Numerical methods for the Vlasov equations

Eric Sonnendrücker
Max-Planck-Institut für Plasmaphysik
Technische Universität München

Lecture notes
Wintersemester 2012/2013

version of January 14, 2013

Contents

CHAPTER 1

Introduction

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`.

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 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. 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.

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

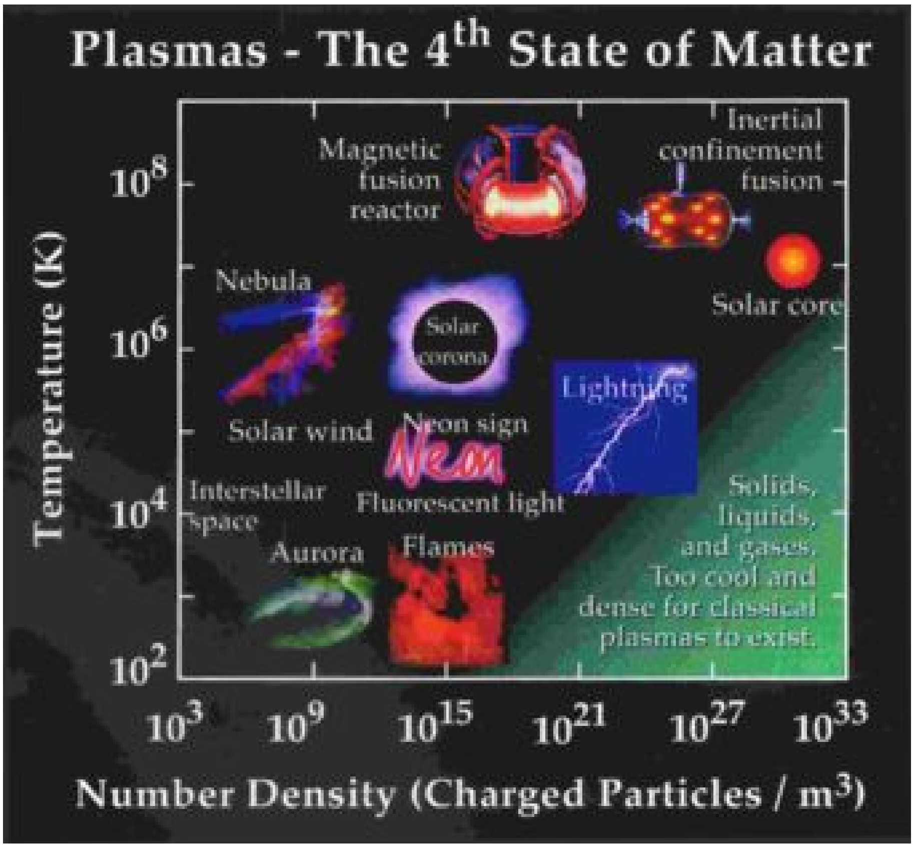

FIGURE 1. Examples of plasmas at different densities and temperatures

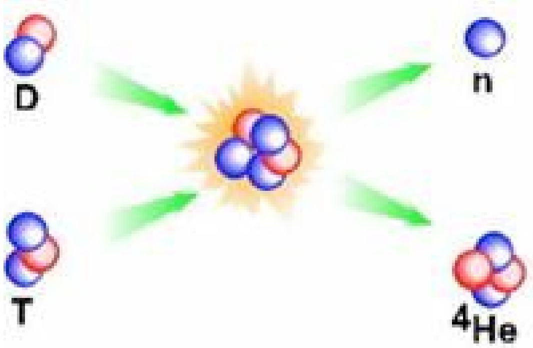

FIGURE 2. The Deuterium-Tritium fusion reaction

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 3.

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.

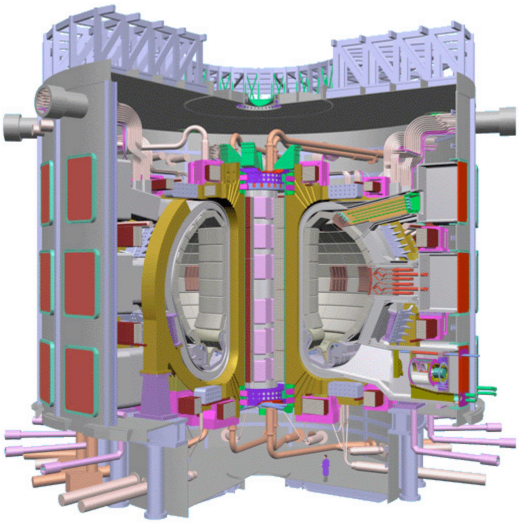

FIGURE 3. Artist view of the ITER Tokamak

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.

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`.

4. 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 non-linear 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.

CHAPTER 2

A hierarchy of models for plasmas

1. 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

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{v}_i, \tag{1}$$

$$\frac{d\gamma_i m\mathbf{v}_i}{dt} = \sum_j q(\mathbf{E}_j + \mathbf{v} \times \mathbf{B}_j), \tag{2}$$

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

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

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. 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. 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

$$\frac{\partial f_s}{\partial t}+\mathbf{v}\cdot\nabla_{\mathbf{x}}f_s+\frac{q_s}{m_s}(\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot\nabla_{\mathbf{v}}f_s=0, \tag{3}$$

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

$$\frac{\partial f_s}{\partial t}+\mathbf{v}(\mathbf{p})\cdot\nabla_{\mathbf{x}}f_s+q_s(\mathbf{E}+\mathbf{v}(\mathbf{p})\times\mathbf{B})\cdot\nabla_{\mathbf{p}}f_s=0. \tag{4}$$

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 $\gamma_s=\sqrt{1+\frac{|\mathbf{p}|^2}{m_s^2c^2}}$.

