Laplace transform. We consider the differential equation

$$\frac{dy}{dt} + y = 1, \quad y(0) = y_0.$$

We start by applying the Laplace transform of the differential equation, multiplying it by e^{-st} and integrating between 0 an $+\infty$. Assuming $\Re(s) > 0$, we have

$$\int_0^{+\infty} \frac{dy}{dt} e^{-st}\, dt = [ye^{-st}]_0^{+\infty} + s \int_0^{+\infty} ye^{-st}\, dt = -y_0 + s\tilde{y}.$$

On the other hand, $\int_0^{+\infty} ye^{-st}\, dt = \tilde{y}(s)$ and

$$\int_0^{+\infty} e^{-st}\, dt = \left[\frac{e^{-st}}{-s}\right]_0^{+\infty} = \frac{1}{s}.$$

The Laplace transform of the differential equation then writes

$$-y_0 + (s+1)\tilde{y} = \frac{1}{s},$$

and thus

$$\tilde{y}(s) = \frac{y_0 + \frac{1}{s}}{1+s} = \frac{1 + sy_0}{s(s+1)}.$$

We notice that $\tilde{y}$ is analytic except at the two poles 0 and -1. To compute the inverse Laplace transform using the formula (A.4), we need to take $u > 0$ in order to have

$$y(t) = \frac{1}{2i\pi} \lim_{A \to +\infty} \int_{u-iA}^{u+iA} \frac{1 + sy_0}{s(s+1)} e^{st}\, ds. \tag{A.5}$$

We shall denote by $g(s) = \frac{1+sy_0}{s(s+1)} e^{st}$ the function we want to integrate and compute the integral using a contour that closes on the left hand side in order to include the two poles so as to be able to apply the residue formula (A.2).

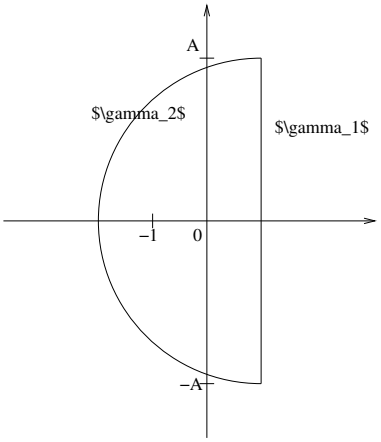

Fig. A.1. Path for the computation of the integral.

We parametrise the line γ_1 of real part u by $\theta \mapsto u + i\theta$, $\theta \in [-A, A]$, and the half-circle γ_2 closing the contour on the left hand side $\theta \mapsto u + Ae^{i\theta}$, $\theta \in [\frac{\pi}{2}, \frac{3\pi}{2}]$, see figure A.1. We assume that $A > u + 1$ so that both poles are inside the half-circle. We then have

$$\int_{\gamma_2} g(s) = \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \frac{1 + (u + Ae^{i\theta})y_0 e^{(u+Ae^{i\theta})t}}{(u + Ae^{i\theta})(u + 1 + Ae^{i\theta})} iAe^{i\theta}\, d\theta.$$

An upper bound of the right hand side is

$$Ce^u \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} e^{At\cos\theta}\, d\theta$$

for A large enough, and this term tends to 0 when A tends to $+\infty$ as $\cos\theta < 0$ for $\theta \in]\frac{\pi}{2}, \frac{3\pi}{2}[$. On the other hand applying the residue theorem, Theorem 7, we have

$$\int_{\gamma_1 \cup \gamma_2} g(s) = 2i\pi \left(\operatorname{Res}_{s=0} g(s) + \operatorname{Res}_{s=-1} g(s) \right) = 2i\pi \left(1 - (1 - y_0)e^{-t} \right).$$

So taking the limit when A tends to $+\infty$ in (A.5), we obtain

$$y(t) = 1 - (1 - y_0)e^{-t}$$

which is the solution of the considered differential equation.

B

Background in probability theory

As the most convenient framework for defining integrals is the Lebegues theory, which starts by defining measurable sets using $\sigma-$algebras, the good framework for abstract probability theory also needs these objects. However, after having defined them to make the connection with the mathematical probability literature, we will only consider probabilities on $\mathbb{R}^n$.

B.1 Probability spaces

Let us recall some standard definitions that can be found in any probability textbook.
Let Ω be a nonempty set.

Definition 12 *A σ-algebra is a collection $\mathcal{F}$ of subsets of Ω with the properties*

(i) $\Omega \in \mathcal{F}$,
(ii)If $A \in \mathcal{F}$ *then* $A^c := \Omega \backslash A \in \mathcal{F}$,
(iii)If $A_1, A_2, \cdots \in \mathcal{F}$ *then* $\bigcup_i A_i \in \mathcal{F}$.

Note that axioms (i) and (ii) imply that $\emptyset \in \mathcal{F}$ and axioms (ii) and (iii) imply that if $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcap_i A_i \in \mathcal{F}$, as $(\bigcup_i A_i^c)^c = \bigcap_i A_i$.

Definition 13 *Let $\mathcal{F}$ be a σ-algebra of subsets of Ω. Then $P : \mathcal{F} \to [0,1]$ is called a probability measure provided:*

(ii)For all $A \in \mathcal{F}$ *we have* $0 \leq P(A) \leq 1$,
(ii)$P(\Omega) = 1$,
(iii)If $A_1, A_2, \cdots \in \mathcal{F}$ *are disjoint then* $P(\bigcup_i A_i) = \sum_i P(A_i)$.

It follows from (ii) and (iii) as Ω and $\emptyset$ are both in $\mathcal{F}$ and disjoint that $P(\emptyset) = 0$. If follows from (i) and (iii) that if $A \subset B$ then B is the disjoint union of A and $B \setminus A$, so $P(A) \leq P(A) + P(B \setminus A) = P(B)$.

Definition 14 *A triple $(\Omega, \mathcal{F}, P)$ is called probability space provided Ω is any set, $\mathcal{F}$ is a σ-algebra and P a probability measure on $\mathcal{F}$.*

Terminology. A set $A \in \mathcal{F}$ is called an *event*, points $\omega \in \Omega$ are called *sample points* and $P(A)$ is the *probability* of event A.

Let $\mathcal{B}$ denote the Borel subsets of $\mathbb{R}^n$ which is the smallest σ-algebra containing all the the open subsets of $\mathbb{R}^n$. In particular it contains all the product intervals (open, semi-open, or closed).

