Monte Carlo Methods with applications to plasma physics

Eric Sonnendrücker
Max-Planck-Institut für Plasmaphysik
and
Zentrum Mathematik der TU München

Lecture notes
Sommersemester 2014

version of July 7, 2014
Contents

Chapter 1. Introduction 5
1. Plasmas 5
2. Controlled thermonuclear fusion 5
3. The ITER project 7
4. The Vlasov-Maxwell and Vlasov-Poisson equations 7
5. Characteristics of a transport equation 9
6. The Particle In Cell (PIC) method 12

Chapter 2. Monte Carlo simulation 15
1. Principle of Monte Carlo method 15
2. Background in probability theory 15
3. Monte Carlo Simulation 21
4. Initialisation of given PDF 28
5. Variance reduction techniques 30
6. Coupling the Monte Carlo Vlasov solver with a grid based Poisson solver 34

Bibliography 41
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 \url{en.wikipedia.org/wiki/Plasma_physics}.

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

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

More information on fusion can be found on wikipedia sites devoted to inertial fusion and magnetic fusion: http://en.wikipedia.org/wiki/Inertial_confinement_fusion http://en.wikipedia.org/wiki/Magnetic_confinement_fusion
Figure 3. Artist view of the ITER Tokamak

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 and Vlasov-Poisson equations

At the microscopic level, a plasma is composed of a number of particles that evolve following the laws of classical or relativistic dynamics. The force that is by far dominant is the electromagnetic interaction coming from the electromagnetic field generated by the particles. This can be mathematically modelled by Newton’s equation of motion for each particle and the direct computation of the electromagnetic field generated by each particle. However there are way too many particles in a plasma for this too be usable for numerical computations. Moreover such a detailed
description is not necessary for most applications. An approximate model which provides a very accurate description of the evolution of a plasma is the Vlasov equation obtained with tools from statistical physics. It is written for non-relativistic particles

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

where \(m\) is the mass of the particles, \(q\) their charge and \(f \equiv f(x,v,t)\) represents the particle density in phase space at point \((x,v)\) 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:

\[
\begin{align*}
-\frac{1}{c^2} \frac{\partial E}{\partial t} + \nabla \times B &= \mu_0 J, \\
\frac{\partial B}{\partial t} + \nabla \times E &= 0, \\
\nabla \cdot E &= \frac{\rho}{\epsilon_0}, \\
\nabla \cdot B &= 0,
\end{align*}
\]

with

\[
\rho(x,t) = \sum_s q_s \int f_s(x,v,t) \, dv, \quad J(x,t) = \sum_s q_s \int f_s(x,v,t) v \, dv.
\]

There are many situations in which the time evolution of the electromagnetic field is slow compared to the phenomenon being investigated. In this case a quasi static approximation of Maxwell’s equations can be used. Often in these cases the electric field is by far dominant in the Lorentz force \(\frac{q}{m} (E + v \times B)\). Then the magnetic field can be neglected and Maxwell’s equations reduce to

\[
\nabla \times E = 0, \quad \nabla \cdot E = \frac{\rho}{\epsilon_0}.
\]

Under some geometric conditions on the computational domain \(\nabla \times E = 0\) implies that there exists a scalar function \(\phi\) called electrostatic potential such that \(E = -\nabla \phi\), then \(\phi\) is a solution of the Poisson equation \(-\Delta \phi = \frac{\rho}{\epsilon_0}\) and our model reduces to the Vlasov-Poisson system

\[
\begin{align*}
\frac{\partial f_s}{\partial t} + v \cdot \frac{\partial f_s}{\partial x} - \frac{q}{m} \nabla \phi \cdot \frac{\partial f_s}{\partial v} &= 0, \\
\rho(x,t) &= \sum_s q_s \int f_s(x,v,t) \, dv, \\
-\Delta \phi &= \frac{\rho}{\epsilon_0}.
\end{align*}
\]

This is the system we will consider throughout the lecture.

The macroscopic quantities which can be measured are defined using the first three velocity moments of the distribution function \(f(x,v,t)\). Taking velocity moments of the Vlasov equation one can also derive relations between these quantities, which define a fluid model for plasmas.

- The particle density is defined by

\[
n(x,t) = \int f(x,v,t) \, dv,
\]

- The mean velocity \(u(x,t)\) verifies

\[
n(x,t) u(x,t) = \int f(x,v,t) v \, dv,
\]
• The pressure tensor \( P(x, t) \) is defined by
\[
P(x, t) = m \int f(x, v, t)(v - u(x, t)) \otimes (v - u(x, t)) \, dv.
\]
• The scalar pressure is one third of the trace of the pressure tensor
\[
p(x, t) = \frac{m}{3} \int f(x, v, t)|v - u(x, t)|^2 \, dv,
\]
• The temperature \( T(x, t) \) is related to the pressure and the density by
\[
T(x, t) = \frac{p(x, t)}{n(x, t)}.
\]

In some cases inter-particle collisions rather than mere interactions through the mean electromagnetic field play an important role for the evolution of the plasma. This is generally added through a collision operator on the right-hand-side of the Vlasov equation. A fairly general collision operator is the Boltzmann operator \( Q(f, f) \) which is quadratic in the distribution function. In this lecture we will consider only a linear Fokker-Planck operator of the form
\[
C(f) = \nu \frac{\partial}{\partial v} \left( \frac{\partial f}{\partial v} + \frac{v - u}{T} f \right),
\]
with \( \nu \) a constant collision frequency, \( u(t, x) \) and \( T(t, x) \) being either given functions or defined self consistently from the distribution function by
\[
n(t, x) = \int f(t, x, v) \, dv, \quad n(t, x)u(t, x) = \int f(t, x, v)v \, dv, \quad n(t, x)e(t, x) = \int f(t, x, v)v^2 \, dv,
\]
and \( T(t, x) = e(t, x) - u(t, x)^2 \).

End of lecture 1.

5. Characteristics of a transport equation

The Vlasov equation is a transport equation in phase space. It can be written in abstract form
\[
\frac{\partial f}{\partial t} + \mathbf{A} \cdot \nabla_{x,v} f = 0,
\]
where the gradient is now with respect to all the phase space variables and \( \mathbf{A} = (v, -\frac{q}{m} \nabla \phi)^T \) for the Vlasov-Poisson equation. Note that \( \nabla_{x,v} \mathbf{A} = 0 \) and therefore the Vlasov equation can also be written in conservative form
\[
\frac{\partial f}{\partial t} + \nabla_{x,v} \cdot (\mathbf{A} f) = 0.
\]

An essential feature of the scalar transport equation as the Vlasov equation is that they can be solved using the characteristic which are the solutions of the ordinary differential equation
\[
\frac{dZ}{dt} = \mathbf{A}(t, Z), \quad Z(s) = z.
\]
where \( Z = (X, V) \) denotes the phase space variable. The solution of the Cauchy problem (ODE + initial condition) \([2]\) will be denoted by \( Z(t; s, z) \) which is the solution of the ODE at time \( t \) whose value was \( z \) at time \( s \). It can be easily seen that the characteristics can be used for computing the solution of the advection equation \([1]\). Indeed, assuming \( Z(t; s, z) \) satisfies \([2]\), we have
\[
\frac{d}{dt} f(t, Z(t; s, z)) = \left( \frac{\partial f}{\partial t} + \frac{dZ}{dt} \cdot \nabla f \right)(t, Z(t; s, z)) = \left( \frac{\partial f}{\partial t} + \mathbf{A} \cdot \nabla f \right)(t, Z(t; s, z)) = 0
\]
as \( f \) is a solution of \([1]\).

The characteristics can be travelled forward or backward in time and are uniquely determined by any time-phase-space couple \((s, z)\) which acts as an initial condition. Because they play an
essential role in the theory and numerical methods for the Vlasov equation we shall study them a bit more formally.

Let us recall the classical theorem of the theory of ordinary differential equations (ODE) which gives existence and uniqueness of the solution of \([2]\). The proof can be found in \([2]\) for example.

**Theorem 1.** Assume that \(A \in C^{k-1}([0,T] \times \mathbb{R}^d)\), \(\nabla A \in C^{k-1}([0,T] \times \mathbb{R}^d)\) for \(k \geq 1\) and that
\[
|A(z,t)| \leq \kappa(1 + |z|) \quad \forall t \in [0,T] \quad \forall z \in \mathbb{R}^d.
\]
Then for all \(s \in [0,T]\) and \(z \in \mathbb{R}^d\), there exists a unique solution \(Z \in C^k([0,T] \times [0,T] \times \mathbb{R}^d)\) of \([2]\).

**Proposition 1.** Assuming that the vector field \(A\) satisfies the hypotheses of the previous theorem, we have the following properties:

(i) \(\forall t_1, t_2, t_3 \in [0,T]\) and \(\forall z \in \mathbb{R}^d\)
\[
Z(t_3; t_2, Z(t_2; t_1, z)) = Z(t_3; t_1, z).
\]

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

(iii) The jacobian \(J(t;s,1) = \det(\nabla Z(t; s, z))\) verifies
\[
\frac{\partial J}{\partial t} = (\nabla \cdot A)(t; Z(t; s, z))J,
\]
and \(J > 0\). In particular if \(\nabla \cdot A = 0\), \(J(t; s, 1) = J(s; s, 1) = \det \mathbb{I}_d = 1\), where \(\mathbb{I}_d\) is the identity matrix of order \(d\).

**Proof.**

(i) The points \(z = Z(t_1; t_1, z), Z(t_2; t_1, z), Z(t_3; t_1, z)\) are on the same characteristic curve. This curve is characterized by the initial condition \(Z(t_1) = z\). So, taking any of these points as initial condition at the corresponding time, we get the same solution of \([2]\). We have in particular \(Z(t_3; t_2, Z(t_2; t_1, z)) = Z(t_3; t_1, z)\).

(ii) Taking \(t_1 = t_3\) in the equality \((i)\) we have
\[
Z(t_3; t_2, Z(t_2; t_3, z)) = Z(t_3; t_3, z) = z.
\]

Hence \(Z(t_3; t_2, .)\) is the inverse of \(Z(t_2; t_3, .)\) (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
\[
J(t; s, 1) = \det(\nabla Z(t; s, z)) = \det(\left(\frac{\partial Z_i(t; s, z)}{\partial z_j}\right)_{1 \leq i, j \leq d}).
\]
But \(Z\) verifies \(\frac{d}{dt} = A(Z(t), t)\). So we get in particular taking the ith line of this equality \(\frac{dZ_i}{dt} = A_i(Z(t), t)\). And taking the gradient we get, using the chain rule,
\[
\frac{d}{dt} \nabla Z_i = \sum_{k=1}^{d} \frac{\partial A_i}{\partial z_k} \nabla Z_k.
\]
For a \(d \times d\) matrix \(M\) the determinant of \(M\) is a \(d\)-linear alternating form taking as arguments the columns of \(M\). So, denoting by \((\ldots, \ldots, \ldots)\) this alternating \(d\)-linear form, we can write \(\det M = (M_1, \ldots, M_d)\) where \(M_j\) is the jth column of \(M\). Using this notation
in our case, we get
\[
\frac{\partial J}{\partial t}(t; s, 1) = \frac{\partial}{\partial t} \det(\nabla Z(t; s, z))
= (\frac{\partial \nabla Z_1}{\partial t}, \nabla Z_2, \ldots, \nabla Z_d) + \cdots + (\nabla Z_1, \nabla Z_2, \ldots, \frac{\partial \nabla Z_d}{\partial t})
= \sum_{k=1}^{d} \frac{\partial A_1}{\partial z_k} \nabla Z_k, \nabla X_2, \ldots, \nabla Z_d) + \cdots + (\nabla Z_1, \nabla Z_2, \ldots, \sum_{k=1}^{d} \frac{\partial A_d}{\partial z_k} \nabla Z_k)
= \frac{\partial A_1}{\partial z_1} J + \cdots + \frac{\partial A_d}{\partial z_d} J,
\]
as \((\ldots, \ldots)\) is alternating and \(d\)-linear. Thus we have
\[
\frac{\partial J}{\partial t}(t; s, 1) = (\nabla \cdot A) J.
\]
On the other hand \(\nabla Z(s; s, z) = \nabla z = I_d\) and so \(J(s; s, 1) = \det I_d = 1\). \(J\) is a solution of the differential equation
\[
\frac{dJ}{dt} = (\nabla \cdot A) J, \quad J(s) = 1,
\]
which admits as the unique solution \(J(t) = e^{\int_s^t (\nabla \cdot A(\tau; s, z)) d\tau} > 0\) and in particular, if \(\nabla \cdot A = 0\), 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 (1) using the characteristics.

**Theorem 2.** Let \(f_0 \in C^1(\mathbb{R}^d)\) and \(A\) satisfying the hypotheses of the previous theorem. Then there exists a unique solution of the linear advection equation (1) associated to the initial condition \(f(z, 0) = f_0(z)\). It is given by
\[
f(z, t) = f_0(Z(0; t, z)),\tag{3}
\]
where \(Z\) represent the characteristics associated to \(A\).

**Proof.** The function \(f\) given by (3) is \(C^1\) as \(f_0\) and \(Z\) are, and \(Z\) is defined uniquely. Let’s verify that \(f\) is a solution of (1) and that it verifies the initial condition. First taking \(t = 0\) in (3) we get
\[
f(z, 0) = f_0(Z(0; 0, z)) = f_0(z)
\]
so that the initial condition is verified.