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

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

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

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

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

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

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

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

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

with $\mu = \frac{m m_s}{m + m_s}$ 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 [**12**] 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*

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

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

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

Then

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

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

3. 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:

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{x}} + \frac{q}{m}(\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot\frac{\partial f}{\partial \mathbf{v}} = \mathcal{Q}(f,f). \tag{5}$$

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 $(a_ib_j)_{1\leq i,j\leq 3}$.

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 (5) 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
$$\frac{\partial n}{\partial t} + \nabla_x\cdot(n\mathbf{u}) = 0. \tag{6}$$

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

$$m\frac{\partial}{\partial t}(n\mathbf{u}) + m\nabla\cdot(n\mathbf{u}\otimes\mathbf{u}) + \nabla\cdot\mathbb{P} = qn(\mathbf{E}+\mathbf{u}\times\mathbf{B}). \tag{7}$$

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

$$\frac{\partial}{\partial t}(\frac{3}{2}p+\frac{1}{2}mn|\mathbf{u}|^2) + \nabla\cdot\,\mathbf{Q} = \mathbf{E}\cdot(qn\mathbf{u}). \tag{8}$$

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 $\mathbf{w}$,

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

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 (6)-(7)-(8) we obtain the fluid equations for one species of particles of a plasma:

$$\frac{\partial n}{\partial t} + \nabla_x \cdot (n\mathbf{u}) = 0 \tag{9}$$

$$m\frac{\partial}{\partial t}(n\mathbf{u}) + m\nabla\cdot(n\mathbf{u}\otimes\mathbf{u}) + \nabla p = qn(\mathbf{E} + \mathbf{u}\times\mathbf{B}) \tag{10}$$

$$\frac{\partial}{\partial t}(\frac{3}{2}p + \frac{1}{2}mn|\mathbf{u}|^2) + \nabla\cdot\,(\frac{5}{2}p\mathbf{u} + \frac{m}{2}n|\mathbf{u}|^2\mathbf{u}) = \mathbf{E}\cdot(qn\mathbf{u}), \tag{11}$$

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.

CHAPTER 3

Some theory on Vlasov systems

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

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + (\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot\nabla_v f = 0, \tag{12}$$

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 (12) becomes

$$\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla_{(x,v)} f = 0. \tag{13}$$

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

$$\begin{aligned}\nabla_{(x,v)}\cdot\mathbf{A} &= \nabla_x\cdot\mathbf{v} + \nabla_v\cdot(\mathbf{E}+\mathbf{v}\times\mathbf{B}) \\ &= 0 + \frac{\partial}{\partial v_1}(E_1+v_2B_3-v_3B_2) + \frac{\partial}{\partial v_2}(E_2+v_3B_1-v_1B_3) \\ &\qquad + \frac{\partial}{\partial v_2}(E_3+v_1B_2-v_2B_1) \\ &= 0.\end{aligned}$$

The Vlasov equation can be written in a conservative form

$$\frac{\partial f}{\partial t} + \nabla_{(x,v)}\cdot(\mathbf{A}f) = 0, \tag{14}$$

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

$$\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla f = 0, \tag{15}$$

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

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}(\mathbf{X},t), \tag{16}$$

$$\mathbf{X}(s) = \mathbf{x}, \tag{17}$$

which is naturally associated to the advection equation (15).

DEFINITION 1. *The solutions of the system* (16) *are called characteristics of the linear advection equation* (15)*. We denote by* $\mathbf{X}(t;s,\mathbf{x})$ *the solution of* (16) - (17)*.*

Let us recall the classical theorem of the theory of ordinary differential equations (ODE) which gives existence and uniqueness of the solution of (16)-(17). 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* (16) - (17)*.*

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 (16)-(17). 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

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

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 $(.,\dots,.)$ this alternated d-linear form, we can write $\det M = (M_1, \dots, M_d)$ where M_j is the jth column of

M. Using this notation in our case, we get

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

as $(.,\dots,.)$ 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 (15) 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 (15) associated to the initial condition $f(\mathbf{x},0) = f_0(\mathbf{x})$. It is given by*

$$f(\mathbf{x},t) = f_0(\mathbf{X}(0;t,\mathbf{x})), \tag{18}$$

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

PROOF. The function f given by (18) 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 (15) and that it verifies the initial condition. We first have using formula (18)

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

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

$$(19)\quad (\frac{\partial f}{\partial t} + \mathbf{A}\cdot\nabla_x f)(\mathbf{x},t) = \frac{\partial X}{\partial s}(0;t,\mathbf{x})\cdot\nabla f_0(0;t,\mathbf{x}) \\ + \mathbf{A}(\mathbf{x},t)\cdot\left(\nabla_x\mathbf{X}(0;t,\mathbf{x})^T\nabla_x f_0(\mathbf{X}(0;t,\mathbf{x}))\right).$$

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 (19) we obtain

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

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 (18) is solution of (15).

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 (18). □

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

$$f(x,v,t) = f_0(x\cos t - v\sin t, x\sin t + v\cos t).$$

2. The Vlasov-Poisson system

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

$$\nabla\times\mathbf{B} = \mathbf{J}, \tag{20}$$

$$\nabla\times\mathbf{E} = 0, \tag{21}$$

$$\nabla\cdot\mathbf{E} = \rho, \tag{22}$$

$$\nabla\cdot\mathbf{B} = 0. \tag{23}$$

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 (21) and (22). Equation (21) implies that $\mathbf{E}$ derives from a scalar potential $\mathbf{E} = -\nabla\phi$, and then (22) 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

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f - \mathbf{E}\cdot\nabla_v f = 0, \tag{24}$$

$$-\Delta\phi = 1-\rho, \quad \mathbf{E} = -\nabla\phi, \tag{25}$$

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 (24) can also be written in conservative form

$$\frac{\partial f}{\partial t} + \nabla_{\mathbf{x},\mathbf{v}}\cdot(\mathbf{F}f) = 0, \tag{26}$$