Example 1. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ be a finite set, and suppose we are given N numbers $0 \leq p_i \leq 1$ for $i = 1, \ldots, N$ satisfying $\sum_{i=1}^{N} p_i = 1$. We take $\mathcal{F}$ to be all the possible subsets of Ω. Then for each $A = \{\omega_{i_1}, \ldots \omega_{i_m}\} \in \mathcal{F}$ with $1 \leq \omega_{i_1} < \cdots < \omega_{i_m} \leq N$ we define

$$P(A) := p_{i_1} + p_{i_2} + \ldots p_{i_m}.$$

Let us consider two concrete examples:

1) Throwing once a dice can be analysed with the following probability space: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}$ consists of all the subsets of Ω and $p_i = \frac{1}{6}$ for $i = 1, \ldots, 6$. An event is a subset of Ω, for example $A = \{2, 5\}$. The probability of a sample point to be in A is then $P(A) = p_2 + p_5 = \frac{1}{3}$.

2) Consider throwing a coin twice. Then the set Ω of all possible events is

$$\{(H, H), (H, T), (T, H), (T, T)\}$$

where H stands for heads and T for tail, $\mathcal{F}$ consists of all the subsets of Ω and $p_i = \frac{1}{4}$ for $i = 1, \ldots, 4$. A possible event A would be to throw heads at least once: $A = \{(H, H), (H, T), (T, H)\}$ and $P(A) = \frac{3}{4}$, and other possible event B would be to throw tail the second time, then $B = \{(H, T), (T, T)\}$ and $P(B) = \frac{1}{2}$.

Example 2. The Dirac mass. Let $\mathbf{z} \in \mathbb{R}^n$ fixed and define for sets $A \in \mathcal{B}$

$$P(A) := \begin{cases} 1 \text{ if } \mathbf{z} \in A, \\ 0 \text{ if } \mathbf{z} \notin A. \end{cases}$$

We call P the Dirac mass at $\mathbf{z}$ and denote it by $P = \delta_{\mathbf{z}}$.

Example 3. Assume f is a non negative integrable function such that $\int_{\mathbb{R}^n} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$. We define for sets $A \in \mathcal{B}$

$$P(A) := \int_A f(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

We call f the *density* of the probability measure P.

B.2 Random variables

A probability space is an abstract construction. In order to define observables it is necessary to introduce mappings $\mathbf{X}$ from Ω to $\mathbb{R}^n$.

Definition 15 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A mapping*

$$\mathbf{X} : \Omega \to \mathbb{R}^n$$

is called a n-dimensional random variable if for each $B \in \mathcal{B}$, we have

$$\mathbf{X}^{-1}(B) \in \mathcal{F}.$$

In other words, $\mathbf{X}$ is n-dimensional random variable on the probability space if it is $\mathcal{F}$-measurable.

This definition enables to define probabilities of events related to $\mathbf{X}$ by inducing a probability law on $(\mathbb{R}^n, \mathcal{B})$.

Proposition 24 *Let $\mathbf{X}$ be an n-dimensional random variable. Then $P_{\mathbf{X}} : \mathcal{B} \to [0,1]$ defined by*

$$P_{\mathbf{X}}(B) = P(\mathbf{X}^{-1}(B))$$

is a probability law on $(\mathbb{R}^n, \mathcal{B})$.

Proof. For $B \in \mathcal{B}$, the measurability of $\mathbf{X}$ implies that $\mathbf{X}^{-1}(B) \in \mathcal{B}$. So the probability $P(\mathbf{X}^{-1}(B))$ is well defined and we just need to check the properties of a probability law, which is straightforward.

Notation 1 *The probability $P_{\mathbf{X}}$ is often denoted conveniently $P_{\mathbf{X}}(B) = P(\mathbf{X} \in B)$.*

B.3 Distribution function

Let $\mathbf{X}$ be a n-dimensional random variable on the probability space $(\Omega, \mathcal{F}, P)$. Let us say that for two vectors $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ all the components of the vectors.

Definition 16 *We call (cumulative) distribution function (CDF) of a random variable $\mathbf{X}$ the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ defined by*

$$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}), \qquad \text{for } x \in \mathbb{R}^n.$$

Definition 17 *Assume $\mathbf{X}$ is a n-dimensional random variable and $F = F_{\mathbf{X}}$ its distribution function. If there exists a non negative, integrable function $f : \mathbb{R}^n \to \mathbb{R}$ such that*

$$F(\mathbf{x}) = F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n)\, dy_1 \ldots dy_n,$$

then f is called the (probability) density function (PDF) for $\mathbf{X}$.

It follows, in this case, that all probabilities related to $\mathbf{X}$ can be expressed as integrals on $\mathbb{R}^n$ using the density function:

$$P_{\mathbf{X}}(B) = P(\mathbf{X} \in B) = \int_B f(\mathbf{x})\, d\mathbf{x} \qquad \text{for all } B \in \mathcal{B}.$$

Probability measures for which a density exists are called absolutely continuous. We shall only consider such probability measures in the sequel.

Note that if we work directly in the probability space $(\mathbb{R}^n, \mathcal{B}, P)$, we can take the random variable to be the identity and a probability density directly defines the probability.

Note also that a random variable $\mathbf{X}$ induces a σ-algebra $\mathcal{F}_X$ on Ω, which is defined by $\mathcal{F}_X = \{X^{-1}(B) | B \in \mathcal{B}\}$. $\mathcal{F}_X$ is the smallest σ-algebra which makes $\mathbf{X}$ measurable.

Examples:

1. The uniform distribution on interval $[a, b]$ is given by the PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{else.} \end{cases}$$

The associated distribution function is

$$F(x) = \int_{-\infty}^{x} f(x)\, \mathrm{d}x = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } x \in [a, b], \\ 1 & \text{if } x > b. \end{cases}$$

2. The *normal* or *gaussian* distribution is defined by a PDF of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

B.4 Expected value, variance.

The integration on a probability space is similar to the definition of the Lebesgues integral. One starts by defining the integral for simple functions of the form $X = \sum_i a_i \chi_{A_i}$, where χ_{A_i} is the characteristic functions of the set $A_i \in \mathcal{F}$, *i.e.* $\chi_{A_i}(\omega) = 1$ if $\omega \in A_i$ and 0 else. Then we define

$$\mathbb{E}(X) := \int X \, \mathrm{d}P := \sum_i a_i P(A_i).$$

Then, because any measurable functions is the limit of a sequence of simple functions, this definition can then be extended by taking limits to any random variable (which is a $\mathcal{F}$-measurable function). For vector valued random variables, the integration is performed component by component.