Let’s take the time derivative of (3)
\[
\frac{\partial f}{\partial t}(z, t) = \frac{\partial Z}{\partial s}(0; t, z) \cdot \nabla f_0(Z(0; t, z)),
\]
and taking the gradient of (3)
\[
\nabla f(z, t) = \nabla (f_0(Z(0; t, z))
= \sum_{k=1}^{d} \frac{\partial f_0}{\partial z_k} \nabla Z_k(0; t, z),
= \nabla Z(0; t, z) T \nabla f_0(Z(0; t, z)),
\]
in the sense of a matrix vector product with the jacobian matrix
\[
\nabla Z(0; t, z) = (\frac{\partial Z_k}{\partial z_l}(0; t, z))_{1 \leq k, l \leq d}.
\]
We then get
\begin{equation}
\frac{\partial f}{\partial t} + A \cdot \nabla f(z, t) = \frac{\partial Z}{\partial s}(0; t, z) \cdot \nabla f_0(Z(0; t, z)) + A(z, t) \cdot (\nabla Z(0; t, z)^T \nabla f_0(Z(0; t, z))) .
\end{equation}
Because of the properties of the characteristics we also have that
\[ Z(t; s, Z(s; r, z)) = Z(t; r, z) \]
and taking the derivative with respect to \( s \), we get
\[
\frac{\partial Z}{\partial s}(t; s, Z(s; r, z)) + \nabla Z(t; s, Z(s; r, z)) \frac{\partial Z}{\partial t}(s; r, z) = 0 .
\]
But by definition of the characteristics
\[
\frac{\partial Z}{\partial s}(s; r, z) = A(Z(s; r, z), s) \]
as this equation is verified for all values of \( t, r, s \) and so in particular for \( r = s \). It becomes in this case
\[
\frac{\partial Z}{\partial s}(t; s, z) + \nabla Z(t; s, z) A(z, s) = 0 .
\]
Plugging this expression into (4) we obtain
\[
\left( \frac{\partial f}{\partial t} + A \cdot \nabla f(z, t) \right) = -\nabla Z(0; t, z) A(z, t) \cdot \nabla f_0(Z(0; t, z)) + A(z, t) \cdot (\nabla Z(0; t, z)^T \nabla f_0(Z(0; t, z))) .
\]
But for a matrix \( M \in M_d(\mathbb{R}) \) and two vectors \( u, v \in \mathbb{R}^d \), on a \( (M u) \cdot v = u^T M^T v = u \cdot (M^T v) \).

Whence we get
\[
\frac{\partial f}{\partial t} + A \cdot \nabla f = 0,
\]
which means that \( f \) defined by (3) is solution of (1).

The problem being linear, if \( f_1 \) and \( f_2 \) are two solutions we have
\[
\frac{\partial}{\partial t} (f_1 - f_2) + A \cdot \nabla (f_1 - f_2) = 0 ,
\]
and using the characteristics \( \frac{d}{dt} (f_1 - f_2)(Z(t), t) = 0 \). So if \( f_1 \) and \( f_2 \) verify the same initial condition, they are identical, which gives the uniqueness of the solution which is thus the function given by formula (3). \( \square \)

6. The Particle In Cell (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 \((x_k(t), v_k(t))_{1 \leq k \leq N}\) 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
\[
f_N(x, v, t) = \sum_{k=1}^{N} w_k \delta(x - x_k(t)) \delta(v - v_k(t)).
\]
Positions \( x_k^0 \), velocities \( v_k^0 \) and weights \( w_k \) are initialised such that \( f_N(x, v, 0) \) is an approximation, in some sense that remains to be precised, of the initial distribution function \( f_0(x, v) \). The time evolution of the approximation is done by advancing the macro-particles along the characteristics of the Vlasov equation, \textit{i.e.} by solving the system of differential equations
\[
\begin{align*}
\frac{dx_k}{dt} &= v_k \\
\frac{dv_k}{dt} &= \frac{q}{m} E(x_k, t) \\
x_k(0) &= x_k^0, \quad v_k(0) = v_k^0.
\end{align*}
\]
PROPOSITION 2. The function $f_N$ is a solution in the sense of distributions of the Vlasov equation associated to the initial condition $f^0_N(x, v) = \sum_{k=1}^{N} w_k \delta(x - x_k^0) \delta(v - v_k^0)$.

PROOF. Let $\varphi \in C_0^\infty(\mathbb{R}^3 \times \mathbb{R}^3 \times [0, +\infty[)$. Then $f_N$ defines a distribution of $\mathbb{R}^3 \times \mathbb{R}^3 \times [0, +\infty[$ in the following way:

\[
\langle f_N, \varphi \rangle = \sum_{k=1}^{N} \int_0^T w_k \varphi(x_k(t), v_k(t), t) \, dt.
\]

We then have

\[
\left( \frac{\partial f_N}{\partial t}, \varphi \right) = -\langle f_N, \frac{\partial \varphi}{\partial t} \rangle = - \sum_{k=1}^{N} w_k \int_0^T \frac{\partial \varphi}{\partial t}(x_k(t), v_k(t), t) \, dt,
\]

but

\[
\frac{d}{dt}(\varphi(x_k(t), v_k(t), t)) = \frac{dx_k}{dt} \cdot \nabla_x \varphi + \frac{dv_k}{dt} \cdot \nabla_v \varphi + \frac{\partial \varphi}{\partial t}(x_k(t), v_k(t), t),
\]

and as $\varphi$ has compact support in $\mathbb{R}^3 \times \mathbb{R}^3 \times [0, +\infty[$, it vanishes for $t = 0$ and $t = T$. So

\[
\int_0^T \frac{d}{dt}(\varphi(x_k(t), v_k(t), t)) \, dt = 0.
\]

It follows that

\[
\langle \frac{\partial f_N}{\partial t}, \varphi \rangle = \sum_{k=1}^{N} w_k \int_0^T \left( \langle v_k \cdot \nabla_x \varphi + \frac{q}{m} E(x_k(t), t) \cdot \nabla_v \varphi \rangle \, dt \right.
\]

\[
= -\langle v \cdot \nabla_x f_N + \frac{q}{m} E(x_k(t) \cdot \nabla_v f_N, \varphi). \rangle
\]

Which means that $f_N$ verifies exactly the Vlasov equation in the sense of distributions. \hfill \Box

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 $f^0_N$ 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 $(x_k^0, v_k^0)$ the barycentres of the cells and for weights $w_k$ associated to the integral of $f_0$ on the corresponding cell: $w_k = \int_{V_k} f_0(x, v) \, dxdv$ so that $\sum_k w_k = \int f_0(x, v) \, dxdv$.
- **Monte-Carlo method:** Pick the initial positions in a random or pseudo-random way using the probability density associated to $f_0$.

**Remark 1.** 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.
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 use a smooth convolution kernel $S$ for this regularization procedure. $S$ could typically be a Gaussian or preferably in practice a smooth piecewise polynomial spline function which has the advantage of having a compact support.

The source for Poisson’s equations $\rho$ is defined from the numerical distribution function $f_N$, for a particle species of charge $q$ by

$$\rho_N = q \sum_k w_k \delta(x - x_k).$$

We then apply the convolution kernel $S$ to define $\rho$ at any point of space and in particular at the grid points:

$$\rho_h(x, t) = \int S(x - x') \rho_N(x') \, dx' = q \sum_k w_k S(x - x_k),$$

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:

$$\frac{dx_k}{dt} = v_k, \quad \frac{dv_k}{dt} = \frac{q}{m} E(x_k, t).$$

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 $x^n_k$, $v^n_k$ and $E^n_k$ known.

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

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

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

We notice that step (7) needs the electric field at time $t_{n+1}$. It can be computed after step (6) by solving the Poisson equation which uses as input $\rho^{n+1}_h$ that needs only $x^{n+1}_k$ and not $v^{n+1}_k$.

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 $\rho_h$ on the grid using relations (6).
2. We update the electrostatic field using a classical mesh based solver (finite differences, finite elements, spectral, ...).
3. We compute the fields at the particle positions.
4. Particles are advanced using a numerical scheme for the characteristics for example Verlet (5)-(7).
CHAPTER 2

Monte Carlo simulation

1. Principle of Monte Carlo method

The basic idea of Monte Carlo methods is to use probability calculations to compute integrals, assuming that a probability can be approximated by a large number of random events. A simple example of a Monte Carlo algorithm is given by the computation of $\pi$ using the area of a quarter circle: Consider the quarter circle $Q$ of radius one centred at zero embedded in the unit square $S = [0, 1]^2$. The the area of $Q$ is $|Q| = \frac{\pi}{4}$ and the area of $S$ is $|S| = 1$. Considering a uniformly distributed sample in $S$, the ratio $\frac{|Q|}{|S|}$ is the probability of an event being in $Q$. This can be approximated by the ratio $\frac{nQ}{n}$, where $n$ is the total number of draws and $nQ$ the number of draws in $Q$. Hence

$$\frac{|Q|}{|S|} = \frac{\pi}{4} \approx \frac{nQ}{n}$$

which yields an approximation of $\pi$ by $4\frac{nQ}{n}$ which is all the better that the number of draws $n$ is large.

Computations of continuous probabilities are strongly related to the computation of integrals, so that in practice one can recast the computation of integrals in the framework of probabilities and then use a large number of samples to approximate them. The purpose of this chapter is to formalise this and also give a way to estimate the error committed in this approximation.

2. Background in probability theory

As the most convenient framework for defining integrals is the Lebégues theory, which starts by defining measurable sets using $\sigma$-algebras, the good framework for abstract probability theory also needs these objects. However, after having defined them to make the connection with the mathematical probability literature, we will only consider probabilities on $\mathbb{R}^n$.

2.1. Probability spaces. Let us recall some standard definitions that can be found in any probability textbook.

Let $\Omega$ be a nonempty set.

**Definition 1.** A $\sigma$-algebra is a collection $\mathcal{F}$ of subsets of $\Omega$ with the properties

(i) $\Omega \in \mathcal{F}$,

(ii) If $A \in \mathcal{F}$ then $A^c := \Omega \setminus A \in \mathcal{F}$,

(iii) If $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcup_{i} A_i \in \mathcal{F}$.

Note that axioms (i) and (ii) imply that $\emptyset \in \mathcal{F}$ and axioms (ii) and (iii) imply that if $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcap_{i} A_i \in \mathcal{F}$, as $(\bigcup_{i} A_i^c)^c = \bigcap_{i} A_i^c$.

**Definition 2.** Let $\mathcal{F}$ be a $\sigma$-algebra of subsets of $\Omega$. Then $P : \mathcal{F} \to [0, 1]$ is called a probability measure provided:

(i) For all $A \in \mathcal{F}$ we have $0 \leq P(A) \leq 1$,

(ii) $P(\Omega) = 1$,
(iii) If \( A_1, A_2, \cdots \in \mathcal{F} \) are disjoint then \( P(\bigcup_i A_i) = \sum_i P(A_i) \).

It follows from (ii) and (iii) as \( \Omega \) and \( \emptyset \) are both in \( \mathcal{F} \) and disjoint that \( P(\emptyset) = 0 \). If follows from (i) and (iii) that if \( A \subset B \) then \( B \) is the disjoint union of \( A \) and \( B \setminus A \), so \( P(A) \leq P(A) + P(B \setminus A) = P(B) \).

End of lecture 2.

**Definition 3.** A triple \((\Omega, \mathcal{F}, P)\) is called probability space provided \( \Omega \) is any set, \( \mathcal{F} \) is a \( \sigma \)-algebra and \( P \) a probability measure on \( \mathcal{F} \).

**Terminology.** A set \( A \in \mathcal{F} \) is called an event, points \( \omega \in \Omega \) are called sample points and \( P(A) \) is the probability of event \( A \).

Let \( \mathcal{B} \) denote the Borel subsets of \( \mathbb{R}^n \) which is the smallest \( \sigma \)-algebra containing all the the open subsets of \( \mathbb{R}^n \). In particular it contains all the product intervals (open, semi-open, or closed).

**Example 1.** Let \( \Omega = \{\omega_1, \omega_2, \ldots, \omega_N\} \) be a finite set, and suppose we are given \( N \) numbers \( 0 \leq p_i \leq 1 \) for \( i = 1, \ldots, N \) satisfying \( \sum_{i=1}^{N} p_i = 1 \). We take \( \mathcal{F} \) to be all the possible subsets of \( \Omega \). Then for each \( A = \{\omega_{i_1}, \ldots, \omega_{i_m}\} \in \mathcal{F} \) with \( 1 \leq i_1 < \cdots < i_m \leq N \) we define
\[
P(A) := p_{i_1} + p_{i_2} + \cdots + p_{i_m}.
\]

Let us consider two concrete examples:
1) Throwing once a dice can be analysed with the following probability space: \( \Omega = \{1, 2, 3, 4, 5, 6\} \), \( \mathcal{F} \) consists of all the subsets of \( \Omega \) and \( p_i = \frac{1}{6} \) for \( i = 1, \ldots, 6 \). An event is a subset of \( \Omega \), for example \( A = \{2, 5\} \). The probability of a sample point to be in \( A \) is then \( P(A) = p_2 + p_5 = \frac{1}{3} \).
2) Consider throwing a coin twice. Then the set \( \Omega \) of all possible events is \( \{(H, T), (T, H), (T, T)\} \) where \( H \) stands for heads and \( T \) for tail, \( \mathcal{F} \) consists of all the subsets of \( \Omega \) and \( p_i = \frac{1}{4} \) for \( i = 1, \ldots, 4 \). A possible event \( A \) would be to throw heads at least once: \( A = \{(H, T), (T, H)\} \) and \( P(A) = \frac{3}{4} \), and other possible event \( B \) would be to throw tail the second time, then \( B = \{(H, T), (T, T)\} \) and \( P(B) = \frac{1}{2} \).

**Example 2.** The Dirac mass. Let \( z \in \mathbb{R}^n \) fixed and define for sets \( A \in \mathcal{B} \)
\[
P(A) := \begin{cases} 
1 & \text{if } z \in A, \\
0 & \text{if } z \notin A.
\end{cases}
\]
We call \( P \) the Dirac mass at \( z \) and denote it by \( P = \delta_z \).

**Example 3.** Assume \( f \) is a non negative integrable function such that \( \int_{\mathbb{R}^n} f(x) \, dx = 1 \). We define for sets \( A \in \mathcal{B} \)
\[
P(A) := \int_A f(x) \, dx.
\]
We call \( f \) the density of the probability measure \( P \).

**2.2. Random variables.** A probability space is an abstract construction. In order to define observables it is necessary to introduce mappings \( X \) from \( \Omega \) to \( \mathbb{R}^n \).