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

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*

$$0 \le f(\mathbf{x},\mathbf{v},t) \le \max_{(\mathbf{x},\mathbf{v})}(f_0(\mathbf{x},\mathbf{v})). \tag{27}$$

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

$$\frac{d}{dt}\left(\int (f(\mathbf{x},\mathbf{v},t))^p \, d\mathbf{x}\, d\mathbf{v}\right) = 0 \tag{28}$$

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

$$\int_V f(\mathbf{x},\mathbf{v},t)\, d\mathbf{x}\, d\mathbf{v} = \int_{F^{-1}(V)} f_0(\mathbf{y},\mathbf{u})\, d\mathbf{y}\, d\mathbf{u}. \tag{29}$$

- *Conservation of momentum*

$$\frac{d}{dt}\int \mathbf{v} f \, d\mathbf{x} d\mathbf{v} = \frac{d}{dt}\int \mathbf{J}\, d\mathbf{x} = 0. \tag{30}$$

- *Conservation of energy*

$$\frac{d}{dt}\left[\frac{1}{2}\int v^2 f \, d\mathbf{x} d\mathbf{v} + \frac{1}{2}\int E^2 \, d\mathbf{x}\right] = 0. \tag{31}$$

Proof. The system defining the associated characteristics writes

$$\frac{d\mathbf{X}}{dt} = \mathbf{V}(t), \tag{32}$$

$$\frac{d\mathbf{V}}{dt} = -\mathbf{E}(\mathbf{X}(t),t). \tag{33}$$

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 (32)-(33), the Vlasov equation (24) 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 (24) 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 (27).

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

$$\int (\nabla\times\mathbf{u})\times\mathbf{u}\, d\mathbf{x} = -\int \left(\mathbf{u}(\nabla\cdot\mathbf{u}) + \frac{1}{2}\nabla u^2\right) d\mathbf{x} = -\int \mathbf{u}(\nabla\cdot\mathbf{u})\, d\mathbf{x}. \tag{34}$$

Let us notice in particular that taking $\mathbf{u} = \mathbf{E}$ in the previous equality with $\mathbf{E}$ solution of the Poisson equation (25), 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

$$\int \mathbf{E}\rho\, d\mathbf{x} = 0. \tag{35}$$

Let us now introduce the Green formula on the divergence:

$$\int_\Omega \nabla\cdot\mathbf{F}q + \int_\Omega \mathbf{F}\cdot\nabla q = \int_{\partial\Omega} (\mathbf{F}\cdot\mathbf{n})\, q \quad \forall\mathbf{F}\in H(div,\Omega),\ q\in H^1(\Omega), \tag{36}$$

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 (24) 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 (36) 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 (35). 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 (36) 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

$$\frac{d}{dt}\int |\mathbf{v}|^2 f\, d\mathbf{x}d\mathbf{v} = -2\int \mathbf{E}\cdot\mathbf{J}\, d\mathbf{x} = 2\int \nabla\phi\cdot\mathbf{J}\, d\mathbf{x}. \tag{37}$$

On the other hand, integrating the Vlasov equation (24) 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 (36), the Poisson equation (25) 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 (37) and using that $\mathbf{E} = -\nabla\phi$, we get the conservation of energy. □

CHAPTER 4

Numerical methods for the Vlasov equation

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:

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f = 0, \tag{38}$$

with $\mathbf{v}$ fixed and

$$\frac{\partial f}{\partial t} + \frac{q}{m}\mathbf{E}(\mathbf{x},t)\cdot\nabla_v f = 0, \tag{39}$$

with $\mathbf{x}$ fixed. We then get two constant coefficient advections that can are easier to solve. This is obvious for (38) as $\mathbf{v}$ does not depend on t and x. On the other hand, integrating (39) 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 4. *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

$$\frac{du}{dt} = (A+B)u, \tag{40}$$

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 (40) into

$$\frac{du}{dt} = Au, \tag{41}$$

$$\frac{du}{dt} = Bu. \tag{42}$$

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 (41) and then (42). 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 (41) on a half time step, and then (42) on a full time step and again (41) 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 6.
- *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 5. *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* [**44**].

Remark 6. *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:*

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

2. 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.

2.1. Splines.

2.1.1. *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 $B - splines$.

2.1.2. *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 2 (B-Spline). *The i-th B-Spline of degree p is defined by the recurrence relation:*

$$N_j^{p+1} = w_j^{p+1} N_j^p + (1 - w_{j+1}^{p+1}) N_{j+1}^p \tag{43}$$

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 $p - 1$

$$N_i^{p\prime}(x) = p \left(\frac{N_i^{p-1}(x)}{t_{i+p} - t_i} - \frac{N_{i+1}^{p-1}(x)}{t_{i+p+1} - t_{i+1}} \right). \tag{44}$$

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 (43) becomes

$$N^{p+1}(x) = \frac{x}{p} N^p(x) + \frac{p + 1 - x}{p} N^p(x - 1), \tag{45}$$

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 7. *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 [**15**] 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 [**39**].

Lemma 1. *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 (45) we have

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

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 $0 \le \alpha \le 1$ 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 (45) 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.