Definition 18 *A random variable* $\mathbf{X}$ *is said integrable with respect to the probability measure* P*, if* $\mathbb{E}(|\mathbf{X}|) < +\infty$*. Then the value* $\mathbb{E}(\mathbf{X}) := \int \mathbf{X}\,\mathrm{d}P$ *is called* expected value *(or* expectation*, or* mean value*) of the random variable* $\mathbf{X}$*.*

If $\mathbb{E}(|\mathbf{X}|^2) < +\infty$*, the value*

$$\mathbb{V}(\mathbf{X}) = \mathbb{E}(|\mathbf{X} - \mathbb{E}(\mathbf{X})|^2) = \int_\Omega |\mathbf{X} - \mathbb{E}(\mathbf{X})|^2\,\mathrm{d}P \geq 0$$

is called variance *of the random variable* $\mathbf{X}$*, and*

$$\sigma(\mathbf{X}) = \sqrt{\mathbb{V}(\mathbf{X})}$$

is called standard deviation *of the random variable* $\mathbf{X}$*.*

The variance can be also expressed by $\mathbb{V}(\mathbf{X}) = \mathbb{E}(|\mathbf{X}|^2) - \mathbb{E}(\mathbf{X})^2$. Indeed

$$\mathbb{V}(\mathbf{X}) = \int_\Omega |\mathbf{X}-\mathbb{E}(\mathbf{X})|^2\,\mathrm{d}P = \int_\Omega (|\mathbf{X}|^2 - 2\mathbf{X}\cdot\mathbb{E}(\mathbf{X}) + |\mathbb{E}(\mathbf{X})|^2)\,\mathrm{d}P = \mathbb{E}(|\mathbf{X}|^2) - \mathbb{E}(\mathbf{X})^2.$$

If the probability measure is absolutely continuous its density provides a convenient way for evaluation of expectations using the so-called transfer theorem.

Theorem 9 (Transfer theorem) *Let* g *be a measurable function of* $\mathbb{R}^n$ *and* $\mathbf{X}$ *an* n*-dimensional random variable. Then, if* f *is the density of the law of* $\mathbf{X}$

$$\mathbb{E}(g(\mathbf{X})) = \int_\Omega g(\mathbf{X})\,dP = \int_{\mathbb{R}^n} g(\mathbf{x})\,\mathrm{d}P_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Formally $\mathrm{d}P_{\mathbf{X}}(\mathbf{x}) = f(\mathbf{x})\,\mathrm{d}\mathbf{x}$. If f depends on $\mathbf{x}$ the probability measure $P_{\mathbf{X}}$ is not translation invariant. This is highlighted by the notation $\mathrm{d}P_{\mathbf{X}}(\mathbf{x})$.

Proof. Let us check the formula for positive simple random variables. The general case is then obtained using the appropriate limit theorems.

So let $g = \sum_{i=1}^n a_i \chi_{A_i}$ be a positive simple function. Then

$$g(\mathbf{X}(\omega)) = \sum_{i=1}^n a_i \chi_{A_i}(\mathbf{X}(\omega)) = \sum_{i=1}^n a_i \chi_{\mathbf{X}^{-1}(A_i)}(\omega).$$

Hence

$$\mathbb{E}(g(\mathbf{X})) = \sum_{i=1}^n a_i P(\mathbf{X}^{-1}(A_i)) = \sum_{i=1}^n a_i P_{\mathbf{X}}(A_i) = \int_{\mathbb{R}^n} g(\mathbf{x})\,\mathrm{d}P_{\mathbf{X}}(\mathbf{x}).$$

The last definition is just the definition of the integral for simple functions. Moreover, if $P_{\mathbf{X}}$ has density f, then by definition of the density

$$\sum_{i=1}^n a_i P_{\mathbf{X}}(A_i) = \sum_{i=1}^n a_i \int_{A_i} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{i=1}^n a_i \int_{\mathbb{R}^n} \chi_{A_i}(\mathbf{x}) f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

The formula given by the transfer theorem will be used for actual computations. In particular for the variance

$$\mathbb{V}(\mathbf{X}) = \int_{\Omega} |\mathbf{X} - \mathbb{E}(\mathbf{X})|^2 \, \mathrm{d}P = \int_{\mathbb{R}^n} (x - \mathbb{E}(\mathbf{X}))^2 f(x) \, \mathrm{d}x.$$

The variance can help quantify the deviation of a random variable $\mathbf{X}$ from its mean:

Proposition 25 (Chebyshev inequality) *Assume* $\mathbb{E}(X^2) < +\infty$. *Then for any* $\epsilon > 0$

$$P\big(|\mathbf{X} - \mathbb{E}(\mathbf{X})| \geq \epsilon\big) \leq \frac{\mathbb{V}(\mathbf{X})}{\epsilon^2}.$$

Proof. Denote by $A = |\mathbf{X} - \mathbb{E}(\mathbf{X})| \geq \epsilon$. Then

$$\mathbb{V}(\mathbf{X}) = \int_{\Omega} |\mathbf{X} - \mathbb{E}(\mathbf{X})|^2 \, \mathrm{d}P \geq \int_{A} |\mathbf{X} - \mathbb{E}(\mathbf{X})|^2 \, \mathrm{d}P \geq \int_{A} \epsilon^2 \, \mathrm{d}P = \epsilon^2 P(A)$$

which gives the result.

B.5 Conditional probabilities and independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let A and B be two events.

Knowing that some random sample point $\omega \in \Omega$ is in A we are interested in obtaining the probability that $\omega \in B$. This defines the conditional probability:

Definition 19 *Let* A *an event of probability* $P(A) > 0$. *Then the probability of* B *given* A *is defined by*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Let us verify that $P(\cdot|A)$ defines a probability measure on $(\Omega, \mathcal{F})$: As $A \cap B \subset A$, $P(A \cap B) \leq P(A)$. Hence $0 \leq P(B|A) \leq 1$ and axiom (i) is verified. $\Omega \cap A = A$ hence $P(\Omega|A) = 1$ and axiom (ii) is verified. If $B_1, B_2, \ldots$ are disjoint, so are their intersections with A and axiom (iii) follows.

Two events A, B are independent if $P(B|A) = P(B)$. Then by definition of the conditional probability $P(B) = \frac{P(A \cap B)}{P(A)}$ and we get the more symmetric definition of the independence of A and B.