**Definition 4.** Let \((\Omega, \mathcal{F}, P)\) be a probability space. A mapping
\[
X : \Omega \to \mathbb{R}^n
\]
is called a \( n \)-dimensional random variable if for each \( B \in \mathcal{B} \), we have
\[
X^{-1}(B) \in \mathcal{F}.
\]
In other words, \( X \) is \( n \)-dimensional random variable on the probability space if it is \( \mathcal{F} \)-measurable.

This definition enables to define probabilities of events related to \( X \) by inducing a probability law on \((\mathbb{R}^n, \mathcal{B})\).
Proposition 3. Let $X$ be an $n$-dimensional random variable. Then $P_X : \mathcal{B} \to [0,1]$ defined by
\[ P_X(B) = P(X^{-1}(B)) \]
is a probability law on $(\mathbb{R}^n, \mathcal{B})$.

Proof. For $B \in \mathcal{B}$, the measurability of $X$ implies that $X^{-1}(B) \in \mathcal{B}$. So the probability $P(X^{-1}(B))$ is well defined and we just need to check the properties of a probability law, which is straightforward. □

Notation 1. The probability $P_X$ is often denoted conveniently $P_X(B) = P(X \in B)$.

2.3. Distribution function. Let $X$ be a $n$-dimensional random variable on the probability space $(\Omega, \mathcal{F}, P)$. Let us say that for two vectors $x \leq y$ if $x_i \leq y_i$ all the components of the vectors.

Definition 5. We call (cumulative) distribution function (CDF) of a random variable $X$ the function $F_X : \mathbb{R}^n \to [0,1]$ defined by
\[ F_X(x) = P(X \leq x), \quad \text{for } x \in \mathbb{R}^n. \]

Definition 6. Assume $X$ is a $n$-dimensional random variable and $F = F_X$ its distribution function. If there exists a non-negative, integrable function $f : \mathbb{R}^n \to \mathbb{R}$ such that
\[ F(x) = F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n) \, dy_1 \cdots dy_n, \]
then $f$ is called the (probability) density function (PDF) for $X$.

It follows, in this case, that all probabilities related to $X$ can be expressed as integrals on $\mathbb{R}^n$ using the density function:
\[ P_X(B) = P(X \in B) = \int_B f(x) \, dx \quad \text{for all } B \in \mathcal{B}. \]

Probability measures for which a density exists are called absolutely continuous. We shall only consider such probability measures in the sequel.

Note that if we work directly in the probability space $(\mathbb{R}^n, \mathcal{B}, P)$, we can take the random variable to be the identity and a probability density directly defines the probability.

Note also that a random variable $X$ induces a $\sigma$-algebra $\mathcal{F}_X$ on $\Omega$, which is defined by $\mathcal{F}_X = \{ X^{-1}(B) | B \in \mathcal{B} \}$. $\mathcal{F}_X$ is the smallest $\sigma$-algebra which makes $X$ measurable.

Examples:

1. The uniform distribution on interval $[a,b]$ is given by the PDF
\[ f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b], \\ 0 & \text{else}. \end{cases} \]

The associated distribution function is
\[ F(x) = \int_{-\infty}^{x} f(x) \, dx = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } x \in [a,b], \\ 1 & \text{if } x > b. \end{cases} \]

2. The normal or gaussian distribution is defined by a PDF of the form
\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \]
2.4. Expected value, variance. The integration on a probability space is similar to the definition of the Lebesgue integral. One starts by defining the integral for simple functions of the form \( X = \sum_i a_i \chi_{A_i} \), where \( \chi_{A_i} \) is the characteristic functions of the set \( A_i \in \mathcal{F} \), i.e. \( \chi_{A_i}(\omega) = 1 \) if \( \omega \in A_i \) and 0 else. Then we define

\[
\mathbb{E}(X) := \int X \, dP := \sum_i a_i P(A_i).
\]

Then, because any measurable functions is the limit of a sequence of simple functions, this definition can then be extended by taking limits to any random variable (which is a \( \mathcal{F} \)-measurable function). For vector valued random variables, the integration is performed component by component.

**Definition 7.** A random variable \( X \) is said integrable with respect to the probability measure \( P \), if \( \mathbb{E}(|X|) < +\infty \). Then the value \( \mathbb{E}(X) := \int X \, dP \) is called expected value (or expectation, or mean value) of the random variable \( X \).

If \( \mathbb{E}(|X|^2) < +\infty \), the value

\[
\mathbb{V}(X) = \mathbb{E}(|X - \mathbb{E}(X)|^2) = \int_\Omega |X - \mathbb{E}(X)|^2 \, dP \geq 0
\]

is called variance of the random variable \( X \), and

\[
\sigma(X) = \sqrt{\mathbb{V}(X)}
\]

is called standard deviation of the random variable \( X \).

The variance can be also expressed by \( \mathbb{V}(X) = \mathbb{E}(|X|^2) - \mathbb{E}(X)^2 \). Indeed

\[
\mathbb{V}(X) = \int_\Omega |X - \mathbb{E}(X)|^2 \, dP = \int_\Omega (|X|^2 - 2X \cdot \mathbb{E}(X) + |\mathbb{E}(X)|^2) \, dP = \mathbb{E}(|X|^2) - \mathbb{E}(X)^2.
\]

If the probability measure is absolutely continuous its density provides a convenient way for evaluation of expectations using the so-called transfer theorem.

**Theorem 3** (Transfer theorem). Let \( g \) be a measurable function of \( \mathbb{R}^n \) and \( X \) an \( n \)-dimensional random variable. Then, if \( f \) is the density of the law of \( X \)

\[
\mathbb{E}(g(X)) = \int_\Omega g(X) \, dP = \int_{\mathbb{R}^n} g(x) \, dP_X(x) = \int_{\mathbb{R}^n} g(x) f(x) \, dx.
\]

Formally \( dP_X(x) = f(x) \, dx \). If \( f \) depends on \( x \) the probability measure \( P_X \) is not translation invariant. This is highlighted by the notation \( dP_X(x) \).

**Proof.** Let us check the formula for positive simple random variables. The general case is then obtained using the appropriate limit theorems.

So let \( g = \sum_{i=1}^n a_i \chi_{A_i} \) be a positive simple function. Then

\[
g(X(\omega)) = \sum_{i=1}^n a_i \chi_{A_i}(X(\omega)) = \sum_{i=1}^n a_i \chi_{X^{-1}(A_i)}(\omega).
\]

Hence

\[
\mathbb{E}(g(X)) = \sum_{i=1}^n a_i P(X^{-1}(A_i)) = \sum_{i=1}^n a_i P_X(A_i) = \int_{\mathbb{R}^n} g(x) \, dP_X(x).
\]

The last definition is just the definition of the integral for simple functions. Moreover, if \( P_X \) has density \( f \), then by definition of the density

\[
\sum_{i=1}^n a_i P_X(A_i) = \sum_{i=1}^n a_i \int_{A_i} f(x) \, dx = \sum_{i=1}^n a_i \int_{\mathbb{R}^n} \chi_{A_i}(x) f(x) \, dx = \int_{\mathbb{R}^n} g(x) f(x) \, dx.
\]

□
The formula given by the transfer theorem will be used for actual computations. In particular for the variance
\[ \mathbb{V}(X) = \int_{\Omega} |X - \mathbb{E}(X)|^2 \, dP = \int_{\mathbb{R}^n} (x - \mathbb{E}(X))^2 f(x) \, dx. \]

The variance can help quantify the deviation of a random variable \( X \) from its mean:

**Proposition 4** (Chebyshev inequality). Assume \( \mathbb{E}(X^2) < +\infty \). Then for any \( \epsilon > 0 \)
\[ P\left( |X - \mathbb{E}(X)| \geq \epsilon \right) \leq \frac{\mathbb{V}(X)}{\epsilon^2}. \]

**Proof.** Denote by \( A = |X - \mathbb{E}(X)| \geq \epsilon \). Then
\[ \mathbb{V}(X) = \int_{\Omega} |X - \mathbb{E}(X)|^2 \, dP \geq \int_A |X - \mathbb{E}(X)|^2 \, dP \geq \int_A \epsilon^2 \, dP = \epsilon^2 P(A) \]
which gives the result. \( \square \)

### 2.5. Conditional probabilities and independence.

Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \( A \) and \( B \) be two events.

Knowing that some random sample point \( \omega \in \Omega \) is in \( A \) we are interested in obtaining the probability that \( \omega \in B \). This defines the conditional probability:

**Definition 8.** Let \( A \) an event of probability \( P(A) > 0 \). Then the probability of \( B \) given \( A \) is defined by
\[ P(B|A) = \frac{P(A \cap B)}{P(A)}. \]

Let us verify that \( P(\cdot|A) \) defines a probability measure on \((\Omega, \mathcal{F})\): As \( A \cap B \subset A \), \( P(A \cap B) \leq P(A) \). Hence \( 0 \leq P(B|A) \leq 1 \) and axiom (i) is verified. \( \Omega \cap A = A \) hence \( P(\Omega|A) = 1 \) and axiom (ii) is verified. If \( B_1, B_2, \ldots \) are disjoint, so are their intersections with \( A \) and axiom (iii) follows.

Two events \( A, B \) are independent if \( P(B|A) = P(B) \). Then by definition of the conditional probability \( P(B) = \frac{P(A \cap B)}{P(A)} \) and we get the more symmetric definition of the independence of \( A \) and \( B \).

**Definition 9.** Two events \( A \) and \( B \) are said to be independent if
\[ P(A \cap B) = P(A)P(B). \]

This definition extends to random variables:

**Definition 10.** We say that the random variables \( X_i : \Omega \to \mathbb{R}^n \), \( i = 1, \ldots, m \) are independent, if for all choices of Borel sets \( B_1, \ldots, B_m \subset \mathbb{R}^n \)
\[ P(X_1 \in B_1, \ldots, X_m \in B_m) = P(X_1 \in B_1) \cdots P(X_m \in B_m). \]

**Theorem 4.** The random variables \( X_i : \Omega \to \mathbb{R}^n \), \( i = 1, \ldots, m \) are independent, if and only if their distribution functions verify
\[ F_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = F_{X_1}(x_1) \cdots F_{X_m}(x_m) \quad \text{for all } x_1,\ldots,x_m \in \mathbb{R}^n. \]

If the random variables have densities this is equivalent to
\[ f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = f_{X_1}(x_1) \cdots f_{X_m}(x_m) \quad \text{for all } x_1,\ldots,x_m \in \mathbb{R}^n. \]

The marginal densities \( f_{X_i} \) are obtained from the joined density \( f_{X_1,\ldots,X_m} \) by integrating on \( \mathbb{R}^n \) over all the other variables, for example
\[ f_{X_1}(x_1) = \int f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \, dx_2 \ldots dx_m. \]

From this theorem follows the following important result:
Theorem 5. If $X_1, \ldots, X_m$ are independent real valued random variables with $\mathbb{E}(|X_i|) < +\infty$ then $\mathbb{E}(|X_1 \ldots X_m|) < +\infty$ and

$$\mathbb{E}(X_1 \cdots X_m) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_m).$$

Proof. The results is easy to prove by applying the previous theorem assuming that each $X_i$ is bounded and has a density:

$$\mathbb{E}(X_1 \cdots X_m) = \int_{\mathbb{R}^m} x_1 \cdots x_m f_{X_1, \ldots, X_m}(x_1, \ldots, x_m) \, dx_1 \cdots dx_m,$$

$$= \int_{\mathbb{R}^m} x_1 f_{X_1}(x_1) \cdots x_m f_{X_m}(x_m) \, dx_1 \cdots dx_m,$$

$$= \mathbb{E}(X_1) \cdots \mathbb{E}(X_m).$$

End of lecture 3.

Moreover for independent variables the variance of the sum is the sum of variances. This is known as Bienaymé’s equality:

Theorem 6 (Bienaymé). If $X_1, \ldots, X_m$ are independent real valued random variables with $\mathbb{V}(|X_i|) < +\infty$ then

$$\mathbb{V}(X_1 + \cdots + X_m) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_m).$$

Proof. This can be proved by induction. We prove it only the case of two random variables. Let $m_1 = \mathbb{E}(X_1), m_2 = \mathbb{E}(X_2)$. Then by linearity of the integral $m_1 + m_2 = \mathbb{E}(X_1 + X_2)$ and

$$\mathbb{V}(X_1 + X_2) = \int_{\Omega} (X_1 + X_2 - (m_1 + m_2))^2 \, dP,$$

$$= \int_{\Omega} (X_1 - m_1)^2 \, dP + \int_{\Omega} (X_2 - m_2)^2 \, dP + 2 \int_{\Omega} (X_1 - m_1)(X_2 - m_2) \, dP,$$

$$= \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2 \mathbb{E}(X_1 - m_1)\mathbb{E}(X_2 - m_2)$$

using the independence of the random variables and the previous theorem in the last line. We then get the desired result by noticing that $\mathbb{E}(X_1 - m_1) = \mathbb{E}(X_2 - m_2) = 0$.

Definition 11. Let $X$ and $Y$ be two square integrable real valued random variables, then their covariance is defined by

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

By linearity of the expected value, we easily get

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Looking at the proof of the Bienaymé equality we see that in the general case we have

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{Cov}(X, Y),$$

the last term vanishing if the two random variables are independent. A more precise measure of the linear independence of two random variables is given by the correlation coefficient defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$
3. Monte Carlo Simulation

3.1. Principle. We want to define a Monte Carlo algorithm to approximate some real number \( a \) which represents for example the value of an integral. To this aim, we need to construct a real valued random variable \( X \) such that

\[
\mathbb{E}(X) = a.
\]

Then we define an approximation by considering a sequence of independent random variables \((X_i)_i\) distributed like \( X \) and approximate \( \mathbb{E}(X) \) by the sample mean

\[
M_N = \frac{1}{N} \sum_{i=1}^{N} X_i.
\]

In order for this procedure to be useful, we need first to be able to recast our problem in the form of the computation of an expected value of an adequate random variable \( X \) that we need to define. Then we need to be able to draw independent variables distributed like \( X \) and finally we need to check that the approximation we defined converges in some sense to the exact value and possibly estimate the speed of convergence.