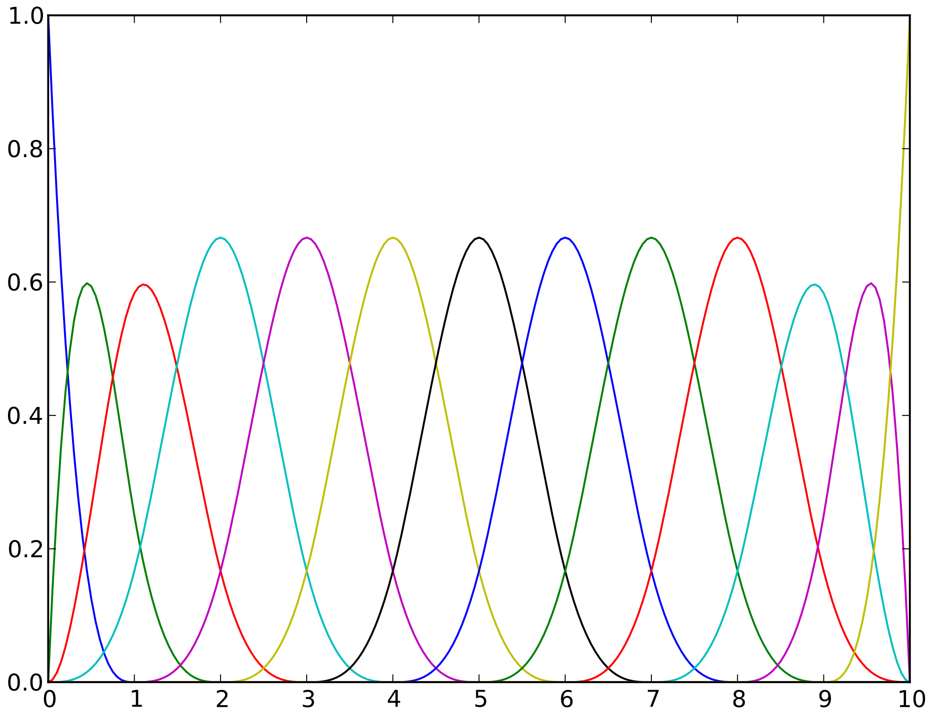

Figure 1. Cubic splines with open boundary conditions with knots at integers

For this a straightforward computation yields

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

□

2.2. 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.

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.

2.3. Cubic spline interpolation. Let us consider a uniform mesh of the interval $[a,b]$ defined by $x_i = a + ih$, $i = 0,\dots,N$, 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 $i = 0,\dots,N$, $f_h \in \mathbb{P}_3([x_i,x_{i+1}])$ 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.

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

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 $|x_i - x_j| \geq 2h$.

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

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

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

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

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.

2.4. Discrete Fourier Transform.

2.4.1. *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 3. *Discrete Fourier Transform.*

- *The* ***dicrete Fourier transform*** *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 2. *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:*

$$(F,G) = F^T \bar{G} = \hat{F}^T \bar{\hat{G}} = (\hat{F}, \hat{G}), \tag{46}$$

- *The discrete Plancherel identity:*

$$\|F\| = \|\hat{F}\|, \tag{47}$$

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

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

2.4.2. *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 intial 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 $\hat{f}_k$ pour $-N/2 \leq k < 0$ 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.

2.4.3. *Computing an approximate solution of the Poisson equation using the FFT.* 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 9. *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* [**11**].

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 10. *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.*

2.4.4. *Circulant matrices.*

DEFINITION 4. *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 7. *The eigenvalues of the circulant matrix* M *are given by*

$$\lambda_k = \sum_{j=0}^{n-1} c_j \omega^{jk}, \tag{48}$$

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 8. *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 (48) 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.

2.4.5. *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 $x_0, \dots, x_N$ 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 [**6**] for further information.

3. Semi-Lagrangian methods

Semi-Lagrangian methods have become, far behing 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.

Most semi-Lagrangian solvers are based on cubic spline interpolation which has proven very efficient in this context. So let us start by introducing this interpolation.

3.1. The classical semi-Lagrangian method. Let us consider an abstract scalar advection equation of the form

$$\frac{\partial f}{\partial t} + \mathbf{a}(\mathbf{x}, t) \cdot \nabla f = 0. \tag{49}$$

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 (49) 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 2.

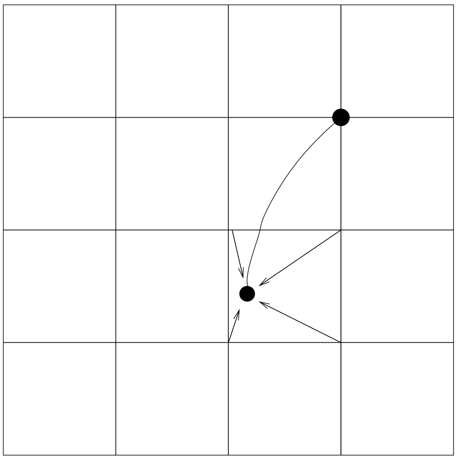

FIGURE 2. Sketch of the classical semi-Lagrangian method.

REMARK 11. *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.*

This semi-Lagrangian method has been initially introduced for the 1D Vlasov-Poisson equation by Cheng and Knorr [**14**] in 1976. It was based on a cubic spline interpolation and a directional Strang splitting method and is still one of the most used methods for this problem.

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
$$\frac{\partial f}{\partial t} + E^{n+1} \frac{\partial f}{\partial v} = 0 \tag{50}$$
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 (50), 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.