Definition 20 *Two events* A *and* B *are said to be independent if*

$$P(A \cap B) = P(A)P(B).$$

This definition extends to random variables:

Definition 21 *We say that the random variables $\mathbf{X}_i : \Omega \to \mathbb{R}^n$, $i = 1, \ldots, m$ are independent, if for all choices of Borel sets $B_1, \ldots, B_m \subseteq \mathbb{R}^n$*

$$P(\mathbf{X}_1 \in B_1, \ldots, \mathbf{X}_m \in B_m) = P(\mathbf{X}_1 \in B_1) \cdots P(\mathbf{X}_m \in B_m).$$

Theorem 10 *The random variables $\mathbf{X}_i : \Omega \to \mathbb{R}^n$, $i = 1, \ldots, m$ are independent, if and only if their distribution functions verify*

$$F_{\mathbf{X}_1,\ldots,\mathbf{X}_m}(\mathbf{x}_1, \ldots, \mathbf{x}_m) = F_{\mathbf{X}_1}(\mathbf{x}_1) \cdots F_{\mathbf{X}_m}(\mathbf{x}_m) \quad \textit{for all } \mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n.$$

If the random variables have densities this is equivalent to

$$f_{\mathbf{X}_1,\ldots,\mathbf{X}_m}(\mathbf{x}_1, \ldots, \mathbf{x}_m) = f_{\mathbf{X}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}_m}(\mathbf{x}_m) \quad \textit{for all } \mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n.$$

The marginal densities $f_{\mathbf{X}_i}$ are obtained from the joined density $f_{\mathbf{X}_1,\ldots,\mathbf{X}_m}$ by integrating on $\mathbb{R}^n$ over all the other variables, for example

$$f_{\mathbf{X}_1}(\mathbf{x}_1) = \int f_{\mathbf{X}_1,\ldots,\mathbf{X}_m}(\mathbf{x}_1, \ldots, \mathbf{x}_m)\, \mathrm{d}\mathbf{x}_2 \ldots\, \mathrm{d}\mathbf{x}_m.$$

From this theorem follows the following important result:

Theorem 11 *If $X_1, \ldots, X_m$ are independent real valued random variables with $\mathbb{E}(|X_i|) < +\infty$ then $\mathbb{E}(|X_1 \ldots X_m|) < +\infty$ and*

$$\mathbb{E}(X_1 \cdots X_m) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_m).$$

Proof. The results is easy to prove by applying the previous theorem assuming that each X_i is bounded and has a density:

$$\begin{aligned}
\mathbb{E}(X_1 \cdots X_m) &= \int_{\mathbb{R}^m} x_1 \cdots x_m f_{X_1,\ldots,X_m}(x_1, \ldots, x_m)\, \mathrm{d}x_1 \ldots\, \mathrm{d}x_m, \\
&= \int_{\mathbb{R}^m} x_1 f_{X_1}(x_1) \cdots x_m f_{X_m}(x_m)\, \mathrm{d}x_1 \ldots\, \mathrm{d}x_m, \\
&= \mathbb{E}(X_1) \cdots \mathbb{E}(X_m).
\end{aligned}$$

Moreover for independent variables the variance of the sum is the sum of variances. This is known as Bienaymé's equality:

Theorem 12 (Bienaymé) *If $X_1, \ldots, X_m$ are independent real valued random variables with $\mathbb{V}(|X_i|) < +\infty$ then*

$$\mathbb{V}(X_1 + \cdots + X_m) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_m).$$

Proof. This can be proved by induction. We prove it only the case of two random variables. Let $m_1 = \mathbb{E}(X_1)$, $m_2 = \mathbb{E}(X_2)$. Then by linearity of the integral $m_1 + m_2 = \mathbb{E}(X_1 + X_2)$ and

$$\begin{aligned}\mathbb{V}(X_1+X_2) &= \int_\Omega (X_1+X_2-(m_1+m_2))^2\,\mathrm{d}P,\\ &= \int_\Omega (X_1-m_1)^2\,\mathrm{d}P + \int_\Omega (X_2-m_2)^2\,\mathrm{d}P + 2\int_\Omega (X_1-m_1)(X_2-m_2)\,\mathrm{d}P,\\ &= \mathbb{V}(X_1)+\mathbb{V}(X_2)+2\mathbb{E}(X_1-m_1)\mathbb{E}(X_2-m_2)\end{aligned}$$

using the independence of the random variables and the previous theorem in the last line. We then get the desired result by noticing that $\mathbb{E}(X_1-m_1) = \mathbb{E}(X_2-m_2) = 0$.

Definition 22 *Let X and Y be two square integrable real valued random variables, then their* covariance *is defined by*

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))).$$

By linearity of the expected value, we easily get

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Looking at the proof of the Bienaymé equality we see that in the general case we have

$$\mathbb{V}(X+Y) = \mathbb{V}(X)+\mathbb{V}(Y)+2\mathrm{Cov}(X,Y),$$

the last term vanishing if the two random variables are independent. A more precise measure of the linear independence of two random variables is given by the correlation coefficient defined by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

B.6 Stochastic processes

Definition 23 *A stochastic process is a collection $\{\mathbf{X}_t | t \geq 0\}$ of random variables. For each $\omega \in \Omega$ the mapping $t \mapsto \mathbf{X}_t(\omega)$ is called sample path.*

The sample path would be the trajectory of a single particle indexed by the continuous index ω.

A natural filtration $\mathcal{F}_X(t)$ of a stochastic process consisting of the σ-algebras induced by all the past values $\mathbf{X}(s)$, $s \leq t$ of the process can be defined. This is call the history of the process until t. If $\mathcal{F}(t)$ is a filtration containing $\mathcal{F}_X(t)$, we say that $\mathbf{X}(t)$ is progressively measurable with respect to $\mathcal{F}(t)$. *A filtration is a collection of σ-algebras, such as needed by for a stochastic process*

B.7 The Itô integral

Let us denote classically by $\mathcal{N}(m,\sigma)$ a normal random variable with mean m and variance σ^2 whose probability density function is $\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/(2\sigma^2)}$.

Definition 24 *A real-valued stochastic process is called a Brownian motion or Wiener Process if the following properties are verified:*

(i) $W(0)=0$ *almost surely,*
(ii) $W(t)-W(s)$ *is* $\mathcal{N}(0,t-s)$ *for all* $t\geq s\geq 0$*,*
(iii) For all times $0<t_1<t_2<\cdots<t_n$*, the random variables* $W(t_1)=W(t_1)-W(0), W(t_2)-W(t_1),\ldots,W(t_n)-W(t_{n-1})$ *are independent.*

Definition 25 *A process* $G\in L^2(0,T)$ *is called a step process if there exists a partition* $P=\{0=t_0<t_1<\cdots<t_n=T\}$ *such that*

$$G(t)=G_k \qquad \text{for } t_k\leq t<t_{k+1} \quad (k=0,\ldots n-1).$$

Definition 26 *Let* $G\in L^2(0,T)$ *be a step process. Then the Itô integral of* G *is defined by*

$$\int_0^T G\,\mathrm{d}W=\sum_{k=0}^{m-1}G_k(W(t_{k+1})-W(t_k)).$$

As general processes $G\in L^2(0,T)$ can be approximated by step processes. The Itô integral of such a process is defined as the limit of the Itô integral of the approximating step processes.