Here the sample mean is an example of what is called an estimator in statistics, which is a rule for computing some statistical quantity, which is a function of the random variable, here the expected value, from sample data.

**Definition 12.** The difference between the expected value of the estimator and the statistical quantity it approximates is called bias. If this difference is zero, the estimator is said to be unbiased.

Let us compute the bias of the sample mean given by (8), we easily get as the \( X_i \) are all distributed like \( X \) and thus have the same expected value that

\[
\mathbb{E}(M_N) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(X_i) = \mathbb{E}(X)
\]

so that the bias is zero and our sample mean is unbiased.

Assuming the sample number \( N \geq 2 \) an unbiased estimator of the variance is given by the following sample variance

\[
V_N = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - M_N)^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \frac{1}{N} \sum_{i=1}^{N} X_i \right)^2.
\]

Indeed, let us compute the expected value of \( V_N \). Denoting by \( a = \mathbb{E}(X_i) \) for \( i = 1, \ldots, N \), we have

\[
V_N = \frac{1}{N-1} \sum_{i=1}^{N} ((X_i - a) + (a - M_N))^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - a)^2 - \frac{N}{N-1} (M_N - a)^2,
\]

as \( 2 \sum_{i=1}^{N} (X_i - a)(a - M_N) = -2N(M_N - a)^2 \). Hence

\[
\mathbb{E}(V_N) = \frac{1}{N-1} \sum_{i=1}^{N} \mathbb{E}((X_i - a)^2) - \frac{N}{N-1} \mathbb{E}((M_N - a)^2) = \frac{1}{N-1} \sum_{i=1}^{N} \mathbb{V}(X_i) - \frac{N}{N-1} \mathbb{V}(M_N).
\]

And because of Bienaymè’s theorem

\[
N^2 \mathbb{V}(M_N) = \mathbb{V}(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} \mathbb{V}(X_i) = N \mathbb{V}(X).
\]

Hence

\[
\mathbb{E}(V_N) = \frac{N}{N-1} \mathbb{V}(X) - \frac{1}{N-1} \mathbb{V}(X) = \mathbb{V}(X).
\]
Remark 2. Note the $1/(N-1)$ factor in the variance estimator instead of the $1/N$ that one would expect at the first glance. Using $1/N$ instead would also yield an estimator of the variance, but this one would be biased, i.e. it would not have the right expected value.

End of lecture 4.

3.2. Estimation of the error in a Monte Carlo simulation. Let us first compute in a general way the mean square error (MSE) of an estimator. The MSE is defined by

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = \int (\hat{\theta} - \theta)^2 dP.$$ 

Note that the root mean square error or RMS error, which is the square root of the MSE, is the classical $L^2$ error.

Assume $\hat{\theta}$ is an estimator for the statistical quantity $\theta$ which is a real number that can be computed as a function of a random variable $X$.

Lemma 1. Assume the random variable $\hat{\theta}$ is an estimator for $\theta$ and $\mathbb{E}(\hat{\theta}^2) < +\infty$. Then

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{V}((\hat{\theta})^2 + Bias(\hat{\theta})^2).$$

Proof. A straightforward calculation yields

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{E}(\hat{\theta}^2) - 2\mathbb{E}(\hat{\theta})\theta + \theta^2 = \mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2 + \theta^2 - 2\mathbb{E}(\hat{\theta})\theta = (\mathbb{E}^2(\hat{\theta}) - \mathbb{E}(\hat{\theta})^2) + (\mathbb{E}(\hat{\theta}) - \theta)^2$$

$$= \mathbb{V}(\hat{\theta}) + (Bias(\hat{\theta}))^2.$$ 

Assume that the random variable $X$ defining our Monte Carlo simulation verifies $\mathbb{E}(X^2) < +\infty$. Then we can apply the previous lemma to $M_N$ as an estimator of $\mathbb{E}(X)$, which yields

$$MSE(M_N) = \mathbb{V}(M_N) + (\mathbb{E}(M_N) - \mathbb{E}(X))^2.$$ 

So the RMS error is composed of two parts, the error coming from the variance of the sample and the possible bias on the sample occurring when the expected value of $M_N$ is not exactly equal to the expected value of the random variable $X$ being approximated.

In many cases the bias can be made to be zero, but in some cases it can be useful to introduce some bias in order to decrease the variance of the sample and the total error.

Lemma 2. Assume $\mathbb{E}(X^2) < +\infty$. Then the RMS error for an unbiased simulation based on the random variable $X$ is

$$e_{rms} = \sigma(M_N) = \frac{\sigma(X)}{\sqrt{N}}.$$ 

Proof. The formula (10) giving the mean squared error of an estimator shows that if the simulation is unbiased $\mathbb{E}(M_N) = \mathbb{E}(X)$ and

$$e_{rms} = \sqrt{\mathbb{V}(M_N)} = \sigma(M_N).$$ 

Now using Bienaymé’s theorem we also have

$$N^2\mathbb{V}(M_N) = \mathbb{V}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \mathbb{V}(X_i) = N\mathbb{V}(X).$$

And thus $\mathbb{V}(M_N) = \mathbb{V}(X)/N$, which gives the result.
On the other hand, Chebyshev’s inequality gives us, assuming $\mathbb{E}(X^2) < +\infty$ that for any $\epsilon > 0$ we have, as $\mathbb{E}(X) = \mathbb{E}(M_N)$

$$P(\left| M_N - \mathbb{E}(X) \right| \geq \epsilon) \leq \frac{\mathbb{V}(M_N)}{\epsilon^2} = \frac{\sigma^2(X)}{N\epsilon^2}.$$ 

Hence when $N \to +\infty$, we have that

$$P(\left| M_N - \mathbb{E}(X) \right| \geq \epsilon) \to 0.$$ 

This means that $M_N$ converges to $\mathbb{E}(X)$ in probability. This is called the weak law of large numbers. The corresponding strong law of large numbers, the proof of which is more involved, states that $M_N$ converges to $\mathbb{E}(X)$ almost surely, which means that

$$P \left[ \omega \mid \lim_{N \to +\infty} M_N(\omega) = \mathbb{E}(X) \right] = 1.$$ 

The law of large numbers, strong or weak, implies that the sample mean converges towards the desired expected value, which justifies the Monte Carlo method.

Another major theorem of probability theory, the central limit theorem, gives a precise estimation of the error committed by an approximation. It claims that

$$\lim_{N \to +\infty} P \left[ \frac{M_N - \mathbb{E}(X)}{\sigma(X)/\sqrt{N}} \leq \lambda \right] = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-u^2/2} du.$$ 

This tells us that the asymptotic distribution of $M_N = \mathbb{E}(X)/\sqrt{N}$ is a unit normal distribution, or equivalently that $M_N$ is a normal distribution with mean $\mathbb{E}(X)$ and standard deviation $\sigma(X)/\sqrt{N}$.

The right hand side of (11) is a number that can be computed explicitly, and that is called confidence coefficient. For $\lambda = 3$ the confidence coefficient is 0.9973 and for $\lambda = 4$ the confidence coefficient is 0.9999 (see e.g. [5] for other values). This is the probability that the true mean lies in the so-called confidence interval $[M_N - \lambda\sigma(X)/\sqrt{N}, M_N + \lambda\sigma(X)/\sqrt{N}]$. Note that as opposite to deterministic error estimates, which are generally of the form $h^p$ or $1/N^p$, where $h$ is a cell size and $N$ a number of discretisation points, and lie on a deterministic curve. The error estimate in a Monte Carlo method is random, but it is always a normal distribution with variance which tends to 0 when the number of sample points tends to $+\infty$. In practice a good estimate of the error is given by $\sigma(X)/\sqrt{N}$, which is all the more interesting that the variance (or standard deviation) can be well estimated by the sample variance (or sample standard deviation), which is an a posteriori estimate that can be directly used in actual computations to measure the error.

### 3.3. Error monitoring in PIC codes.

In order to check the validity of simulation, it is important to monitor the evolution of some key quantities. In particular quantities that are conserved in the continuous model should be computed and the accuracy with which they are conserved will give a good indicator of the accuracy of the code: For the Vlasov Poisson system, key conserved quantities are total number of particles $N = 1 = \int f \, dx \, dv$, total momentum $P = \int fv \, dx \, dv$ and total energy $E = \frac{1}{2} \int f v^2 \, dx \, dv + \frac{1}{2} \int f\rho \phi \, dx$.

In our Monte Carlo approximation, assuming the particles are distributed according to the PDF $f$, we have

$$N = \mathbb{E}(1), \quad P = \mathbb{E}(V), \quad E = \mathbb{E}(\frac{1}{2}(V^2 + \phi(X))).$$

$\mathbb{E}(1) = \frac{1}{N}N = 1$ is conserved by construction, so there is nothing to monitor. For the others we can compute for a given initial condition the error due to sampling and for subsequent time steps the sample mean and sample standard deviation divided by $\sqrt{N}$ can be monitored to give a measure of the error. This can of course be compared to the error given by the actual sample with respect to the conserved value known from the initial condition.
Example. Consider a Landau damping initial condition, on a 1-periodic interval in $x$:

$$f_0 = (1 + \alpha \cos(kx)) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad (\alpha < 1).$$

The random variable $(X, V)$ is then randomly drawn according to this distribution. We can then compute

$$\mathbb{E}(V) = \int v f_0(x, v) \, dx \, dv = 0, \quad \mathbb{V}(V) = \int v^2 f_0(x, v) \, dx \, dv - 0 = 1.$$

So the RMS error committed by approximating $\mathcal{P}$ by the unbiased estimator $\mathcal{P}_N = \frac{1}{N} \sum_{i=1}^{N} V_i$ will be $1/\sqrt{N}$.

Let us now consider the total energy. First for the kinetic energy, we need to compute

$$\mathbb{E}(V^2) = \int v^2 f_0(x, v) \, dx \, dv = 1, \quad \mathbb{V}(V^2) = \int v^4 f_0(x, v) \, dx \, dv - \mathbb{E}(V^2)^2 = 3 - 1 = 2.$$

On the other hand he potential $\phi$ associated to the initial condition is solution of $\phi'' = \alpha \cos(kx)$. Assuming 0 average, we get $\phi(x) = -\frac{\alpha}{k^2} \cos(kx)$. We then can compute

$$\mathbb{E}(\phi(X)) = -\frac{\alpha}{k^2} \int \cos(kx) f_0(x, v) \, dx \, dv = -\frac{\alpha^2}{2k^2},$$

$$\mathbb{V}(\phi(X)) = \frac{\alpha^2}{k^4} \int \cos^2(kx) f_0(x, v) \, dx \, dv - \frac{\alpha^4}{4k^4} = \frac{\alpha^2(2 - \alpha^2)}{4k^4}.$$

It follows that the total energy of the initial condition is

$$\mathcal{E} = \mathbb{E} \left( \frac{1}{2} (V^2 + \phi(X)) \right) = \frac{1}{2} - \frac{\alpha^2}{4k^2},$$

and as the random variable $V$ and $\phi(X)$ are independent, the variance of the total energy is the sum of the variances of the kinetic energy and the potential energy.

A natural estimator for the energy based on the sample $(X_i, V_i)_{1 \leq i \leq N}$, distributed like $(X, V)$, used for the Monte Carlo simulation is here

$$\mathcal{E}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} (V_i^2 + \phi(X_i)),$$

from which it easily follows that $\mathbb{E}(\mathcal{E}_N) = \mathbb{E}(\mathcal{E})$ so that the estimator is unbiased. Moreover we can compute the variance of the estimator using Bienaymé’s equality

$$\mathbb{V}(\mathcal{E}_N) = \frac{1}{N} \mathbb{V}(\frac{1}{2} (V^2 + \phi(X)))/N = \frac{1}{4N} \left( \mathbb{V}(V^2) + \mathbb{V}(\phi(X)) \right)/N = \frac{1}{4N} \left( 2 + \frac{\alpha^2(2 - \alpha^2)}{4k^4} \right),$$

which is also the MSE error of the estimator as the simulation is unbiased.

After the initial time step, the exact distribution is not known, so that only empirical estimations can be computed. In order to monitor the noise (or error) on each computed quantity $\theta(X)$, we define the relative error

$$R = \frac{\sigma(\theta_N)}{\theta_N},$$

which is the inverse ratio of the estimated value and its standard deviation. We have

$$R = \frac{1}{N-1} \frac{\sqrt{\theta^2 - \bar{\theta}^2}}{\bar{\theta}},$$

where $\bar{\theta}(X) = \frac{1}{N} \sum_{i=1}^{N} \theta(X_i)$.
3.4. Error on the probability density. A standard way of estimating a probability density in \( \mathbb{R}^d \) from a sample is the kernel density estimator. It relies on a kernel which we shall call \( S_d \), which is a real function in \( \mathbb{R}^d \), that we shall assume to be the product of \( d \) identical functions: 
\[
S_d(x_1, x_2, \ldots, x_d) = S(x_1)S(x_2) \ldots S(x_d)
\]
verifying \( \int S(x) \, dx = 1 \) and \( S(x) = S(-x) \) which implies \( \int xS(x) \, dx = 0 \).