3.2. Conservation properties of the split 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.

3.2.1. *Conservation of mass.*

PROPOSITION 9. *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

$$\sum_j f_j^{k+1} = \sum_j S(x_j - a\Delta t) = \sum_{i,j} c_i N^p(j-i-a\Delta t/h) = \sum_i c_i \sum_j N^p(j-i-a\Delta t/h) = \sum_i c_i,$$

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. □

3.2.2. *Conservation of total momentum.*

PROPOSITION 10. *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 $\sum_i n_i E_i = 0$ 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

$$\begin{aligned}\sum_j j N^p(j-k+E_i\Delta t/\Delta v) &= \sum_j (j - l + E_i \Delta t/\Delta v) N^p(j - l + E_i\Delta t/\Delta v) \\ &\quad + (l - E_i\Delta t/\Delta v)\sum_j N^p(j - l + E_i \Delta t/\Delta v),\end{aligned}$$

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 1 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 11. *The FFT spectral Poisson solver introduced previously 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. □

3.3. A semi-Lagrangian method without splitting. 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 simplifies the algorithm. However the splitting itself is a source of errors giving a even greater importance to the axes directions. In some cases it is interesting to develop a semi-Lagrangian method without splitting. In this case the origins of the characteristics need to be computed numerically as the solutions for different initial conditions of the following ordinary differential equation (ODE):

$$\frac{dV}{dt} = E(X(t),t), \quad \frac{dX}{dt} = V.$$

Note that this ODE needs to be solved backward in time as we are backtracking the characteristics. The algorithm enabling to go from time step t_n to time step t_{n+1} is then the following:
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}. An order 2 predictor-corrector scheme can be defined as follows:

(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}.

2D spline interpolation.

3.4. 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 decomposing $\mathbf{z}$ 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

$$\frac{\partial \rho}{\partial t} + \mathbf{v}_D \cdot \nabla \rho = 0, \; -\Delta \phi = \rho,$$

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, are there are filaments appearing and then vortex roll-ups. An illustration of this phenomenon is displayed in Figure 3 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.

3.5. 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 proved by showing that the resulting scheme is algebraically equivalent to a conservative scheme. See [**19**] 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.

A scheme of principle is given in figure 4.

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 3.

For this reason, and also because the conservative method becomes very simple in this case, as the 1D cell is completely determined by its enpoints, we shall only use it for split 1D equations in the conservative form

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial x}(a(x,t)f) = 0.$$

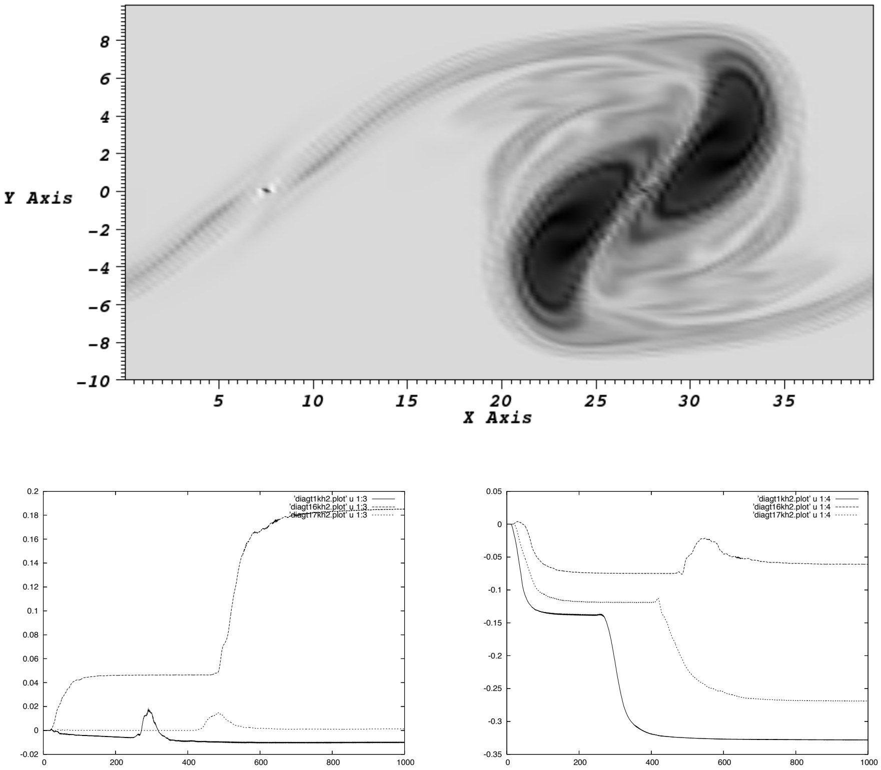

Figure 3. Evolution of a Kelvin-Helmhotz instability for the guiding-center model. The top figure displays a snapshot of the distribution function during the creation of a vortex. The snapshot is taken a the time corresponding to the large increase of the L^1 norm on the bottom left figure. The bottom 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).

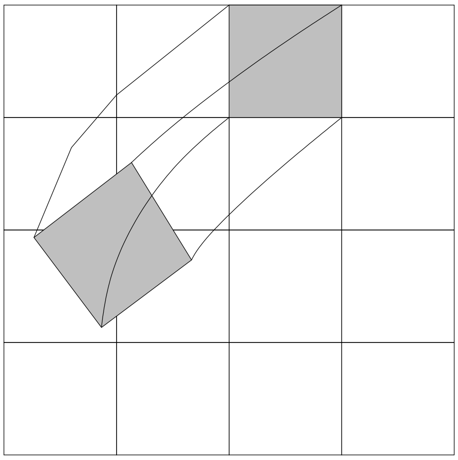

Figure 4. Idea of conservative semi-Lagrangian method.

Let us now detail the 3 steps of the algorithm in the 1D case starting with step 1. This step consists in construction on each cell 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

$$W(x_{j+\frac{1}{2}}) - W(x_{j-\frac{1}{2}}) = h_j f_j^n = \tilde{p}_m(x_{j+\frac{1}{2}}) - \tilde{p}_m(x_{j-\frac{1}{2}}).$$

Let us take for $\tilde{p}_m$ an interpolation polynomial at points $x_{j+\frac{1}{2}}$ of the function W, which will yield 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.

It remains to choose the type of interpolation. The simplest way is to use Lagrange interpolation with as many neighboring points as needed for the chosen degree. One can then choose a centered stencil or try an choose a stencil enabling to reduce the oscillations. ENO type stencil have not proved efficient for the Vlasov equation, as they increase the diffusivity, but well designed WENO methods have had some success [**37, 27, 28**]. 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 [**21**] only such very weak limiting is performed.