From the definition of the Itô integral, the following properties can be easily derived, for step processes first and then by passing to the limit for general processes:

Theorem 13 (Properties of the Itô integral) *Let* λ,μ *be constants and* $G,H\in L^2(0,T)$*. Then we have*

(i) $\int_0^T(\lambda G+\mu H)\,\mathrm{d}W=\lambda\int_0^T G\,\mathrm{d}W+\mu\int_0^T H\,\mathrm{d}W$,
(ii) $\mathbb{E}\big(\int_0^T G\,\mathrm{d}W\big)=0$,
(iii) $\mathbb{E}\big((\int_0^T G\,\mathrm{d}W)^2\big)=\mathbb{E}\big(\int_0^T G^2\,\mathrm{d}t\big)$,

B.8 Stochastic differential equations

The Itô integral can now be used to defined stochastic differential equations in the following way:

Definition 27 *We say that an* $\mathbb{R}^d$ *valued stochastic process* $\mathbf{X}(t)$ *is solution of the Itô stochastic differential equation*

$$d\mathbf{X}(t) = \mathbf{A}(t, \mathbf{X}(t))\,\mathrm{d}t + \nu(t, \mathbf{X}(t))\,\mathrm{d}\mathbf{W}(t), \tag{B.1}$$

$$\mathbf{X}(0) = \mathbf{x}, \tag{B.2}$$

provided $X(t)$ *is progressively measurable with respect to* $\mathcal{F}(t)$, $\mathbf{F} = \mathbf{A}(t, \mathbf{X}(t)) \in L^1(0,T)$, $G = \nu(t, \mathbf{X}(t)) \in L^2(0,T)$ *and*

$$\mathbf{X}(t) = \mathbf{x} + \int_0^t \mathbf{A}(s, \mathbf{X}(s))\,\mathrm{d}s + \int_0^t \nu(s, \mathbf{X}(s))\,\mathrm{d}\mathbf{W}(s) \text{ almost surely.}$$

Stochastic differential equations using the Itô formalism provide a natural extension to ordinary differential equations. Moreover they enable to extend the usual notion of characteristics for transport equations to equations with diffusion. Indeed, let $\mathbf{X}(t)$ be a solution of the stochastic differential equation (B.1)–(B.2). Taking the derivative of a smooth function along the solutions of this SDE can be done analogously to taking the derivative along the characteristics by using the the following rule

$$du(t, \mathbf{X}(t)) = \frac{\partial u}{\partial t}(t, \mathbf{X}(t))\,\mathrm{d}t + \nabla u(t, \mathbf{X}(t)) \cdot d\mathbf{X}(t) + \frac{1}{2}\nu(t, \mathbf{X}(t))^2 \Delta u(t, \mathbf{X}(t))\,\mathrm{d}t.$$

Notice the additional second order term coming from the Wiener process in comparison to the classical formula. Separating the terms in dt and $d\mathbf{W}(t)$ yields the **Itô formula**

$$du(t, \mathbf{X}(t)) = (\frac{\partial u}{\partial t} + \mathbf{A} \cdot \nabla u + \frac{1}{2}\nu^2 \Delta u)(t, \mathbf{X}(t))\,\mathrm{d}t + \nu(t, \mathbf{X}(t)) \nabla u(t, \mathbf{X}(t)) \cdot d\mathbf{W}. \tag{B.3}$$

B.9 The Kolmogorov forward and backward equations

Let u be a smooth function such that $\mathbf{G}(t) = \nu(t, \mathbf{X}(t)) \nabla u(t, \mathbf{X}(t)) \in L^2(0,T)$. A property of the Itô integral given previously implies that

$$\mathbb{E}(\int_0^T \mathbf{G} \cdot d\mathbf{W}) = \mathbb{E}(\int_0^T \nu(t, \mathbf{X}(t)) \nabla u(t, \mathbf{X}(t)) \cdot d\mathbf{W}) = 0.$$

Then it follows from the Itô formula, integrating between s and t commuting the expectation with the time integration, that

$$\mathbb{E}(u(t, \mathbf{X}(t))) = \mathbb{E}(u(s, \mathbf{X}(s))) + \int_s^t \mathbb{E}(\frac{\partial u}{\partial t} + \mathbf{A} \cdot \nabla u + \frac{1}{2}\nu^2 \Delta u)(\sigma, \mathbf{X}(\sigma))\,\mathrm{d}\sigma. \tag{B.4}$$

In particular if $\mathbf{X}(s) = \mathbf{x}$ is known, and not a random variable, and if u is a solution of the partial differential equation

$$\frac{\partial u}{\partial t} + \mathbf{A} \cdot \nabla u + \frac{1}{2}\nu^2 \Delta u = 0,$$

formula (B.4) provides an expression for u as a function of its value at a final time T

$$u(s, \mathbf{x}) = \mathbb{E}(u(T, \mathbf{X}(T))).$$

This expression is known as the Kolmogorov backward equation. It is in particular useful for some finance modelling problem where a given value needs to be reached.

On the other hand for physics problems where in general one is interested in the evolution of the system from an initial condition, formula (B.4) also provides an evolution equation for the probability density of the random variable $\mathbf{X}(t)$. Indeed, denote by $f(t, \mathbf{x})$ the probability density of the random variable $\mathbf{X}(t)$. Then for any smooth function ψ the expectation $\mathbb{E}(\psi(\mathbf{X}(t)))$ is defined by

$$\mathbb{E}(\psi(\mathbf{X}(t))) = \int \psi(\mathbf{x}) f(t, \mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Then using formula (B.4) for $u(t, \mathbf{x}) = \psi(\mathbf{x})$ yields

$$\int \psi(\mathbf{x}) f(t, \mathbf{x}) \, \mathrm{d}\mathbf{x} = \int \psi(\mathbf{x}) f(s, \mathbf{x}) \, \mathrm{d}\mathbf{x} + \int_s^t \int (\mathbf{A} \cdot \nabla \psi + \frac{1}{2} \nu^2 \Delta \psi)(\mathbf{x}) f(\sigma, \mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\sigma.$$

As ψ is an arbitrary test function, taking the limit $s \to t$ shows that f is a weak solution of the equation

$$\frac{\partial f}{\partial t} + \nabla \cdot (\mathbf{A} f) - \frac{1}{2} \Delta (\nu^2 f) = 0. \tag{B.5}$$

Indeed if f is smooth:

$$\int (\mathbf{A} \cdot \nabla \psi + \frac{1}{2} \nu^2 \Delta \psi) f(t, \mathbf{x}) \, \mathrm{d}\mathbf{x} = \int (-\nabla \cdot (\mathbf{A} f) + \frac{1}{2} \Delta (\nu^2 f)) \psi \, \mathrm{d}\mathbf{x}.$$

Equation (B.5) is called the Kolmogorov forward equation, and in physics the Fokker-Planck equation. This justifies that the solution of the stochastic differential equation can be used to get a solution of the Fokker-Planck equation we are interested in.