We then define for a given density \( f : \mathbb{R}^d \rightarrow \mathbb{R} \)
\[
f_h(x_1, \ldots, x_d) = \frac{1}{h^d} \int S \left( \frac{x_1 - y_1}{h} \right) \ldots S \left( \frac{x_d - y_d}{h} \right) f(y_1, y_2, \ldots, y_d) \, dy_1 \ldots dy_d,
\]
where \( (Y_1, \ldots, Y_d) \) are distributed according to the density \( f \). From this we can define the following estimator for \( f(x_1, \ldots, x_d) \), we also use the fact that \( S \) is even:
\[
f_{h,N}(x_1, \ldots, x_d) = \frac{1}{Nh^d} \sum_{i=1}^{N} S \left( \frac{Y_{1,i} - x_1}{h} \right) \ldots S \left( \frac{Y_{d,i} - x_d}{h} \right).
\]

As usual to estimate the mean squared error committed by this estimator, we compute its bias and variance
\[
\text{Bias}(f_{h,N}(x_1, \ldots, x_d)) = \mathbb{E}(f_{h,N}(x_1, \ldots, x_d)) - f(x_1, \ldots, x_d)
\]
\[
= \frac{1}{h^d} \mathbb{E} \left( S \left( \frac{Y_1 - x_1}{h} \right) \ldots S \left( \frac{Y_d - x_d}{h} \right) \right) - f(x_1, \ldots, x_d)
\]
\[
= \frac{1}{h^d} \int S \left( \frac{y_1 - x_1}{h} \right) \ldots S \left( \frac{y_d - x_d}{h} \right) f(y_1, \ldots, y_d) \, dy_1 \ldots dy_d - f(x_1, \ldots, x_d)
\]
\[
= \int S(z_1) \ldots S(z_d) f(x_1 + hz_1, \ldots, x_d + hz_d) \, dz_1 \ldots dz_d - f(x_1, \ldots, x_d),
\]
where \( H(f) = (\frac{\partial^2 f}{\partial x_i \partial x_j})_{1 \leq i, j \leq d} \) is the Hessian matrix of \( f \) and \( z = (z_1, \ldots, z_d)^T \). Because of the symmetry of \( S \), the terms in \( h \) as well as the off-diagonal second order terms and the third order terms vanish. Hence the bias can be written
\[
\text{Bias}(f_{h,N}(x_1, \ldots, x_d)) = \int S(z_1) \ldots S(z_d) h(z_1) \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_d) dz_1 \ldots dz_d + \frac{h^2}{2} z^T H(f) z + O(h^4),
\]
where \( H(f) = (\frac{\partial^2 f}{\partial x_i \partial x_j})_{1 \leq i, j \leq d} \) is the Hessian matrix of \( f \) and \( z = (z_1, \ldots, z_d)^T \). Because of the symmetry of \( S \), the terms in \( h \) as well as the off-diagonal second order terms and the third order terms vanish. Hence the bias can be written
\[
\text{Bias}(f_{h,N}(x_1, \ldots, x_d)) = \frac{h^2}{2} \kappa_2(S) \Delta f(x_1, \ldots, x_d) + O(h^4),
\]
where \( \kappa_2(S) = \int x^2 S(x) \, dx \) is the second order moment of the kernel \( S \) and \( \Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_d^2} \) the Laplace operator. We note that the bias depends only on \( h \) the width of the kernel, and not on the number of particles \( N \). It goes to zero when \( h \) goes to 0.

Let us now compute the variance of the estimator. With Bienaymé’s equality we get
\[
\mathbb{V}(f_{h,N}(x_1, \ldots, x_d)^2) = \frac{1}{N} \mathbb{V} \left( \frac{1}{h^d} S \left( \frac{Y_1 - x_1}{h} \right) \ldots S \left( \frac{Y_d - x_d}{h} \right) \right).
\]
Higher order B-splines are then defined by:

\[
MSE(14) = \frac{1}{h^d} \int S^2(z_1) \ldots S^2(z_d) f(x_1 + hz_1, \ldots, x_d + hz_d) \, dz_1 \ldots dz_d = \frac{1}{h^d} \int S^2(z_1) \ldots S^2(z_d) \, dz_1 \ldots dz_d(f(x_1, \ldots, x_d) + O(h)) = \frac{1}{h^d} R(S)^d(f(x_1, \ldots, x_d) + O(h))
\]

where the \( R(S) = \int S^2(x) \, dx \) is called the roughness of the kernel \( S \). On the other hand, using the previous computation of the bias and the fact that \( \int S(x) \, dx = 1 \), we have

\[
\mathbb{E}(\frac{1}{h^d} S \left( \frac{Y_1 - x_1}{h} \right) \ldots S \left( \frac{Y_d - x_d}{h} \right))^2 = (f(x_1, \ldots, x_d) + O(h))^2.
\]

When \( h \to 0 \) this term can be neglected compared to the other contribution to the variance. Hence

\[
(13) \quad \mathbb{V}(f_{h,N}(x_1, \ldots, x_d)) = \frac{R(S)^d}{Nh^d} f(x_1, \ldots, x_d) + O(\frac{1}{N}).
\]

And finally, the mean squared error of the estimator is the sum of its variance and squared bias, which yields

\[
(14) \quad MSE(f_{h,N}(x_1, \ldots, x_d)) = \frac{R(S)^d}{Nh^d} f(x_1, \ldots, x_d) + \frac{h^d}{4} \kappa_2(S)(\Delta f)^2(x_1, \ldots, x_d) + O(N^{-1}) + O(h^6).
\]

Note that for the MSE to converge, one needs obviously the number of samples \( N \to +\infty \), \( h \to 0 \), but also \( Nh^d \to +\infty \) for the first term to tend to 0. As \( h^d \) is a measure of the cell size in a \( d \)-dimensional space, this means that the number of particles per cell needs to converge to \( +\infty \). In general in PIC methods, one is not really interested in the convergence of the distribution function, but it is essential to have a good convergence of the density in physical space. For this reason, one generally imposes the number of particles per cell in physical space to be large enough, and all the larger that the cells become smaller. Keeping the number of particles per cell constant when \( h \) decreases does not yield convergence of the method.

To get a unique parameter yielding an order of convergence, one can minimise the dominating terms of \( MSE(f_{h,N}(x_1, \ldots, x_d)) \) with respect to \( h \), yielding an expression of \( h \) in function of \( N \).

Standard kernels in statistics beyond the top hat kernel, are the Gaussian kernel and Epanechnikov type kernels of the form \( S(x) = c_s(1 - x^2)^s \) for \( |x| < 1 \) and 0 else, where \( c_s \) is a normalisation constant insuring that \( \int S(x) \, dx = 1 \). \( s \) is a small integer, typically 1.2 or 3 giving the smoothness of the kernel.

In PIC codes \( S \) is generally chosen to be a spline function. A spline function of degree \( m \) is a piecewise polynomial of degree \( m \) and which is in \( C^{m-1} \). It can be defined by recurrence: The degree 0 B-spline that we shall denote by \( S^0 \) is defined by

\[
S^0(x) = \begin{cases} 
1 & \text{if } -\frac{1}{2} \leq x < \frac{1}{2}, \\
0 & \text{else.}
\end{cases}
\]

Higher order B-splines are then defined by:

For all \( m \in \mathbb{N}^* \),

\[
S^m(x) = (S^0)^m(x),
= S^0 \ast S^{m-1}(x),
= \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} S^{m-1}(y) \, dy.
\]
In particular the degree 1 spline is

\[ S^1(x) = \begin{cases} 
(1 - |x|) & \text{if } |x| < 1, \\
0 & \text{else},
\end{cases} \]

the degree 2 spline is

\[ S^2(x) = \begin{cases} 
\frac{1}{2}(\frac{3}{2} - |x|)^2 & \text{if } \frac{1}{2} < |x| < \frac{3}{2}, \\
\frac{3}{4} - x^2 & \text{if } |x| < \frac{1}{2}, \\
0 & \text{else},
\end{cases} \]

the degree 3 spline is

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

3.5. Aliasing. In PIC codes, where the evolution of the density function given by the Vlasov equation, needs to be coupled with the computation of the electric field on a grid, aliasing which is inherent to sampling on a grid plays an important role in the choice of the kernel.

**Theorem 7 (Shannon).** If support of \( \hat{f} \) is included in \( [-\frac{\pi}{h}, \frac{\pi}{h}] \), then

\[ f(t) = \sum_{k=-\infty}^{+\infty} f(kh) \text{sinc} \left( \frac{\pi(t-kh)}{h} \right), \]

where \( \text{sinc}(t) = \frac{\sin(t)}{t} \) is called the sinus cardinal function.

This means that \( f \) is completely determined by sampling with uniform step \( h \) if it has bounded support in \( [-\frac{\pi}{h}, \frac{\pi}{h}] \). However the support of an arbitrary function is generally not in \( [-\frac{\pi}{h}, \frac{\pi}{h}] \). If the support is bounded, it is enough to take \( h \) small enough. If the support is not bounded but \( f \) tends to 0 fast enough at infinity one also gets a good approximation if \( h \) is small enough. The question is what happens when \( h \) is not small enough to get a good approximation of \( \hat{f} \) in \( [-\frac{\pi}{h}, \frac{\pi}{h}] \).

In the case when \( \text{supp}(f) \not\subset [-\frac{\pi}{h}, \frac{\pi}{h}] \), in the formula giving the Fourier transform of a sampled function

\[ \hat{f}_h(\omega) = \frac{1}{h} \sum_{n=-\infty}^{+\infty} \hat{f}(\omega - \frac{2n\pi}{h}). \]

the supports of \( \hat{f}(\omega - \frac{2n\pi}{h}) \) of different \( n \) will have a non empty intersection. In particular \( \hat{f}(\omega - \frac{2n\pi}{h}) \) intersects \( [-\frac{\pi}{h}, \frac{\pi}{h}] \) for \( |n| \geq 1 \). Which means that high frequencies will appear in a low frequency interval. This is called ‘aliasing.

In this case with the reconstruction formula of Shannon’s theorem

\[ \tilde{f}(t) = (g_h * f_h)(t) = \sum_{k=-\infty}^{+\infty} f(kh)g_h(t-kh), \]

whose Fourier is

\[ \hat{\tilde{f}}(\omega) = \hat{f}_h(\omega)\hat{g}_h(\omega) = h\hat{f}_h(\omega)\chi_{[-\frac{\pi}{h}, \frac{\pi}{h}]} = \chi_{[-\frac{\pi}{h}, \frac{\pi}{h}]} \sum_{k=-\infty}^{+\infty} \hat{f}(\omega - \frac{2n\pi}{h}) \]

which can be very different of \( \hat{f}(\omega) \) because of the high frequency contributions.

To suppress aliasing, \( f \) needs to be approximated by \( \hat{f} \) which is the closest function in \( L^2 \) whose Fourier transform is in \( [-\frac{\pi}{h}, \frac{\pi}{h}] \).
Due to Plancherel’s formula
\[
\| f - \hat{f} \|_2^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega) - \hat{\tilde{f}}(\omega)|^2 d\omega
\]
\[
= \frac{1}{2\pi} \int_{|\omega| > \frac{\pi}{h}} |\hat{\tilde{f}}(\omega)|^2 d\omega + \frac{1}{2\pi} \int_{|\omega| < \frac{\pi}{h}} |\hat{\tilde{f}}(\omega) - \hat{\tilde{f}}(\omega)|^2 d\omega.
\]
The distance between the two functions is minimal when the second integral vanishes. We hence take \( \hat{\tilde{f}} \) to be the restriction of \( \hat{f} \) to \([ -\frac{\pi}{h}, \frac{\pi}{h}] \), which writes
\[
\hat{\tilde{f}}(\omega) = \hat{f}(\omega) \chi_{[-\frac{\pi}{h}, \frac{\pi}{h}]}(\omega) = \frac{1}{h} \hat{f}(\omega) \hat{g}_h(\omega),
\]
with \( g_h(t) = \text{sinc} \frac{\pi t}{h} \). It then follows \( \hat{\tilde{f}} = \frac{1}{h} f \ast g_h \).

Using the sinc would thus suppress all aliasing problems. However it has the problem that its support in physical space is unbounded, which means that all particles would contribute to all grid points, which is very time consuming in practice. Such an algorithm can only be used in practice, when working directly in Fourier space and only a few Fourier modes are needed.

On the other hand, the Fourier transforms of the B-splines are
\[
S_m(k) = \text{sinc}^{m+1} \left( \frac{k}{2} \right),
\]
which means that \( S_m \) decays like \( 1/|k|^{m+1} \) in Fourier space, which is quite fast for quadratic or cubic splines, thus limiting the aliasing problems.

4. Initialisation of given PDF

A Monte Carlo simulation relies on a random sequence following some given probability law. Such a sequence can be generated from a uniform random sequence on \([0, 1]\). Obtaining a good approximations of uniform random sequence is a complex task, but some good solutions are given by libraries included with standard compilers or numerical software. We will rely on those. Let us just mention that a computer cannot generate a truly random sequence, but generates two kind of random sequences: the first one called pseudo-random has the objective to provide good approximations of truly random sequences and the other one called quasi-random is designed to fill in the interval as uniformly as possible, yielding a smaller variance.

Then having a good random generator for a uniform random sequence in \([0, 1]\), there are different ways to draw values for any other probability density function. Some are specific to a given form of PDF like normal distributions, other are limited to some class of PDF like products of 1D functions and others are very general. A large number of techniques is described in the book \([5]\). We will describe the techniques that are the most useful for PIC simulations.

4.1. Inversion of the CDF. Let \( F \) be the cumulative distribution function (CDF) of the random variable we wish to simulate.

**Proposition 5.** Assume \( F : [a,b] \to [0,1] \) is a strictly increasing function. Let \( U \) be a uniformly distributed random variable on \([0,1]\), then \( X = F^{-1}(U) \) is a real value random variable with distribution function \( F \).

**Proof.** Let \( x \in [a,b] \). Then \( F^{-1}(U) \leq x \iff U \leq F(x) \).

The distribution function of \( X \) is defined by
\[
F_X(x) = P(X \leq x) = P(U \leq F(x)) = F(x)
\]
as \( U \) has a uniform distribution. \( \square \)
In many cases $F$ can be inverted analytically and when $F(x)$ can be computed, it can be inverted numerically using a fine grid and the assumption that $F$ grows linearly between two grid points.

