Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 129 (2016) 50–68
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Data mining and probabilistic models for error estimate analysis of finite element method

Joël Chaskalovic[a,*], Franck Assous[b]

[a] *D'Alembert, University Pierre and Marie Curie, Paris, France*
[b] *Ariel University, 40700 Ariel, Israël*

Received 26 November 2014; received in revised form 29 December 2015; accepted 21 March 2016
Available online 24 May 2016

Abstract

In this paper, we propose a new approach based on data mining techniques and probabilistic models to compare and analyze finite element results of partial differential equations. We focus on the numerical errors produced by linear and quadratic finite element approximations. We first show how error estimates contain a kind of numerical uncertainty in their evaluation, which may influence and even damage the precision of finite element numerical results. A model problem, derived from an elliptic approximate Vlasov–Maxwell system, is then introduced. We define some variables as physical predictors, and we characterize how they influence the odds of the linear and quadratic finite elements to be locally "same order" accurate. Beyond this example, this approach proposes a method to compare, between several approximation methods, the accuracy of numerical results.

Keywords: Data mining; Probabilistic models; Error estimate; Finite element

1. Introduction

In many physical problems, we are limited first by our ability to measure observations and then to define the gap between the real problem and the mathematical model used to describe it. This is often referred as the modeling error. In previous papers [1,2], we have proposed to apply data mining techniques to evaluate this error. We relied on the fact that data mining techniques have already proved to be efficient in other contexts which deal with massive data, like in biology [23], medicine [28,27], marketing [25], advertising and communications [9,11].

When we attempt to simulate the problem numerically, we must identify the possible limitations of the numerical techniques employed. This is sometimes referred as the approximation error. In this article, using data mining techniques and probabilistic models, we extend our approach to compare numerical errors produced by different finite element approximations. Indeed, there is a need to investigate ways to quantify uncertainty and its impact, also in the case of numerical approximations, albeit supported by a mathematical theory, i.e. error estimate analysis.

Following [30] or [16], we distinguish between *errors* and *uncertainty* as follows: *errors* are detectable insufficiencies not due to lack of knowledge, whereas *uncertainties* are linked with lack of knowledge. As a consequence,

* Corresponding author.
E-mail address: jch1826@gmail.com (J. Chaskalovic).

http://dx.doi.org/10.1016/j.matcom.2016.03.013

one can consider errors as deterministic quantities, whereas uncertainties are inherently stochastic. This justified the probabilistic framework we will use in this article.

In that spirit, numerical simulations can be subject to uncertainty for instance in boundary conditions, geometry of the physical domain, etc., but they are also subject to uncertainty in their accuracy estimation. Indeed, an error estimate, even accurate, is an approximation to the actual unknown error, due to the presence of uncertainty that we will highlight in the following sections. Hence, uncertainty quantification of a given numerical method can be an interesting step towards its certification, in addition to the well known error estimate analysis.

For this purpose, we consider finite elements, and we aim at identifying a kind of stochastic behavior in finite element error estimates, which justify our stochastic approach later on. Accordingly, large databases that contain numerical approximations will be explored, in order to see whether and where precise finite elements are needed to guarantee accuracy. We will also show the limits of the approach, mainly due to the lack of the reference solution.

To illustrate our method, we introduce a model problem – a quasi-static Vlasov–Maxwell model – which models charged particle beams in plasma physics problems. We derive two numerical approximation methods, based on linear and quadratic finite elements [10], denoted respectively P_1 and P_2. Then, we compare the accuracy between the two implemented methods by modeling the dependency of the odds, which characterizes when P_1 and P_2 finite elements produce equivalent results.

More specifically, our objective is to propose data mining and probabilistic tools to ascertain situations where P_1 and P_2 finite elements lead to equivalent results. Our method will mine stored data of computed approximations to identify the predictors responsible of such a situation. However, at this point, we do not seek under what circumstances one leaves the P_2 finite element for the P_1 ones, to get accurate results for a given concrete application.