References

1. Lars V Ahlfors. *Complex analysis: an introduction to the theory of analytic functions of one complex variable*. McGraw-Hill, 1979.
2. Simon J Allfrey and Roman Hatzky. A revised δf algorithm for nonlinear PIC simulation. *Computer physics communications*, 154(2):98–104, 2003.
3. Herbert Amann. *Ordinary differential equations: an introduction to nonlinear analysis*, volume 13. Walter de Gruyter, 1990.
4. Douglas Arnold, Richard Falk, and Ragnar Winther. Finite element exterior calculus: from hodge theory to numerical stability. *Bulletin of the American mathematical society*, 47(2):281–354, 2010.
5. Douglas N Arnold, Richard S Falk, and Ragnar Winther. Finite element exterior calculus, homological techniques, and applications. *Acta numerica*, 15:1–155, 2006.
6. Ahmet Y. Aydemir. A unified monte carlo interpretation of particle simulations and applications to non-neutral plasmas. *Physics of Plasmas*, 1(4):822–831, 1994.
7. R. Barthelmé, P. Ciarlet Jr, and E. Sonnendrücker. Generalized formulations of Maxwell's equations for numerical Vlasov–Maxwell simulations. *Mathematical Models and Methods in Applied Sciences*, 17(05):657–680, 2007.
8. R. Barthelmé and C. Parzani. Numerical charge conservation in particle-in-cell codes. *Numerical Methods for Hyperbolic and Kinetic Problems*, pages 7–28, 2005.
9. William B. Bateson and Dennis W. Hewett. Grid and particle hydrodynamics: beyond hydrodynamics via fluid element particle-in-cell. *J. Comput. Phys.*, 144(2):358–378, 1998.
10. J.-P. Beirut and L. N. Trefethen. Barycentric lagrange interpolation. *SIAM Review*, 46(3):501–517, 2004.
11. Richard Bellman and Robert S Roth. *The Laplace Transform*, volume 3. World Scientific, 1984.
12. N. Besse. Convergence of a semi-lagrangian scheme for the one-dimensional Vlasov-Poisson system. *SIAM J. Numer. Anal.*, 42(1):350–382, 2004.
13. N. Besse and M. Mehrenberger. Convergence of classes of high-order semi-lagrangian schemes for the Vlasov-Poisson system. *Math. Comp.*, 77:93–123, 2008.

14. Nicolas Besse. Convergence of a high-order semi-lagrangian scheme with propagation of gradients for the one-dimensional vlasov-poisson system. *SIAM Journal on Numerical Analysis*, 46(2):639–670, 2008.
15. Nicolas Besse and Eric Sonnendrücker. Semi-lagrangian schemes for the vlasov equation on an unstructured mesh of phase space. *Journal of Computational Physics*, 191(2):341–376, 2003.
16. Prabhu Lal Bhatnagar, Eugene P Gross, and Max Krook. A model for collision processes in gases. i. small amplitude processes in charged and neutral one-component systems. *Physical review*, 94(3):511, 1954.
17. Charles K Birdsall and A Bruce Langdon. *Plasma physics via computer simulation*. CRC Press, 2004.
18. D. Boffi. Compatible Discretizations for Eigenvalue Problems. In *Compatible Spatial Discretizations*, pages 121–142. Springer New York, New York, NY, 2006.
19. D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010.
20. D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, 2013.
21. J.P. Boris. Relativistic plasma simulations - optimization of a hybrid code. In *Proc. 4th Conf. Num. Sim. of Plasmas, (NRL Washington, Washington DC)*, pages 3–67, 1970.
22. A. Buffa and I. Perugia. Discontinuous Galerkin Approximation of the Maxwell Eigenproblem. *SIAM Journal on Numerical Analysis*, 44(5):2198–2226, January 2006.
23. M. Campos Pinto, , M. Mounier, and E. Sonnendrücker. Handling the divergence constraints in maxwell and vlasov-maxwell simulations. *Appl. Math. Comput*, 2015.
24. M. Campos Pinto, S. Jund, S. Salmon, and E. Sonnendrücker. Charge conserving fem-pic schemes on general grids. *C.R. Mecanique*, 342(10-11):570–582, 2014.
25. Martin Campos Pinto. Towards smooth particle methods without smoothing. *Journal of Scientific Computing*, pages 1–29, 2014.
26. Martin Campos Pinto and Michel Mehrenberger. Convergence of an adaptive semi-lagrangian scheme for the vlasov-poisson system. *Numerische Mathematik*, 108(3):407–444, 2008.
27. Martin Campos Pinto, Eric Sonnendrücker, Alex Friedman, David P Grote, and Steve M. Lund. Noiseless vlasov-poisson simulations with linearly transformed particles. *J. Comput. Phys.*, 275:236–256, 2014.
28. Claudio Canuto, M Hussaini, A Quarteroni, and TAJ Zang. *Spectral Methods in Fluid Dynamics (Scientific Computation)*. Springer-Verlag, New York-Heidelberg-Berlin, 1987.
29. S. Caorsi, P. Fernandes, and M. Raffetto. On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM Journal on Numerical Analysis*, 38(2):580–607 (electronic), 2000.
30. J.A. Carrillo and F. Vecil. Nonoscillatory interpolation methods applied to Vlasov-based models. *SIAM J. Sci. Comput.*, 29(3):1179–1206, 2007.
31. Carlo Cercignani. *The Boltzmann equation*. Springer, 1988.
32. Francis F Chen. Introduction to plasma physics and controlled fusion volume 1: Plasma physics. *Physics Today*, 38:87, 1985.