**Examples:**

1. Uniform distribution on $[a, b]$

   The uniform distribution on $[a, b]$ has the distribution function $F(x) = \frac{x-a}{b-a}$, and to get its inverse we solve the equation $y = F(x) = \frac{x-a}{b-a}$. The solution is
   
   $$x = a + (b-a)y = F^{-1}(y).$$

2. Numerical inversion of an analytically known distribution function $F$.

   This amounts for a given point $y$ which is obtained from a uniform distribution in $[0, 1]$ to compute $x$ such that $F(x) = y$, which means solving $y - F(x) = 0$. The most efficient way, in general, to do this numerically is Newton's method which computes $x$ as the limit of the iterations
   
   $$x_{n+1} = x_n - \frac{y - F(x)}{-F'(x)}.$$

3. Numerical inversion of a function known at discrete grid points.

   We assume the values of $F$ are known on a grid $a = x_0 < x_1 < \cdots < x_{N_x} = b$. Because an approximation is involved in interpolating the values between the grid points, rather than computing directly the inverse for each value $y$ given by the uniform random generator, we start by computing $F^{-1}(y_j)$ where $0 = y_0 < y_1 < \cdots < y_{N_y} = 1$ is a uniform grid of $[0, 1]$. This can be done very easily as $F$ is an increasing function using the following algorithm:

   - $F^{-1}(0) = a$, $F^{-1}(1) = a$, i=0
   - For $j = 1, \ldots, N_y - 1$
     - Find $i$ such that $F(x_{i-1}) < y_i \leq F(x_i)$ (while $(F(x_i) < y_j)$ do $i = i + 1$)
     - Interpolate $F^{-1}(y_j)$ linearly (in order to maintain that $F^{-1}$ is non decreasing between $F(x_{i-1})$ and $F(x_i)$).
   
   Once $F^{-1}(y_j)$ is known on the grid $0 = y_0 < y_1 < \cdots < y_{N_y} = 1$, for any $y$ drawn uniformly on $[0, 1]$, find $j$ such that $y_j \leq y < y_{j+1}$ and interpolate linearly $F^{-1}(y)$.

   **Remark 3.** This method can also be used when $F$ is analytically known by first computing its values on a fine grid. This is generally more efficient that Newton's method and most of the time accurate enough.

### 4.2. Acceptance-rejection method.

This also sometimes simply called the rejection method. Assume, we want to draw according to the PDF $f$ and we know how to draw from the PDF $g$ with $f(x) \leq cg(x)$ for some given constant $c$. If the support of $f$ vanishes outside of a compact set $F$ we can take for example $g$ uniform in $K$ and $c = \max(f/g)$.

The the rejection algorithm is the following

1. Draw $x$ from $g$
2. Draw a uniform random number on $[0, 1]$ $u$
3. If $u \leq f(x)/(cg(x))$, accept $x$,
4. else reject $x$ and start again from (1).

The rejection method is very general, but in order to be efficient the number of rejections should be held as small as possible and $cg$ chosen as close as possible to $f$, with the constraint of course that one needs to be able to draw from $g$. 
4.3. Composition method. This method is also known as the probability mixing method and can be used when the PDF that one wants to sample from is the sum of two simpler PDF. Given two PDF $f_1, f_2$ that we know how to sample from, and

$$f(x) = \alpha f_1(x) + (1-\alpha)f_2(x), \quad \text{with } \alpha < 1.$$  

A value $x$ can be sampled from $f$ by the following procedure

1. Select a random number $r_i$ from a uniform distribution on $[0,1]$,
2. If $r_i < \alpha_1$ draw $x_i$ according to the PDF $f_1$,
3. Else draw $x_i$ according to the PDF $f_2$.

This can be extended to the weighted sum of an arbitrary number of probability density functions. If

$$f(x) = \alpha_1 f_1(x) + \cdots + \alpha_n f_n(x), \quad \text{with } \alpha_1 + \cdots + \alpha_n = 1,$$

one can then draw from $f$ by drawing a random number $r$ from a uniform distribution on $[0,1]$ and then drawing from $f_i$ if $\alpha_0 + \cdots + \alpha_{i-1} < r < \alpha_0 + \cdots + \alpha_i, 1 \leq i \leq n$, denoting by $\alpha_0 = 0$.

5. Variance reduction techniques

As we saw the Monte Carlo error for the approximation of the expected value of a random variable $X$ is

$$\sigma(X) / \sqrt{N}.$$  

Apart from increasing the number of realisations $N$, the most efficient method to reduce the error is to use available information to replace $X$ by another random variable with the same expected value but a lower variance. We shall describe a few techniques to do that in the context of Particle in Cell methods.

5.1. Control variates. Consider the standard Monte Carlo problem of approximating $a = \mathbb{E}(X)$, for a given random variable $X$, by a sample mean.

Assume now that there exists a random variable $Y$ the expected value of which is known, that is somehow correlated to $X$. For a given $\alpha \in \mathbb{R}$, let us define the new random variable

$$Z_\alpha = X - \alpha(Y - \mathbb{E}(Y)).$$

Obviously, we have for any $\alpha$ that $\mathbb{E}(Z_\alpha) = \mathbb{E}(X) = a$, which means that the sample mean of $Z_\alpha$

$$M_{N,\alpha} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \alpha(Y_i - \mathbb{E}(Y))) = \alpha\mathbb{E}(Y) + \frac{1}{N} \sum_{i=1}^{N} (X_i - \alpha Y_i)$$

could be used instead of the sample mean of $X$ to approximate $a$. The random variable $\alpha Y$ is called a control variate for $X$.

End of lecture 5.

Let us now look under what conditions the variance of $Z_\alpha$ is lower than the variance of $X$. We assume that both $\mathbb{V}(X) > 0$ and $\mathbb{V}(Y) > 0$.

**Lemma 3.** If the random variables $X$ and $Y$ are not independent, there exists a value of $\alpha$ for which the variance of $Z_\alpha$ is smaller than the variance of $X$. More precisely

$$\min_{\alpha \in \mathbb{R}} \mathbb{V}(Z_\alpha) = \mathbb{V}(X)(1 - \rho^2(X,Y)) = \mathbb{V}(Z_{\alpha^*}), \quad \text{with } \alpha^* = \frac{\text{Cov}(X,Y)}{\mathbb{V}(Y)}.$$  

Moreover

$$\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \iff \begin{cases} \alpha < 2\alpha^* & \text{if } \alpha > 0, \\ \alpha > 2\alpha^* & \text{if } \alpha < 0. \end{cases}$$
PROOF. As $Z_\alpha = X - \alpha Y + \alpha \mathbb{E}(Y)$, and $\mathbb{E}(Z_\alpha) = \mathbb{E}(X)$ we have

$$\mathbb{V}(Z_\alpha) = \mathbb{E}(Z_\alpha^2) - \mathbb{E}(X)^2,$$

$$= \mathbb{E}((X - \alpha Y)^2) + 2\alpha \mathbb{E}(Y)\mathbb{E}(X - \alpha Y) + \alpha^2 \mathbb{E}(Y)^2 - \mathbb{E}(X)^2,$$

$$= \mathbb{E}(X^2) - 2\alpha \mathbb{E}(XY) + \alpha^2 \mathbb{E}(Y)^2 + 2\alpha \mathbb{E}(Y)\mathbb{E}(X) - 2\alpha^2 \mathbb{E}(Y)^2 + \alpha^2 \mathbb{E}(Y)^2 - \mathbb{E}(X)^2,$$

$$= \mathbb{V}(X) - 2\alpha \text{Cov}(X, Y) + \alpha^2 \text{Var}(Y),$$

$$= \sigma^2(X) - 2\alpha \sigma(X)\sigma(Y)\rho(X, Y) + \alpha^2 \sigma^2(Y),$$

introducing the standard deviation of a random variable $\sigma^2(X) = \mathbb{V}(X)$ and the correlation coefficient of two random variables $\rho(X, Y) = \text{Cov}(X, Y)/\sigma(X)\sigma(Y))$.

So the variance of $Z_\alpha$ is a second order polynomial in $\alpha$ the minimum of which is reached for

$$\alpha^* = \frac{\sigma(X)}{\sigma(Y)}\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma^2(Y)},$$

and plugging this into the expression of $\mathbb{V}(Z_\alpha)$, we get

$$\mathbb{V}(Z_{\alpha^*}) = \sigma^2(X) - 2\sigma(X)^2\rho(X, Y)^2 + \sigma^2(X)\rho(X, Y)^2 = \mathbb{V}(X)(1 - \rho^2(X, Y)).$$

On the other hand

$$\mathbb{V}(Z_\alpha) - \mathbb{V}(X) = \alpha \sigma(Y)(\alpha \sigma(Y) - 2\sigma(X)\rho(X, Y)).$$

Hence for $\alpha > 0$,

$$\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \Leftrightarrow \alpha < 2\frac{\sigma(X)}{\sigma(Y)}\rho(X, Y) = 2\alpha^*,$$

and for $\alpha < 0$, $\mathbb{V}(Z_\alpha) < \mathbb{V}(X) \Leftrightarrow \alpha > 2\alpha^*$.

**Remark 4.** This results means that provided $\text{Cov}(X, Y) \neq 0$, i.e. $X$ and $Y$ are not independent, there is always an interval around the optimal value $\alpha^*$ for which $Z_\alpha$ has a lower variance than $X$. The more correlated $X$ and $Y$ are, the larger this interval is. So the most important is to find a random variable $Y$ the expectation of which is known, that is as correlated with $X$ as possible. Then if a good approximation of $\text{Cov}(X, Y)$ can be computed, one can use this to get closer to $\alpha^*$ and minimise the variance as much as possible with the random variable $Y$.

A typical example is when $X = Y + \epsilon \tilde{Y}$, where $\epsilon$ is small and $\mathbb{E}(Y)$ is known and for simplicity $Y$ and $\tilde{Y}$ are independent. Plugging this in the expression of $\mathbb{V}(Z_\alpha)$ in the above proof yields

$$\mathbb{V}(Z_\alpha) = \mathbb{V}(Y) + \epsilon^2 \mathbb{V}(\tilde{Y}) - 2\alpha \mathbb{V}(Y) + \alpha^2 \mathbb{V}(Y) = (1 - \alpha)^2 \mathbb{V}(Y) + \epsilon^2 \mathbb{V}(\tilde{Y}).$$

So that taking $\alpha = 1$ yields that $\mathbb{V}(Z_\alpha)$ is of order $\epsilon^2$ assuming $\mathbb{V}(\tilde{Y})$ of order 1. This is typically the form that is used in PIC simulations.

**5.2. Importance sampling.** We are interested in computing, for some given probability density $f$, quantities of the form

$$\int \psi(z) f(z) \, dz.$$

The standard Monte Carlo method for doing this is to define our integral as an expected value using a random variable $Z$ of density $f$. Then

$$\int \psi(z) f(z) \, dz = \mathbb{E}(\psi(Z)).$$

Depending on the function $\psi$ it might not be the best approach to use directly the density $f$ for drawing the random variable used in the simulation. Indeed if $g$ is any other probability density
that does not vanish in the support of \( f \) one can express our integral as an expectation using a random variable \( \tilde{Z} \) of density \( g \):

\[
\int \psi(z) f(z) \, dz = \int \psi(z) \frac{f(z)}{g(z)} g(z) \, dz = \mathbb{E}(W(\tilde{Z}) \psi(\tilde{Z})),
\]

where the random variable \( W(\tilde{Z}) = \frac{f(\tilde{Z})}{g(\tilde{Z})} \) is called weight.

The Monte Carlo approximation using independent random variables distributed identically with density \( g \) can be expressed as

\[
\tilde{M}_N = \frac{1}{N} \sum_{i=1}^{N} W(\tilde{Z}_i) \psi(\tilde{Z}_i),
\]

from which we get

\[
\mathbb{E}(\tilde{M}_N) = \mathbb{E}(W(\tilde{Z}) \psi(\tilde{Z})) = \int \psi(z) f(z) \, dz.
\]

So \( \tilde{M}_N \) is another unbiased estimator of the integral we wish to compute and the approximation error for a given number of samples \( N \) is determined by its variance.

Let us now investigate how \( g \) can be chosen to get a smaller variance. For this we need to compare the variance of \( W(\tilde{Z}) \psi(\tilde{Z}) \) and the variance of \( \psi(Z) \) knowing that both have the same expected value.

\[
\mathbb{E}(W(\tilde{Z})^2 \psi(\tilde{Z})^2) = \int \psi(z)^2 W(z)^2 g(z) \, dz = \int \psi(z)^2 W(z) f(z) \, dz.
\]

On the other hand

\[
\mathbb{E}(\psi(Z)^2) = \int \psi(z)^2 f(z) \, dz.
\]

So we see that there is a factor \( W \) difference between the two expressions and obviously if \( W < 1 \) in regions where \( \psi \) is larger, the procedure will lead to a smaller variance. Note that because \( f \) and \( g \) both have an integral one, we cannot have \( W < 1 \) everywhere.

We also remark that, assuming that \( \psi(z) \) does not vanish, if we take \( W(z) = \mathbb{E}(\psi(Z))/\psi(z) \) which corresponds to \( g(z) = f(z) \psi(z)/\mathbb{E}(\psi(Z)) \), we get

\[
\mathbb{E}(W(\tilde{Z})^2 \psi(\tilde{Z})^2) = \mathbb{E}(\psi(Z)) \int \psi(z) f(z) \, dz = \mathbb{E}(\psi(Z))^2 = \mathbb{E}(W(\tilde{Z}) \psi(\tilde{Z}))^2
\]

so that \( \nabla(W(\tilde{Z}) \psi(\tilde{Z})) = 0 \). This of course cannot be done in practice as \( \mathbb{E}(\psi(Z)) \) is the unknown quantity we wish to approximate, but it can be used as a guideline to find a density \( g \) that reduces the variance as much as possible and tells us that the density \( g \) should be proportional to the integrand \( f \psi \), i.e. that markers should be distributed according to the integrand.

5.3. Application to the PIC method. For the PIC method, we can combine the importance sampling method and the control variates method.

5.3.1. Importance sampling. The choice of a density for importance sampling depends on the expected value that we are interested in. There are many of those in a PIC code, but arguably the accurate computation of the electric field, which determines the self-consistent dynamics is the most important. Depending on the physical problem we want to deal with more particles will be needed in some specific phase space areas, like for example in some region of the tail for a bump-on-tail instability. For this reason, it is interesting in a PIC code to have the flexibility of drawing the particles according to any density, but one needs to be careful with the choice of this density as the results can become better or worse.