This article is organized as follows. In Section 2, we consider the global framework of a general elliptic variational formulation, and we highlight where and how a quantitative uncertainty appears in finite elements error estimates. Then, in Section 3, we introduce a model problem to illustrate our approach, namely a quasi-static paraxial Vlasov–Maxwell approximation. Numerical solutions are then computed by a P_1 and P_2 finite element Particle-In-Cell method. Section 4 will be devoted to the data mining and probabilistic models. After a brief presentation of the data mining tools involved, we will use them on our model problem approximations. Conclusions regarding equivalent numerical P_1 and P_2 approximations will be drawn.

2. Uncertainty in finite elements error estimate

We propose an approach based on statistical and probabilistic models to compare and analyze finite element results of partial differential equations. More precisely, we aim to explore large datasets in order to see whether high-order finite elements are required to guarantee accuracy. Our purpose in this section is to define "the same order" notion. Starting from finite element error analysis, we illustrate how two different finite element approximations can be of the "same order", due to a quantitative uncertainty appearing in the error estimates.

To begin with, let us recall some familiar notions regarding finite element error estimate of partial differential equations. Our main focus will be on a comparison of numerical errors produced by P_1 and P_2 finite element approximations (for more details see [3]). The aim is to show a kind of stochastic behavior in finite element error estimates, justifying a stochastic approach later on.

Since our study focuses mainly on space-dependent part of the model (not the time-dependent one, if any), we consider in what follows an elliptic standard problem, that can be viewed as the stationary problem associated with the time-dependent one. Indeed, when looking at a time-dependent problem, after time discretization, one generally solves a sequence of stationary problems, one for each time step. Hence, for a time-dependent problem, the method proposed here, derived at each time step, can be then accumulated over time.

Let V be a Hilbert space (with norm $\|.\|_V$). Throughout this section, $a(\cdot,\cdot)$ denotes a bilinear, continuous and V-elliptic form defined on $V \times V$, and $L(\cdot)$ a linear continuous form defined on V. We introduce the abstract elliptic variational formulation

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u, v) = L(v), \quad \forall v \in V. \end{cases} \tag{1}$$

Existence and uniqueness of a solution u of (1) is guaranteed by the Lax–Milgram theorem [12]. Let us introduce a finite dimension subset V_h of V, and consider the approximate solution u_h of u, that solves the approximate variational

formulation

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a(u_h, v_h) = L(v_h), \quad \forall v_h \in V_h. \end{cases} \tag{2}$$

Our aim is to estimate the error between the exact solution u and its approximation u_h. The first step is given by Céa's Lemma [12] that we recall here:

Lemma 2.1. *Let u be the solution to (1) and u_h its approximation, solution to (2). Then, the error $\|u - u_h\|_V$ is bounded by*

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V \,, \tag{3}$$

where C is a positive constant.

The constant C which appears in (3) is equal to the ratio between the continuity constant (continuity of $a(\cdot, \cdot)$), and the ellipticity constant (ellipticity of $a(\cdot, \cdot)$). Basically, C is unknown since the ellipticity constant cannot be accurately computed. It is part of the uncertainty that will appear later in the error estimate, and will create situations where accurate finite elements are required to guarantee a sufficient accuracy, or similarly, situations where more accurate finite elements do not add extra precision.

As a result of Céa's Lemma, the second step is to estimate $\inf_{v_h \in V_h} \|u - v_h\|_V$. For this purpose, the well-known technique consists in using standard interpolation estimates. Given a bounded and regular domain Ω subset of $\mathbb{R}^2$ in which the problem is set, we first introduce a regular finite element partition of Ω, assuming to coincide exactly with Ω. We denote by T a triangle of this mesh, and we first consider the linear P_1 finite element.

Classical finite element theorems (see for instance [3,12]) evaluate the distance between a sufficiently smooth function $v \in V$ and its interpolate on T, $\pi_T(v)$, that is the unique function of V_h that has the same values as u at the nodes of the partition. Applying them to the exact solution u of (1), one readily gets the following inequalities, assuming u is regular enough, and denoting $\mathbf{x}$ a point $(x, y) \in T$

$$\forall \mathbf{x} \in T, \quad |u(\mathbf{x}) - \pi_T(u)(\mathbf{x})| \leq 2\mathrm{diam}(T)^2 \ \|D^2(u)\|_{L^\infty(T)}, \tag{4}$$

$$\forall \mathbf{x} \in T, \quad |\nabla u(\mathbf{x}) - \nabla \pi_T(u)(\mathbf{x})| \leq 6\frac{\mathrm{diam}(T)^2}{\rho(T)} \ \|D^2(u)\|_{L^\infty(T)}. \tag{5}$$