33. Chio-Zong Cheng and Georg Knorr. The integration of the vlasov equation in configuration space. *Journal of Computational Physics*, 22(3):330–351, 1976.
34. Charles K Chui. *An introduction to wavelets*, volume 1. Academic press, 1992.
35. Philippe G Ciarlet. *The finite element method for elliptic problems*. Elsevier, 1978.
36. G. Cohen and P. Monk. Gauss point mass lumping schemes for Maxwell's equations. *Numerical Methods for Partial Differential Equations*, 14(1):63–88, 1998.
37. Gary Cohen. *Higher-order numerical methods for transient wave equations*. Springer, 2002.
38. Gary Cohen and Peter Monk. Efficient edge finite element schemes in computational electromagnetism. In *Proceedings of the Third International Conference on Mathematical and Numerical Aspects of Wave Propagation Phenomena, SIAM, Philadelphia*, 1995.
39. CJ Cotter, J Frank, and S Reich. The remapped particle-mesh semi-lagrangian advection scheme. *Quarterly Journal of the Royal Meteorological Society*, 133(622):251–260, 2007.
40. G.-H. Cottet and P.-A. Raviart. Particle methods for the one-dimensional Vlasov-poisson equations. *SIAM J. Numer. Anal.*, 21(1):52–76, 1984.
41. Georges-Henri Cottet and Petros D Koumoutsakos. *Vortex methods: theory and practice*. Cambridge university press, 2000.
42. N. Crouseilles, M. Mehrenberger, and E. Sonnendrücker. Conservative semi-lagragian methods for Vlasov-type equations. *J. Comput. Phys.*, 229:1927–1953, 2010.
43. N. Crouseilles, P. Navaro, and E. Sonnendrücker. Charge conserving grid based methods for the Vlasov-Maxwell equations. *C. R. Mecanique*, 342(10-11):636–646, 2014.
44. N. Crouseilles, Th. Respaud, and E. Sonnendrücker. A forward semi-lagrangian scheme for the numerical solution of the Vlasov equation. *Comput. Phys. Comm.*, 180:1730–1745, 2009.
45. Ronald C Davidson and Hong Qin. *Physics of intense charged particle beams in high energy accelerators*. World Scientific, 2001.
46. Brian Davies. Locating the zeros of an analytic function. *Journal of computational physics*, 66(1):36–49, 1986.
47. A. Dedner, F. Kemm, D. Kröner, C.-D. Munz, T. Schnitzer, and Wesenberg M. Hyperbolic divergence cleaning for the MHD equations. *J. Comput. Phys.*, 175:645–673, 2002.
48. P. Degond and P.-A. Raviart. On the paraxial approximation of the stationary Vlasov-maxwell system. *Math. Models Meth. Appl. Sciences*, 3:513–562, 1993.
49. P Degond and PA Raviart. On the paraxial approximation of the stationary vlasov-maxwell system. *Mathematical Models and Methods in Applied Sciences*, 3(04):513–562, 1993.
50. LM Delves and JN Lyness. A numerical method for locating the zeros of an analytic function. *Mathematics of computation*, pages 543–560, 1967.
51. Jacques Denavit. Numerical simulation of plasmas with periodic smoothing in phase space. *Journal of Computational Physics*, 9(1):75–98, 1972.
52. S. Depeyre and D. Issautier. A new constrained formulation of the Maxwell system. *Rairo-Mathematical Modelling and Numerical Analysis-Modelisation Mathematique Et Analyse Numerique*, 31(3):327–357, 1997.

53. William L Dunn and J Kenneth Shultis. *Exploring Monte Carlo Methods*. Elsevier, 2011.
54. J.W. Eastwood. The virtual particle electromagnetic particle-mesh method. *Computer Physics Communications*, 64(2):252–266, 1991.
55. J.W. Eastwood, W. Arter, NJ Brealey, and RW Hockney. Body-fitted electromagnetic PIC software for use on parallel computers. *Computer Physics Communications*, 87(1):155–178, 1995.
56. Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*. Springer-Verlag, New York, 2004.
57. T.Zh. Esirkepov. Exact charge conservation scheme for particle-in-cell simulation with an arbitrary form-factor. *Computer Physics Communications*, 135(2):144–153, 2001.
58. F. Filbet and E. Sonnendrücker. Comparison of eulerian Vlasov solvers. *Comput. Phys. Comm.*, 151:247–266, 2003.
59. F. Filbet, E. Sonnendrücker, and P. Bertrand. Conservative numerical schemes for the Vlasov equation. *J. Comput. Phys.*, 172:166–187, 2001.
60. Francis Filbet. Convergence of a finite volume scheme for the vlasov–poisson system. *SIAM Journal on Numerical Analysis*, 39(4):1146–1169, 2001.
61. Francis Filbet and Eric Sonnendrücker. Modeling and numerical simulation of space charge dominated beams in the paraxial approximation. *Mathematical Models and Methods in Applied Sciences*, 16(05):763–791, 2006.
62. Burton D Fried and Samuel Daniel Conte. The plasma dispersion function. *The Plasma Dispersion Function, New York: Academic Press, 1961*, 1, 1961.
63. X Garbet, Y Idomura, L Villard, and TH Watanabe. Gyrokinetic simulations of turbulent transport. *Nuclear Fusion*, 50(4), 2010.
64. Marc Gerritsma. Edge functions for spectral element methods. In *Spectral and High Order Methods for Partial Differential Equations*, pages 199–207. Springer, 2011.
65. V. Grandgirard and Y. Sarazin. Gyrokinetic simulations of magnetic fusion plasmas. In *Numerical models for fusion*, volume 39–40 of *Panoramas et synthèses*. Société Mathématique de France, Paris, 2013.
66. Virginie Grandgirard, Maura Brunetti, Pierre Bertrand, Nicolas Besse, Xavier Garbet, Philippe Ghendrih, Giovanni Manfredi, Yanick Sarazin, Olivier Sauter, Eric Sonnendrücker, et al. A drift-kinetic semi-lagrangian 4d code for ion turbulence simulation. *Journal of Computational Physics*, 217(2):395–423, 2006.
67. Martin Greenwald. Verification and validation for magnetic fusiona). *Physics of Plasmas (1994-present)*, 17(5):058101, 2010.
68. J.S. Hesthaven and T. Warburton. High-order nodal discontinuous Galerkin methods for the Maxwell eigenvalue problem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 362(1816):493–524, March 2004.
69. Ralf Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, 11:237–339, 2002.
70. Roger W Hockney and James W Eastwood. *Computer simulation using particles*. CRC Press, 2010.
71. F. Huot, A. Ghizzo, P. Bertrand, E. Sonnendrcker, and O. Coulaud. Instability of the time splitting scheme for the one-dimensional and relativistic Vlasov-maxwell system. *J. Comput. Phys.*, 185(2):512–531, 2003.