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.

The second step of the method consists in computing the origin of the characteristics ending at the grid points. This step is in principle identical as for the classical semi-Lagrangian algorithm. 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 $\frac{\partial f}{\partial t} + \frac{\partial}{\partial t}(a(t,x)f) = 0$.

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):

$$x_i - x_i^* = \frac{\Delta t}{2}[a(t_{n+1}, x_i) + a(t_n, x_i^*)]. \tag{51}$$

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 that linear interpolation. And in this case, as described in [**18**] 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 $x_i - x_i^* = (i - i_0 - \beta)\Delta x$, 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

$$i - i_0 - \beta_i = \frac{\Delta t}{2\Delta x}[a(t_{n+1}, x_i) + (1-\beta_i)a(t_n, x_{i_0}) + \beta_i a(t_n, x_{i_0+1})].$$

From this we can extract the following formula for β_i:

$$\beta_i = \frac{i - i_0 - \frac{\Delta t}{2\Delta x}[a(t_{n+1}, x_i) + a(t_n, x_{i_0})]}{1 + \frac{\Delta t}{2\Delta x}[a(t_n, x_{i_0+1}) - a(t_n, x_{i_0})]}.$$

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.

The third and last step consists in computing the average value on the cell at time t_{n+1} by using the relation

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f^{n+1}(x)\,dx = \int_{X(t_n; x_{i-\frac{1}{2}}, t_{n+1})}^{X(t_n; x_{i+\frac{1}{2}}, t_{n+1})} f^n(x)\,dx,$$

where $f^n(x)$ is the polynomial function reconstructed on each cell in step 1.

Positivity. 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 [**21, 18**].

4. Particle methods

4.1. 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 intial 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 [**8**] and Hockney and Eastwood [**24**]. Mathematical convergence proofs of the algorithms have also been performed in some special cases, see Neunzert and Wick [**26**], Cottet and Raviart [**17**], Victory and Allen [**36**] and Wollman [**42**].

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 dynamics has been introduced by Bateson and Hewett [**4**], 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.

4.2. 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 initialized such that $f_N(\mathbf{x}, \mathbf{v}, 0)$ is an approximation, in some sense that remains to be precized, of the intial 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 12. *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. □

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:

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 intial positions in a random or pseudo-random way using the probability density associated to f_0.

Remark 12. *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.*

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

$$S^0(x) = \begin{cases} \frac{1}{\Delta x} & \text{si } -\frac{\Delta x}{2} \le x < \frac{\Delta x}{2}, \\ 0 & \text{else.} \end{cases}$$

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

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

B-splines verify the following important properties

Proposition 13. • *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:

$$\rho_h(\mathbf{x}, t) = \int S(\mathbf{x} - \mathbf{x}')\rho_N(\mathbf{x}')\, d\mathbf{x}' = q \sum_k w_k S(\mathbf{x} - \mathbf{x}_k), \tag{52}$$

$$\mathbf{J}_h(\mathbf{x}, t) = \int S(\mathbf{x} - \mathbf{x}')\mathbf{J}_N(\mathbf{x}')\, d\mathbf{x}' = q \sum_k w_k S(\mathbf{x} - \mathbf{x}_k)\mathbf{v}_k. \tag{53}$$

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

$$\mathbf{E}_h(\mathbf{x}, t) = \sum_j \mathbf{E}_j(t) S(\mathbf{x} - \mathbf{x}_j), \quad \mathbf{B}_h(\mathbf{x}, t) = \sum_j \mathbf{B}_j(t) S(\mathbf{x} - \mathbf{x}_j). \tag{54}$$

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 $\frac{1}{6}\mathbf{E}_{j-1} + \frac{2}{3}\mathbf{E}_j + \frac{1}{6}\mathbf{E}_{j+1}$.

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 13. *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.*

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.

$$\mathbf{v}_k^{n+\frac{1}{2}} = \mathbf{v}_k^n + \frac{q\Delta t}{2m}\mathbf{E}_k^n(\mathbf{x}_k^n), \tag{55}$$

$$\mathbf{x}_k^{n+1} = \mathbf{x}_k^n + \Delta t \mathbf{v}_k^{n+\frac{1}{2}}, \tag{56}$$

$$\mathbf{v}_k^{n+1} = \mathbf{v}_k^{n+\frac{1}{2}} + \frac{q\Delta t}{2m}\mathbf{E}_k^{n+1}(\mathbf{x}_k^{n+1}). \tag{57}$$

We notice that step (57) needs the electric field at time t_{n+1}. It can be computed after step (56) by solving the Poisson equation which uses as input ρ_h^{n+1} that needs only $\mathbf{x}_k^{n+1}$ and not $\mathbf{v}_k^{n+1}$.

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 (52)-(53).
(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 (54).
(4) Particles are advanced using a numerical scheme for the characteristics for example Verlet (55)-(57).

Probabilistic interpretation of the PIC method. Even when the particles are initialized in a non random manner, the non linear interactions along with numerical errors introduce random effects after some time. The right tool for the mathematical investigation of PIC methods is thus probability and the analysis is similar to this of other Monte-Carlo numerical methods.

A Monte-Carlo description of the PIC method can be introduced as follows: We start by drawing macro-particles in phase space as a random realization of the probability law associated to the probability density p obtained from the intial distribution f_0 by

$$p = \frac{1}{\mathcal{N}} f_0, \qquad \text{where } \mathcal{N} = \int f_0 \, dx \, dv.$$

When the particles are initialized, they are advance using the equations of motion which are deterministic, but at each time t the particles given by their positions in phase space $(\mathbf{x}_k(t), \mathbf{v}_k(t))_{1\leq k\leq N}$ represent a realization of the probability law associated to the density $p(\mathbf{x}, \mathbf{v}, t) = \frac{1}{\mathcal{N}} f(\mathbf{x}, \mathbf{v}, t)$ linked to the solution at time t of the Vlasov equation.