Above, $\mathrm{diam}(T)$ denotes the diameter of the triangle T, $\rho(T)$ the roundness of T, i.e. the diameter of the biggest circle inside T, and $\|D^2(u)\|_{L^\infty(T)}$ is defined by

$$\|D^2(u)\|_{L^\infty(T)} = \max\left(\sup_{\mathbf{x}\in T} \left| \frac{\partial^2 u}{\partial x^2}(\mathbf{x}) \right|, \sup_{\mathbf{x}\in T} \left| \frac{\partial^2 u}{\partial y^2}(\mathbf{x}) \right|, \sup_{\mathbf{x}\in T} \left| \frac{\partial^2 u}{\partial x \partial y}(\mathbf{x}) \right| \right). \tag{6}$$

The local interpolation estimates (4) and (5), written in an element T of the mesh, are obtained by using Taylor's expansion with the Lagrange remainder, in which appear the second derivatives computed in an *unknown* point of T. In practice, exact values are generally impossible to obtain, and this remainder is overestimated by using, in each element T, the L_∞-norm, replacing the second derivatives by an upper bound as defined in (6). This *overestimation* is responsible of a *quantitative uncertainty*, which will be investigated by a stochastic approach later on.

From the local interpolation estimates (4) and (5), one can derive the global error estimates, known as Bramble–Hilbert's Lemma, by "gathering" the local estimates. This is a standard procedure (see for instance [12]), that gives, assuming the exact solution u regular enough, $V = H^1(\Omega)$, and Ω being a convex polygonal domain, the following estimate:

Lemma 2.2. *Let $u_h^{(1)}$ be the P_1 finite element approximation of u. Then, the error $\|u - u_h^{(1)}\|_{H^1(\Omega)}$ is bounded by*

$$\|u - u_h^{(1)}\|_{H^1(\Omega)} \leq \gamma_1 h, \tag{7}$$

where $\gamma_1 \equiv \beta_1 \|D^2 u\|_{L^\infty(\Omega)}$ is an unknown constant which does not depend on the mesh size h.

In a very similar way, analogous results can be derived for $u_h^{(2)}$, the P_2 finite element approximation of u. In this case, assuming the exact solution u regular enough, one obtains that the error $\|u - u_h^{(2)}\|_{H^1(\Omega)}$ is bounded by

$$\|u - u_h^{(2)}\|_{H^1(\Omega)} \leq \gamma_2 h^2, \tag{8}$$

where γ_2 is also an unknown constant, analogous to γ_1 for the P_2 finite element approximation.

Therefore, the presence of the two unknown constants γ_1 and γ_2 could lead to the situation where $\gamma_2 h^2 \gg \gamma_1 h$, for a given mesh size h. As a consequence, we suspect the following numerical situation to occur

$$|u(\mathbf{x}) - u_h^{(1)}(\mathbf{x})| \leq |u(\mathbf{x}) - u_h^{(2)}(\mathbf{x})|, \tag{9}$$

or at least

$$|u(\mathbf{x}) - u_h^{(1)}(\mathbf{x})| \simeq |u(\mathbf{x}) - u_h^{(2)}(\mathbf{x})|, \tag{10}$$

which means that *locally* for some $\mathbf{x}$, P_1 finite elements might be either more accurate than P_2 finite elements (inequality (9)), or equivalent (inequality (10)).

In the following section, our purpose will be to explore the possibilities of such situations by implementing data mining and probabilistic models on numerical approximations, collected inside a convenient database. More precisely, we will characterize the regions where two numerical approximations $u_h^{(1)}$ and $u_h^{(2)}$ could satisfy (10). This will allow us to see where P_2 finite elements are overqualified to guarantee accuracy.

3. The model problem

To illustrate our approach, the first step consists in introducing a model problem (1) and its finite element P_1 and P_2 approximations (2). To this end, we consider a quasi-static elliptic approximation of the Vlasov–Maxwell equations in a relativistic case. This system is discretized by a Particle-In-Cell method: the Vlasov equation is approximated by a particle method [6], whereas the electromagnetic field is discretized by P_1 and P_2 finite elements.