Initialisation. Assume we know the density \( g_0 \) according to which we want to draw the markers. Then we initialise the marker’s phase space positions \( z^0_i = (x^0_i, v^0_i) \) as realisations of a random variable \( Z^0 \) with density \( g_0 \).
Time stepping. The markers evolve along the characteristics of the Vlasov equation so that at time \( t \) the random variable \( \mathbf{Z}_t = (\mathbf{X}_t, \mathbf{V}_t) \) is distributed according to the density \( g(t, \mathbf{z}) \), that is the solution of the Vlasov-Poisson equation with initial condition \( g_0 \).

Then as we saw, the different quantities we need to compute using the Monte Carlo approximation are of the form

\[
\int \psi(\mathbf{z}) f(t, \mathbf{z}) \, d\mathbf{z} = \int \psi(\mathbf{z}) \frac{f(t, \mathbf{z})}{g(t, \mathbf{z})} g(t, \mathbf{z}) \, d\mathbf{z} = \mathbb{E} \left( \psi(\mathbf{Z}) \frac{f(t, \mathbf{Z})}{g(t, \mathbf{Z})} \right)
\]

for some analytically known function \( \psi(\mathbf{z}) \). This means that we need to simulate the random variable \( Y_t = \psi(\mathbf{Z}_t) \frac{f(t, \mathbf{Z}_t)}{g(t, \mathbf{Z}_t)} = \psi(\mathbf{Z}_t)W \), where the random variable \( W \) is defined by \( W = f(t, \mathbf{Z}_t)/g(t, \mathbf{Z}_t) \).

Because \( f \) and \( g \) are conserved along the same characteristics we have

\[
W = \frac{f(t, \mathbf{Z}_t)}{g(t, \mathbf{Z}_t)} = \frac{f_0(\mathbf{Z}^0)}{g_0(\mathbf{Z}^0)},
\]

so that the random variable \( W \) does not depend on time and is set once for all at the initialisation.

Using importance sampling, we obtain the so-called weighted PIC method, in which the particles or markers are advance like in the standard PIC method, but have in addition an importance weight which does not evolve in time. The drawback of this method is that the variance can increase when large importance weights and small importance weights are mixed close together in phase space which often happens in long nonlinear simulations.

5.3.2. Control variates. We combine here control variates with importance sampling for most generality, but it can also be used without importance sampling by taking \( g_0 = f_0 \).

In the PIC method expected values of the form (15) cannot be exactly computed because the particle density in phase space \( f(t, \mathbf{z}) \) is not analytically known except at the initial time. However in many problems, e.g. Landau damping, bump-on-tail instability the distribution function stays close to an analytically known distribution function \( \tilde{f}(t, \mathbf{z}) \). Next to the control variate \( Y_t \) associated to \( f(t, \mathbf{z}) \), this can be used to build the control variate \( \bar{Y}_t \) associated to \( \tilde{f}(t, \mathbf{z}) \) such that

\[
Y_t = \psi(\mathbf{Z}) \frac{f(t, \mathbf{Z})}{g(t, \mathbf{Z})}, \quad \bar{Y}_t = \psi(\mathbf{Z}) \frac{\tilde{f}(t, \mathbf{Z})}{g(t, \mathbf{Z})}.
\]

Indeed we have

\[
\mathbb{E}(\bar{Y}_t) = \int \psi(\mathbf{z}) \frac{\tilde{f}(t, \mathbf{z})}{g(t, \mathbf{z})} g(t, \mathbf{z}) \, d\mathbf{z} = \int \psi(\mathbf{z}) \tilde{f}(t, \mathbf{z}) \, d\mathbf{z}
\]

which can be computed analytically for simple enough functions \( \psi \) and \( \tilde{f} \). Moreover if \( \tilde{f} \) is close enough to \( f \) then \( \bar{Y}_t \) will be close to \( Y_t \) and from the previous discussion a variance reduction of the order of the squared distance between the two random variables can be expected.

Let us now explain how this can be implemented in a PIC simulation.

Initialisation. As for importance sampling, the initial phase space positions of the markers are sampled as realisations \( (\mathbf{Z}_i^0)_{1 \leq i \leq N} \) of the random variable \( \mathbf{Z}^0 \) of density \( g_0 \). The importance weights are then defined by the corresponding realisations of the random variable \( W = f_0(\mathbf{Z}^0)/g_0(\mathbf{Z}^0) \), i.e. \( w_i = f_0(\mathbf{Z}_i^0)/g_0(\mathbf{Z}_i^0) \).

We also initialise the importance weights for \( \delta f = f - \tilde{f} \), which are defined by the random variable

\[
W_\alpha^0 = \frac{f_0(\mathbf{Z}^0) - \alpha \tilde{f}(t_n, \mathbf{Z}^0)}{g_0(\mathbf{Z}^0)} = W - \alpha \frac{\tilde{f}(0, \mathbf{Z}^0)}{g_0(\mathbf{Z}^0)}.
\]

Time stepping. The markers \( \mathbf{Z} \) are advanced by numerically solving the characteristics of the Vlasov equation. This means that given their positions \( \mathbf{Z}^n \) at time \( t_n \), an ODE solver is used to compute an approximation of their position \( \mathbf{Z}^{n+1} \) at time \( t^{n+1} \). Because \( f \) and \( g \) satisfy the
same Vlasov-Poisson equation, they are conserved along the same characteristics so that, as for importance sampling

\[ W = \frac{f(t_n, \mathbf{Z}^n)}{g(t_n, \mathbf{Z}^n)} = \frac{f_0(\mathbf{Z}^0)}{g_0(\mathbf{Z}^0)} \]

is a random variable which does not depend on time. On the other hand, we know \( \tilde{f} \) analytically and know that \( f \) and \( g \) are conserved along the characteristics, so that we can compute the importance weight for \( \delta f \) at time \( t_n \) from the phase space positions of the markers at the same time:

\[ W^n_\alpha = \frac{f(t_n, \mathbf{Z}^n) - \alpha \tilde{f}(t_n, \mathbf{Z}^n)}{g(t_n, \mathbf{Z}^n)} = \frac{f_0(\mathbf{Z}^0) - \alpha \tilde{f}(t_n, \mathbf{Z}^n)}{g_0(\mathbf{Z}^0)} = W - \alpha \frac{\dot{f}(t_n, \mathbf{Z}^n)}{g_0(\mathbf{Z}^0)}. \]

So \( W^n_\alpha \) is a time dependent random variable which can be computed explicitly using the analytical functions \( \tilde{f}, f_0 \) and \( g_0 \). These values can be used to express the sample mean for the new simulated random variable \( \tilde{Y}_\alpha = Y - \alpha(\tilde{Y} - \mathbb{E}(\tilde{Y})) \). This is defined by

\[ M^n_{\alpha,N} = \frac{1}{N} \sum_{i=1}^{N} (Y^n_i - \alpha \tilde{Y}_i^n) + \alpha \mathbb{E}(\tilde{Y}). \]

Plugging in the values for \( Y^n_i \) and \( \tilde{Y}_i^n \) we get

\[ M^n_{\alpha,N} = \frac{1}{N} \sum_{i=1}^{N} \left( \psi(\mathbf{Z}^n_i) \frac{f(t_n, \mathbf{Z}^n_i) - \alpha \tilde{f}(t_n, \mathbf{Z}^n_i)}{g(t_n, \mathbf{Z}^n_i)} \right) + \alpha \mathbb{E}(\tilde{Y}) = \frac{1}{N} \sum_{i=1}^{N} W^n_{\alpha,i} \psi(\mathbf{Z}^n_i) + \alpha \mathbb{E}(\tilde{Y}). \]

This yields an estimator for \( \psi(\mathbf{Z}) \) based on the weights \( W^n_\alpha \) and the expected value that can be computed analytically \( \mathbb{E}(\tilde{Y}) \). If no estimation of the optimal \( \alpha^* \) is available this method is used with \( \alpha = 1 \).

This is classically known as the \( \delta f \) method in the PIC literature [3, 1], as its interest lies in the expression \( f = \tilde{f} + \delta f \) with \( \tilde{f} \) known. A large variance reduction for \( \alpha = 1 \) is obtained as long as \( \delta f \ll \tilde{f} \), else one can also achieve some variance reduction by optimising for \( \alpha \) [4].

6. Coupling the Monte Carlo Vlasov solver with a grid based Poisson solver

The steps of the PIC algorithm are the following

1. Initialisation:
   (a) Draw markers \((\mathbf{x}_i, \mathbf{v}_i)\) according to the probability density \( g_0(\mathbf{x}, \mathbf{v}) \), if \( g_0 \) is not the initial particle distribution \( f_0 \) compute the importance weights \( w_i = f_0(\mathbf{x}_i, \mathbf{v}_i)/g_0(\mathbf{x}_i, \mathbf{v}_i) \).
   (b) Compute the initial electric field corresponding to the particles positions by solving the Poisson equation on a grid of physical space. For this a discrete value, depending on the Poisson solver being used, of the charge density \( \rho(t, \mathbf{x}) = 1 - \int f(t, \mathbf{x}, \mathbf{v}) \, d\mathbf{v} \) is needed.

2. Time stepping to go from \( t_n \) to \( t_{n+1} \):
   (a) Push the particles from \( t_n \) to \( t_{n+1} \) using the known discrete electric field. For this the electric field needs to be evaluated at the particle positions.
   (b) Compute the electric field corresponding to the new particle positions.

Next to the Monte Carlo solver for the Vlasov equation an important building block is the grid based Poisson solver and the interaction between the two. We shall distinguish here Poisson solvers needing values at discrete points like Finite Difference or spectral collocation methods and solvers using finite dimensional function spaces like Finite Elements which are coupled with markers in a different manner.

The two steps linked to the coupling, are on the one hand the computation of the discrete charge density needed by the Poisson solver from the particle positions and on the other hand the computation of the electric field at the particle positions.
6.1. Finite Difference PIC methods. We consider the 1D Poisson equation on the interval \([0, L]\)

\[-\Delta \phi = \rho = 1 - \int f(x, v) \, dv,\]

with periodic boundary conditions. This is well posed provided the average of \(\phi\), \(\int_0^L \phi(x) \, dx = 0\). We consider a uniform \(N_x\) points discretisation of the periodic interval \([0, L]\), \(x_j = j\Delta x = jL/N_x\) for \(0 \leq j \leq N_x - 1\). Because of the periodicity we have for any discrete function \((g_j)_{0 \leq j \leq N_x - 1}\) that \(g_{j+kN_x} = g_j\) for any \(k \in \mathbb{Z}\), where we denote by \(g_j\) an approximation of \(g(x_j)\). The standard second order centred Finite Difference for solving this equation reads

\[(16) \quad \frac{-\phi_{j+1} + 2\phi_j - \phi_{j-1}}{\Delta x^2} = \rho_j \text{ for } 0 \leq j \leq N_x - 1.\]

This yields a system of \(N\) equation with \(N\) unknowns. However all constant vectors are in the kernel of the associated matrix. Hence we need to set the constant to get a unique solution. This can be done thanks to the vanishing average hypothesis on \(\phi\), which implies on the discrete function that \(\sum_{j=0}^{N_x-1} \phi_j = 0\). A second order Finite Difference formula for computing the electric field then writes

\[(17) \quad E_j = -\frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} \text{ for } 0 \leq j \leq N_x - 1.\]

**Proposition 6.** Assume the electrostatic potential is computed from \((\rho_0, \ldots, \rho_{N_x-1})\) using \((16)\) and the electric field using \((17)\) with periodic boundary conditions. Then we have the following relations:

\[
\sum_{j=0}^{N_x-1} E_j \rho_j = 0, \quad \sum_{j=0}^{N_x-1} \phi_j \rho_j = \sum_{j=0}^{N_x-1} \frac{(\phi_{j+1} - \phi_j)^2}{\Delta x^2}.
\]

**Proof.** Using \((16)\) and \((17)\) we first compute

\[
\sum_{j=0}^{N_x-1} E_j \rho_j = -\sum_{j=0}^{N_x-1} \frac{(-\phi_{j+1} + 2\phi_j - \phi_{j-1})}{\Delta x^2} \frac{(\phi_{j+1} - \phi_{j-1})}{2\Delta x},
\]

\[
= \frac{1}{2\Delta x^2} \sum_{j=0}^{N_x-1} (\phi_{j+1}^2 - \phi_{j-1}^2 - 2\phi_j\phi_{j+1} + 2\phi_j\phi_{j-1}),
\]

\[
= 0.
\]

Indeed, by change of index, using the periodicity, we have

\[
\sum_{j=0}^{N_x-1} \phi_{j+1}^2 = \sum_{j=0}^{N_x-1} \phi_{j-1}^2 \text{ and } \sum_{j=0}^{N_x-1} \phi_j \phi_{j+1} = \sum_{j=0}^{N_x-1} \phi_j \phi_{j-1}.
\]

Now multiplying \((16)\) by \(\phi_j\) we get, using again periodicity and change of index

\[
\sum_{j=0}^{N_x-1} \phi_j \rho_j = -\sum_{j=0}^{N_x-1} \phi_j \frac{(\phi_{j+1} - \phi_j) - (\phi_j - \phi_{j-1})}{\Delta x^2},
\]

\[
= \sum_{j=0}^{N_x-1} \frac{(\phi_{j+1} - \phi_j)^2}{\Delta x^2}.
\]

\(\square\)
Remark 5. The properties in this proposition are discrete versions of the following properties verified by the continuous equations with periodic boundary conditions:

\[- \int_0^L \rho \nabla \phi \, dx = \int_0^L \Delta \phi \, dx = 0, \text{ and } \int_0^L \rho \phi \, dx = - \int_0^L \Delta \phi \, dx = \int_0^L (\nabla \phi)^2 \, dx.\]

These are necessary conditions for the conservation laws and to have them satisfied at the discrete level, one needs a discrete version of these. As we verified a standard centred second order scheme provides them, but there are many others, like higher order centred schemes or classical spectral Fourier schemes.