72. Didier Issautier, Frédéric Poupaud, Jean-Pierre Cioni, and Loula Fezoui. A 2-D Vlasov-Maxwell solver on unstructured meshes. In *Third international conference on mathematical and numerical aspects of wave propagation*, pages 355–371, 1995.
73. Ralf Kleiber, Roman Hatzky, Axel Könies, Karla Kauffmann, and Per Helander. An improved control-variate scheme for particle-in-cell simulations with collisions. *Comput. Phy. Comm.*, 182:1005–1012, 2011.
74. Peter Kravanja, Marc Van Barel, O Ragos, MN Vrahatis, and FA Zafiropoulos. Zeal: A mathematical software package for computing zeros of analytic functions. *Computer physics communications*, 124(2):212–232, 2000.
75. LD Landau. The transport equation in the case of coulomb interactions. *Collected Papers of LD Landau, Pergamon press, Oxford*, pages 163–170, 1981.
76. A.B. Langdon. On enforcing Gauss' law in electromagnetic particle-in-cell codes. *Comput. Phys. Comm.*, 70:447–450, 1992.
77. R. Leis. *Initial boundary value problems in mathematical physics*. John Wiley & Sons Ltd, 1986.
78. Andrew Lenard and Ira B Bernstein. Plasma oscillations with diffusion in velocity space. *Physical Review*, 112(5):1456, 1958.
79. P.J. Mardahl and J.P. Verboncoeur. Charge conservation in electromagnetic pic codes; spectral comparison of boris/dadi and langdon-marder methods. *Computer physics communications*, 106(3):219–229, 1997.
80. B. Marder. A method for incorporating Gauss's law into electromagnetic PIC codes. *J. Comput. Phys.*, 68:48–55, 1987.
81. P. Monk. An analysis of Nédélec's method for the spatial discretization of Maxwell's equations. *Journal of Computational and Applied Mathematics*, 47(1):101–121, 1993.
82. C-D Munz, P Omnes, R Schneider, E Sonnendrücker, and U Voss. Divergence correction techniques for maxwell solvers based on a hyperbolic model. *Journal of Computational Physics*, 161(2):484–511, 2000.
83. Claus-Dieter Munz, Rudolf Schneider, Eric Sonnendrücker, and Ursula Voss. Maxwell's equations when the charge conservation is not satisfied. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(5):431–436, 1999.
84. Ramachandran D Nair, Jeffrey S Scroggs, and Frederick HM Semazzi. A forward-trajectory global semi-lagrangian transport scheme. *Journal of Computational Physics*, 190(1):275–294, 2003.
85. Takashi Nakamura and Takashi Yabe. Cubic interpolated propagation scheme for solving the hyper-dimensional Vlasov–Poisson equation in phase space. *Computer Physics Communications*, 120(23):122–154, August 1999.
86. Jean-Claude Nédélec. Mixed finite elements in in $\mathbb{R}^3$. *Numerische Mathematik*, 35(3):315–341, 1980.
87. Jean-Claude Nédélec. A new family of mixed finite elements in $\mathbb{R}^3$. *Numerische Mathematik*, 50(1):57–81, 1986.
88. H Neunzert and J Wick. The convergence of simulation methods in plasma physics. *Mathematical methods of plasmaphysics (Oberwolfach, 1979)*, 20:271–286, 1980.
89. J.M. Qiu and A. Christlieb. A conservative high order semi-lagrangian weno method for the Vlasov equation. *J. Comput. Phys.*, 229(4):1130–1149, 2010.

90. J.M. Qiu and C.W. Shu. Conservative semi-lagrangian finite difference weno formulations with applications to the Vlasov equation. *Commun. Comput. Phys.*, 10:979–1000, 2011.
91. Sebastian Reich. An explicit and conservative remapping strategy for semi-lagrangian advection. *Atmospheric Science Letters*, 8(2):58–63, 2007.
92. Marshall N Rosenbluth, William M MacDonald, and David L Judd. Fokker-planck equation for an inverse-square force. *Physical Review*, 107(1):1, 1957.
93. Laurent Schwartz. Mathematics for the physical sciences. *New York*, 1966.
94. Magdi M Shoucri and Real RJ Gagné. Splitting schemes for the numerical solution of a two-dimensional vlasov equation. *Journal of Computational Physics*, 27(3):315–322, 1978.
95. Chi-Wang Shu. High order weighted essentially nonoscillatory schemes for convection dominated problems. *SIAM Review*, 51(1):82–126, 1.
96. Eric Sonnendrücker, Jean Roche, Pierre Bertrand, and Alain Ghizzo. The semi-lagrangian method for the numerical resolution of the vlasov equation. *Journal of Computational Physics*, 149(2):201–220, 1999.
97. Andrew Staniforth and Jean Côté. Semi-lagrangian integration schemes for atmospheric models – a review. *Monthly Weather Review*, 119(9):2206–2223, 1991.
98. Thomas Howard Stix. The theory of plasma waves. *The Theory of Plasma Waves, New York: McGraw-Hill, 1962*, 1, 1962.
99. A. Stock, J. Neudorfer, R. Schneider, C. Altmann, and C.-D. Munz. Investigation of the Purely Hyperbolic Maxwell System for Divergence Cleaning in Discontinuous Galerkin based Particle-In-Cell Methods. In *COUPLED PROBLEMS 2011 IV International Conference on Computational Methods for Coupled Problems in Science and Engineering*, 2011.
100. Zlatko Udovičić. Calculation of the moments of the cardinal b-spline. *Sarajevo journal of mathematics*, 5(18), 2009.
101. T. Umeda, Y. Omura, T. Tominaga, and H. Matsumoto. A new charge conservation method in electromagnetic particle-in-cell simulations. *Computer Physics Communications*, 156(1):73–85, 2003.
102. Michael Unser, Akram Aldroubi, and Murray Eden. Fast b-spline transforms for continuous image representation and interpolation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):277–285, 1991.
103. H. D. Victory Jr., , and Edward J. Allen. The convergence theory of particle-in-cell methods for multidimensional Vlasov-poisson systems. *SIAM J. Numer. Anal.*, 28(5):1207–1241, 1991.
104. J. Villasenor and O. Buneman. Rigorous charge conservation for local electromagnetic field solvers. *Comput. Phys. Commun.*, 69:306–316, 1992.
105. S. Wollman. On the approximation of the Vlasov-Poisson system by particle methodes. *SIAM J. Numer. Anal.*, 37(4):1369–1398, 4.
106. K.S. Yee. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE Trans. Antennas Propag.*, 14(3):302–307, May 1966.
107. H. Yoshida. Construction of higher order symplectic integrators. *Phys. Lett. A*, 150:262, 1990.

Index

B-splines, 71

circulant matrix, 68

discrete fourier transform (DFT), 66

fast Fourier transform (FFT), 67

operator splitting, 63

splines, 70
splitting, 63

time splitting, 63

validation, 49
verification, 49