In this framework, the different physical variables can be expressed as expectancies under the probability law of density p. For a given function g its expectancy is then defined by

$$\mathbb{E}_p(g) = \int g(\mathbf{x}, \mathbf{v}) p(\mathbf{x}, \mathbf{v}, t) \, d\mathbf{x} \, d\mathbf{v} = \frac{1}{\mathcal{N}} \int g(\mathbf{x}, \mathbf{v}) f(\mathbf{x}, \mathbf{v}, t) \, d\mathbf{x} \, d\mathbf{v}.$$

Thanks to the law of large numbers, these expectancies can be approximated by random realizations:

$$\mathbb{E}_f(g(\mathbf{x}, \mathbf{v})) \approx \frac{1}{N} \sum_{k=1}^{N} g(\mathbf{x}_k, \mathbf{v}_k).$$

Moreover, the central limit theorem enables to obtain an error estimate $\mathbb{E}_f(g(\mathbf{x}, \mathbf{v})^2) < +\infty$, and so the error

$$\epsilon_N = \mathbb{E}_f(g(\mathbf{x}, \mathbf{v})) - \frac{1}{N} \sum_{k=1}^{N} g(\mathbf{x}_k, \mathbf{v}_k)$$

is such that $\frac{\sqrt{N}}{\sigma} \epsilon_N$ converges to the centered reduced Gaussian whose variance σ is given by

$$\begin{aligned} \sigma^2 &= \mathbb{E}_f(g(\mathbf{x}, \mathbf{v})^2) - \mathbb{E}_f(g(\mathbf{x}, \mathbf{v}))^2, \\ &= \int g(\mathbf{x}, \mathbf{v})^2 f(\mathbf{x}, \mathbf{v}) \, dx \, dv - \Big(\int g(\mathbf{x}, \mathbf{v}) f(\mathbf{x}, \mathbf{v}) \, dx \, dv\Big)^2. \end{aligned}$$

We can deduce that this approximation converges as $1/\sqrt{N}$ when N goes to $+\infty$ and the approximation is all the better that the variance is small. In particular variance reduction techniques used in statistics are a mean to improve the approximation.

The different physical quantities can be interpreted thanks to the probabilistic terminology. The kinetic energy is

$$\mathcal{N} \mathbb{E}_f(|\mathbf{v}|^2) = \int |\mathbf{v}|^2 f(\mathbf{x}, \mathbf{v}) \, d\mathbf{x} \, d\mathbf{v},$$

the charge density at point $\mathbf{x}_i$ is

$$\rho_i = \mathcal{N} \mathbb{E}_f(S_i) = \int S(\mathbf{x} - \mathbf{x}_i) f(\mathbf{x}, \mathbf{v}) \, d\mathbf{x} \, d\mathbf{v},$$

and the current density at point $\mathbf{x}_i$ is

$$J_i = \mathcal{N} \mathbb{E}_f(\mathbf{v} S_i) = \int S(\mathbf{x} - \mathbf{x}_i) \mathbf{v} f(\mathbf{x}, \mathbf{v}) \, d\mathbf{x} \, d\mathbf{v}.$$

Moreover these expectancies can be approximated with a random realization in the following way:

$$\mathbb{E}_f(|\mathbf{v}|^2) \approx \frac{1}{N}\sum \mathbf{v}_k^2, \quad \rho_i = \mathbb{E}_f(S_i) \approx \frac{1}{N}\sum S(\mathbf{x}_k - \mathbf{x}_i),$$

$$J_i = \mathbb{E}_f(vS_i) \approx \frac{1}{N}\sum S(\mathbf{x}_k - \mathbf{x}_i)\mathbf{v}_k.$$

Bibliography

[1] S.J. Allfrey, R. Hatzky, A revised delta f algorithm for nonlinear PIC simulations, Comp. Phys. Commun. 154 (2) (2003) 98-104.

[2] L. Ahlfors. *Complex Analysis.* McGraw-Hill, 1979.

[3] H. Amman, *Ordinary differential equations*, De Gruyter, New-York 1990.

[4] William B. Bateson, Dennis W. Hewett, *Grid and particle hydrodynamics: beyond hydrodynamics via fluid element particle-in-cell.* J. Comput. Phys. 144 (1998), no. 2, 358–378.

[5] Régine Barthelmé, Le problème de conservation de la charge dans le couplage des équations de Vlasov et de Maxwell, thèse de l'Université Louis Pasteur, 2005, spécialité Mathématiques. `http://www-irma.u-strasbg.fr/irma/publications/2005/05014.shtml`

[6] J.-P. Beirut, L. N. Trefethen, Barycentric Lagrange Interpolation, SIAM Review 46 (3), (2004), 501–517.

[7] R.E. Bellman, R.S. Roth. *The Laplace Transform.* World Scientific, 1984.

[8] C. K. Birdsall, A. B. Langdon, *Plasma physics via computer simulation*, Institute of Physics, Bristol (1991) p. 359.

[9] N. Besse, Convergence of a semi-Lagrangian scheme for the one-dimensional Vlasov-Poisson system. SIAM J. Numer. Anal. 42 (2004), no. 1, 350–382.

[10] N. Besse, M. Mehrenberger, Convergence of classes of high-order semi-Lagrangian schemes for the Vlasov-Poisson system. Math. Comp. 77 (2008), no. 261, 93–123.

[11] C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag New-York (1988).

[12] C. Cercignani, The Boltzmann Equation and its Applications, Springer-Verlag New York, 1988.

[13] F. Chen, *Introduction to plasma physics and controlled fusion*, Springer New-York 1984.