6.2. Finite Element PIC methods. Still for the 1D Poisson equation on the interval \([0, L]\)

\[ -\frac{d^2 \phi}{dx^2} = \rho = 1 - \int f(x, v) \, dv, \]

with periodic boundary conditions.

A variational formulation of this equation is obtained by multiplied by a smooth test function and integrating by parts the left hand side. Then the variational formulation reads:

Find \( \phi \in H^1_\sharp(0, L) \) such that

\[ \int_0^L \phi'(x) \psi'(x) \, dx = \int_0^L \rho(x)\psi(x) \, dx, \quad \forall \psi \in H^1_\sharp(0, L), \]

where we denote \( H^1_\sharp(0, L) \) the space of \( L \)-periodic functions with vanishing mean.

A Finite Element approximation, is a Galerkin approximation of (18), which means that we look for a function \( \phi_h \in V_h \), with \( V_h \) a finite dimensional subspace of \( H^1_\sharp(0, L) \), the test functions \( \psi_h \) also being in \( V_h \). Expressing the unknown functions \( \phi_h \) and the test functions \( \psi_h \) in the same finite dimensional basis of size \( N_x \), the variational formulation in the finite dimensional space is algebraically equivalent to a non singular linear system of size \( N_x \).

We consider now a Finite Element discretisation using the finite dimensional subspace of periodic spline functions of degree \( p \) on the uniform grid \( x_j = j \Delta x = jL/N_x \):

\[ S^p_h = \{ \phi_h \in C^{p-1}_\sharp(0, L) \mid \phi_h|_{[x_j, x_{j+1}]} \in \mathbb{P}^p([x_j, x_{j+1}]) \}, \]

where \( C^{p-1}_\sharp(0, L) \) is the space of \( L \)-periodic \( p-1 \) time continuously derivable functions and \( \mathbb{P}^p([x_j, x_{j+1}]) \) the space of polynomials of degree \( p \) on the interval \([x_j, x_{j+1}]\). Then a finite dimensional subspace of \( H^1_\sharp(0, L) \) is

\[ V_h = \{ \phi_h \in S^p_h \mid \int_0^L \phi_h(x) \, dx = 0 \}. \]

A basis of \( S^p_h \) can be defined using the B-splines of degree \( p \) on a uniform periodic grid of step \( \Delta x \). Those are defined by induction by the de Boor recursion formula \( S^0_j = 1 \) if \( j \Delta x \leq x < (j+1)\Delta x \) and 0 else. And for all \( p \in \mathbb{N}^* \),

\[ S^p_j(x) = \frac{x/j \Delta x - j S^{p-1}_j(x) + (j + p + 1) - x/j \Delta x S^{p-1}_{j+1}(x)}{p}. \]

From this definition, it also easily follows the formula for the derivative of a uniform B-spline

\[ \frac{d S^p_j(x)}{dx} = \frac{S^{p-1}_j(x) - S^{p-1}_{j+1}(x)}{\Delta x}. \]

Using this B-spline basis a function \( \phi_h \in V_h \) writes \( \phi_h = \sum_{j=0}^{N_x-1} \phi_j S^p_j(x) \), with \( \sum_{j=0}^{N_x-1} \phi_j = 0 \) so that the average of \( \phi \) vanishes. Plugging this into the variational formulation (18) with test
function \( \psi_h = S^p_k \) for \( k = 0, \ldots, N_x - 1 \) we get the Galerkin approximation of the Poisson equation

\[
\sum_{j=0}^{N_x-1} \phi_j \int_0^L S_j'(x) S_k'(x) \, dx = \int_0^L \rho(t,x) S_k(x) \, dx = 1 - \int_0^L \int_{-\infty}^{+\infty} f(t,x,v) S_k(x) \, dx \, dv.
\]

Now for a random variable \((X_t, V_t)\) having density \(g(t,x,v)\) and importance weight \(W\) with respect to the density \(f(t,x,v)\), we have

\[
\int_0^L \int_{-\infty}^{+\infty} f(t,x,v) S_k(x) \, dx \, dv = \mathbb{E}(WS_k(X)) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i)
\]

with our Monte Carlo approximation. Hence the Poisson equation we need to solve, with a source term coming from the Monte Carlo approximation for Vlasov becomes

\[
(21) \quad \sum_{j=0}^{N_x-1} \phi_j \int_0^L S_j'(x) S_k'(x) \, dx = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i).
\]

This yields the linear system \(K \tilde{\phi} = b\), the coefficients of \(K\) being \(\int_0^L S_j'(x) S_k'(x) \, dx\), the components of the column vector \(\tilde{\phi}\) being \(\phi_j\) and the components of the column vector \(b\) being \(\frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i)\).

The matrix \(K\) is called the stiffness matrix in Finite Element terminology. In our case because of the periodic boundary conditions, \(K\) is singular of rank \(N_x - 1\) as all constant vectors are in its kernel. To get a unique solution of the system we need the additional condition \(\sum_{j=0}^{N_x-1} \phi_j = 0\) (for translation invariant basis functions).

As opposed to the Finite Difference discretisation where we need an additional smoothing kernel, the Finite Element basis functions provide the needed regularisation naturally.

**Remark 6.** The Galerkin procedure also provides a kernel density estimate by projecting orthogonally in \(L^2\) the density on the span of \(S_j\), \(0 \leq j \leq N_x - 1\). This reads

\[
\sum_{j=0}^{N_x-1} \rho_j \int_0^L S_j(x) S_k(x) \, dx = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_k(x_i),
\]

or as a linear system \(M \tilde{\rho} = b\), where \(M\) is the matrix with coefficients \(\int_0^L S_j(x) S_k(x) \, dx\) \(\tilde{\rho}\) is the column vector of components \(\rho_j\) and \(b\) is defined as above. \(M\) is called mass matrix in the Finite Element terminology.

### 6.3. Conservation properties.

#### 6.3.1. Total number of particles.

The total number of particles, in the importance sampling formulation, is

\[
N = \int f(t,x,v) \, dx \, dv = \int \frac{f(t,x,v)}{g(t,x,v)} g(t,x,v) \, dx \, dv = \mathbb{E}(W) = 1,
\]

the importance weight being defined by \(W = f(t,X,V) / g(t,X,V)\). This is not depending on time as \(f\) and \(g\) evolve along the same characteristics.

Its estimator is defined by

\[
N_N = \frac{1}{N} \sum_{i=1}^{N} W_i.
\]

Because the weights \(W_i\) are drawn at the initialisation and not changed during the time stepping, \(N_N\) is exactly conserved during the simulation.
\[ \hat{N}_N^\alpha = \frac{1}{N} \sum_{i=1}^{N} W_{\alpha,i}^t + \alpha, \quad \text{with} \quad W_{\alpha,i}^t = W_i - \alpha \frac{\tilde{f}(t, X_i^t, V_i^t)}{g_0(X_0, V_0)}. \]

Here \( W_{\alpha,i}^t \) evolves in time and there is no reason why \( \hat{N}_N^\alpha \) should be exactly conserved. However, as the \( W_{\alpha,i}^t \) are identically distributed, we can verify the the estimator is unbiased:

\[ \mathbb{E}(\hat{N}_N^\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(W_{\alpha,i}^t) + \alpha = \mathbb{E}(W_i) + \alpha = \mathbb{E}(W_i) + \alpha(1 - \mathbb{E}\left( \frac{\tilde{f}(t, X_i^t, V_i^t)}{g_0(X_0, V_0)} \right)) = \mathbb{E}(W_i) = 1. \]

Indeed, as \( g \) is conserved along the characteristics

\[ \mathbb{E}\left( \frac{\tilde{f}(t, X_i^t, V_i^t)}{g_0(X_0, V_0)} \right) = \mathbb{E}\left( \frac{\tilde{f}(t, x, v)}{g(t, x, v)} \right) g(t, x, v) \, dx \, dv = \int \tilde{f}(t, x, v) \, dx \, dv = 1. \]

As it is important in practice to have exact conservation of the number of particles in order to avoid the build up of a spurious electric field which can become important in long times, we prefer to modify the estimator in order to have exact conservation of the number of particles. To this aim, we replace \( \bar{P}_i^t = \tilde{f}(t, X_i^t, V_i^t)/g_0(X_0, V_0) \) by

\[ \tilde{P}_i^t = 1 + \frac{\tilde{f}(t, X_i^t, V_i^t)}{g_0(X_0, V_0)} - \frac{1}{N} \sum_{j=1}^{N} \left( \frac{\tilde{f}(t, X_j^t, V_j^t)}{g_0(X_0, V_0)} \right). \]

As the \( (X_i^t, V_i^t) \) are identically distributed and the expected value is linear it follows that \( \mathbb{E}(\tilde{P}_i^t) = 1 \).

Then, for any function \( \psi \), the estimator

\[ \hat{M}_N^\alpha = \frac{1}{N} \sum_{i=1}^{N} (W_i - \alpha \bar{P}_i^t) \psi(Z_i^t) + \alpha \mathbb{E}(\bar{Y}). \]

is unbiased and the total number of particles is exactly conserved by construction as \( \frac{1}{N} \sum_{i=1}^{N} \bar{P}_i^t = 1 \).

### 6.4. Total momentum.

In order to couple this Poisson solver to our Monte Carlo Vlasov solver, we need a way to define \( \rho_j \), the value of the density at the grid points \( x_j \). This is classically done using a kernel density estimator \( S_{\Delta x} \), where \( S \) is any smooth function with integral one, typically in PIC simulations \( S \) are chosen to be B-spline functions, or tensor products of B-spline functions in several dimensions, and \( S_{\Delta x} = S(x/\Delta x)/\Delta x \). Then we get that

\[ (22) \quad \rho_j(t) = 1 - \int S_{\Delta x}(x_j - y) f(t, y, v) \, dy = 1 - \int S_{\Delta x}(x_j - y) w(t, y, v) g(t, y, v) \, dy \]

\[ = 1 - \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_{\Delta x}(x_j - x_i(t)) \]

where \( (X_i(t), V_i(t)) \) are random variables of density \( g(t, y, v) \), and the importance weight is \( W_j = f(t, X_i(t), V_i(t))/g(t, X_i(t), V_i(t)) \), which is one for standard PIC method where the particles density is \( f \). The realisations of the random variables are the particles, or markers, positions at time \( t \) that are computed by numerical integrating the characteristics of the Vlasov equation.

The reverse part of the coupling consists in obtaining consists in computing the electric field at the particle positions. The electric field field is defined by \( E(t, x) = -\frac{\partial \psi}{\partial x}(x, t) \) and can be obtained at the grid points \( x_j \) by finite differences of the potential, for example centred finite differences for second order accuracy \( E_j = \frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} \). Then to compute the electric field at the particle position a standard Lagrange interpolation could be used. However, in order to achieve conservation of total
momentum, we need to use a kernel smoothing using the same kernel as we use for computing \( \rho_j \). Hence we define for any \( x \in [0, L] \)

\[
E(x) = \sum_{j=0}^{N_x} E_j S_{\Delta x}(x - x_j).
\]

Note that in practice for a given \( x \) only very few \( j \) such that \( x - x_j \) are in the support of \( S_{\Delta x} \) have a non vanishing contribution.

**Proposition 7.** A PIC code for which \( \rho_j \) is computed from the particles using [22] and the electric field is computed at the particle position using [23] with the same smoothing kernel \( S_{\Delta x} \), with a velocity update with steps of the form \( v_{i}^{n+1} = v_i^n - hE(x_i) \) and such that the Poisson solver verifies \( \sum_{i=0}^{N_p} E_j \rho_j = 0 \) and \( \sum_{j=0}^{N_x} E_j \), conserves the total momentum, i.e.

\[
\frac{1}{N_p} \sum_{i=1}^{N_p} w_i v_i^{n+1} = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i v_i^n.
\]

**Proof.** Multiplying by \( w_i \) and summing over all the particles \( v_i^{n+1} = v_i^n - hE(x_i) \) becomes

\[
\frac{1}{N_p} \sum_{i=1}^{N_p} w_i v_i^{n+1} = \frac{1}{N_p} \sum_{i=1}^{N_p} w_i v_i^n - \frac{h}{N_p} \sum_{i=1}^{N_p} w_i E(x_i).
\]

We hence need to proof that \( \frac{1}{N_p} \sum_{i=1}^{N_p} w_i E(x_i) = 0 \). Using [23]

\[
\frac{1}{N_p} \sum_{i=1}^{N_p} w_i E(x_i) = \frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{j=0}^{N_x} w_i E_j S_{\Delta x}(x_i^n - x_j)
\]

\[
= \sum_{j=0}^{N_x} E_j \frac{1}{N_p} \sum_{i=1}^{N_p} w_i S_{\Delta x}(x_i^n - x_j)
\]

\[
= \sum_{j=0}^{N_x} E_j (1 - \rho_j^n)
\]

\[
= 0.
\]

using [22], \( \sum_{i=0}^{N_x} E_j \rho_j = 0 \) and \( \sum_{j=0}^{N_x} E_j \).

**Remark 7.** Total momentum conservation conservation implies that a particle cannot generate a force on itself as is physically correct. This is obtained by applying the proposition for only one particle. This is known as the self force issue that has raised many discussions in the Particle In Cell literature [4]. Obviously having only one particle is not useful in a non trivial Monte Carlo, however having an unbiased estimator of the total momentum is very important avoid non physical drifts. Following the above proof this is achieved if the electric field verifies \( E(WE(X)) = 0 \).

6.4.1. Total energy. At the continuous level, the total energy defined by

\[
\mathcal{E} = \frac{1}{2} \int v^2 f(t, x, v) \, dx \, dv + \frac{1}{2} \int \phi'(x)^2 \, dx,
\]

is conserved.

Taking \( \psi_h = \phi_h \) in the variational formulation [18], immediately yields

\[
\int \phi_h'(x)^2 \, dx = \int \rho \phi_h \, dx = - \int f(t, x, v) \phi_h \, dx \, dv = E(\phi_h(X)).
\]
So that

$$\mathcal{E}_h = \mathbb{E}(V^2/2) - \mathbb{E}(\phi_h(X)).$$
Bibliography