[14] C.Z. Cheng, G. Knorr, The integration of the Vlasov equation in configuration space, J. Comput. Phys. 22 (1976) 330-351.

[15] Charles K. Chui, An introduction to Wavelets, Academic Press, Inc, San Diego, 1992.

[16] P. Colella, P. R. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulations.* J. Comput. Phys., **54**, 174-201, (1984).

[17] G.-H. Cottet, P.-A. Raviart, *Particle methods for the one-dimensional Vlasov-Poisson equations.* SIAM J. Numer. Anal. 21 (1984), no. 1, 52–76.

[18] N. Crouseilles, M. Mehrenberger, E. Sonnendrücker, Conservative semi-Lagragian methods for Vlasov-type equations. J. Comput. Phys., 229, (2010), pp 1927-1953.

[19] N. Crouseilles, Th. Respaud, E. Sonnendrücker, *A forward semi-Lagrangian scheme for the numerical solution of the Vlasov equation*, Comput. Phys. Comm., 180, (2009), pp. 1730-1745.

[20] P. Degond and P.-A. Raviart, On the paraxial approximation of the stationary Vlasov-Maxwell system *Math. Models Meth. Appl. Sciences* **3** (1993), 513–562.

[21] F. Filbet, E. Sonnendrücker, P. Bertrand, Conservative numerical schemes for the Vlasov equation, J. Comput. Phys., **172**, pp. 166-187, (2001).

[22] F. Filbet, E. Sonnendrücker, *Comparison of Eulerian Vlasov solvers*, Comput. Phys. Comm., **151**, pp. 247-266, (2003).

[23] B. D. Fried and S. D. Conte, *The Plasma Dispersion Function.* Academic, New York, 1961.

[24] RW Hockney and JW Eastwood. *Computer Simulation Using Particles.* Adam Hilger, Philadelphia, 1988.

[25] W. Magnus, S. Winkler, *Hill's equation*, John Wiley and Sons 1966.

[26] H. Neunzert, J. Wick, *The convergence of simulation methods in plasma physics.* Mathematical methods of plasmaphysics (Oberwolfach, 1979), 271–286, Methoden Verfahren Math. Phys., 20, Lang, Frankfurt, 1980.

[27] J.M. Qiu, A. Christlieb, *A conservative high order semi-Lagrangian WENO method for the Vlasov equation.* J. Comput. Phys. 229 , no. 4, 1130?1149, (2010).

[28] J.M. Qiu, C.W. Shu, *Conservative Semi-Lagrangian Finite Difference WENO Formulations with Applications to the Vlasov Equation.* Commun. Comput. Phys., 10 (2011), pp. 979-1000.

[29] R. Rieben, D. White, and G. Rodrigue, *High Order Symplectic Integration Methods for Finite Element Solutions to Time Dependent Maxwell Equations*, IEEE Transactions on Antennas and Propagation, Vol. 52, No. 8, pp. 2190-2195, 2004.

[30] G. Rodrigue, D. White. A vector finite element time-domain method for solving Maxwell's equation on unstructured hexaedral grids. SIAM J. Sci. Comput. Vol. 23, No. 3, pp. 683?706, (2001).

[31] G.R. Shubin, J.B. Bell, A modified equation approach to constructing fourth-order methods for acoustic wave propagation. SIAM J. Sci. Statist. Comput. 8 (1987), no. 2, 135–151.

[32] P.-J. Sacherer, P.-M. Lapostolle. *IEEE Transaction of nuclear science* **N-S 18** p. 1101.

[33] D. Sármány, M.A. Botchev, J.J.W. van der Vegt, Dispersion and dissipation error in high-order Runge-Kutta discontinuous Galerkin discretisations of the Maxwell equations. J. Sci. Comput. 33 (2007), no. 1, 47–74.

[34] L. Schwartz. *Méthodes mathématiques pour les sciences physiques*, Hermann, Paris, 1961.

[35] T.H. Stix, *Waves in plasma*, American Institute of Physics, 1992.

[36] H. D. Victory, Jr., Edward J. Allen, *The convergence theory of particle-in-cell methods for multidimensional Vlasov-Poisson systems.* SIAM J. Numer. Anal. 28 (1991), no. 5, 1207–1241.

[37] J.A. Carrillo and F. Vecil, Nonoscillatory interpolation methods applied to Vlasov-based models, SIAM Journal on Scientific Computing 29 (2007), p. 1179

[38] J. Villasenor and O. Buneman, *Rigorous charge conservation for local electromagnetic field solvers*, Comput. Phys. Commun. 69 (1992) pp. 306-316.

[39] Zlatko Udovičić, Calculation of the moments of the cardinal B-spline, Sarajevo journal of mathematics Vol.5 (18) (2009), 291–297.

[40] T. Warburton, An Explicit Construction for Interpolation Nodes on the Simplex. *Preprint* `www.caam.rice.edu/~timwar/preprints/JEMnodesV1.pdf` (2005).

[41] Eric W. Weisstein. "Jacobi Polynomial." From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/JacobiPolynomial.html`

[42] S. Wollman, On the approximation of the Vlasov-Poisson system by particle methodes. SIAM J. Numer. Anal. **37** no. 4 (2000), pp. 1369–1398.

[43] K. S. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, IEE Trans. Ant. Propagat., Vol. 14, 302, 1966.

[44] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A **150**, p. 262, (1990).

[45] M. Zerroukat, N. Wood, A. Staniforth, *A monotonic and positive-definite filter for a Semi-Lagrangian Inherently Conserving and Efficient (SLICE) scheme*, Q.J.R. Meteorol. Soc., **131**, pp 2923-2936, (2005).

[46] M. Zerroukat, N. Wood, A. Staniforth, *The Parabolic Spline Method (PSM) for conservative transport problems*, Int. J. Numer. Meth. Fluids, **51**, pp. 1297-1318, (2006).