This model, derived from plasma simulations, has the advantage to be "rich enough", leading to a database constituted by a large number of different variables. Indeed, in an "elementary" model problem, as for instance the Laplace problem, the P_1 and P_2 approximations may be too predictable, so that our investigation method does not add much.

Remark 3.1. Even if the Vlasov–Maxwell system is time-dependent, the model problem we used here is quasi-static. The difference between the quasi-static and the static model is that, in the first, the right-hand sides are (explicitly) time-dependent. This difference is not a terminology subtlety. Indeed, from a numerical point of view, solving a quasi-static problem with a time-dependent right-hand side, amounts to solve a series of static problems after the time-discretization is performed. As a consequence, the database we will construct will contain the collection of solutions obtained at each time step, by the two finite element methods. This will positively enriched the database. Moreover, the problem being elliptic at each time step, this study fits within Section 2 framework.

Actually, the model we consider is an asymptotic paraxial Vlasov–Maxwell model (cf. [24,5]), derived by exploiting the physical and geometrical configuration of our problem. This paraxial model has a controlled accuracy and is obtained in the same spirit as in [15,26,31,32].

Let us briefly describe the problem configuration. Consider a beam of charged particles, for instance electrons, with a mass m and a charge q which moves inside a perfectly conducting cylindrical tube of boundary Γ. Let us denote by z the axis of the tube, which is also the optical axis of the beam. Due to the axisymmetric features, we will use the cylindrical coordinates (r, θ, z). The boundary Γ is thus written $\Gamma = \{(r, \theta, z);\ r = R\}$, R being the radius of the tube. For the sake of simplicity, we assume here that there are no external fields.

Classically, the motion of these particles is described in terms of a particle distribution function $f(\mathbf{X}, \mathbf{P}, t)$, which satisfies the relativistic axisymmetric Vlasov equation. Solving this equation needs to compute the electromagnetic field, which appears in this equation through the Lorentz force. In the paraxial model considered here, this electromagnetic field does not satisfy the classical axisymmetric Maxwell equations, but an approximate paraxial quasi-static problem.

Basically, it is derived by assuming that the beam – the dimension of which being small compared to the longitudinal length of the device – is highly relativistic, i.e. the relativistic factor γ satisfies $\gamma \gg 1$. We also

assume that the longitudinal particle velocities v_z are close to the light velocity c, whereas the modulus of transverse particle velocities $(v_r^2+v_\theta^2)^{1/2}$ is small compared to c. This model also requires to introduce the new position-velocity coordinates $\zeta = ct - z,\ v_\zeta = c - v_z$.

In these conditions, one can obtain a model in which appear only the transverse electric field E_r, the longitudinal electromagnetic field (E_z, B_z), and the transverse so-called "pseudo-fields" $\mathcal{E} = (\mathcal{E}_r, \mathcal{E}_\theta)$, defined by $\mathcal{E}_r = E_r - cB_\theta, \mathcal{E}_\theta = E_\theta + cB_r$. All these quantities depend implicitly on the time t and on the axisymmetric coordinates (r, ζ). The equations of this model are written

$$\begin{cases} \dfrac{1}{r}\dfrac{\partial}{\partial r}(r\,E_r) = s\dfrac{1}{\varepsilon_0}\rho(r,\zeta,t), \\ E_r(0,\zeta,t) = 0, \end{cases} \qquad \begin{cases} \dfrac{\partial E_z}{\partial r} = \dfrac{1}{c}\dfrac{\partial E_r}{\partial t}, \\ E_z(R,\zeta,t) = 0, \end{cases} \qquad \begin{cases} \dfrac{\partial B_z}{\partial r} = \mu_0 J_\theta(r,\zeta,t), \\ \displaystyle\int_0^R B_z r dr = 0, \end{cases}$$

$$\begin{cases} \dfrac{1}{r}\dfrac{\partial}{\partial r}(r\,\mathcal{E}_r) = \mu_0\, c\, J_\zeta(r,\zeta,t) - \dfrac{1}{c}\dfrac{\partial E_z}{\partial t}, \\ \mathcal{E}_r(0,\zeta,t) = 0, \end{cases} \qquad \begin{cases} \dfrac{1}{r}\dfrac{\partial}{\partial r}(r\,\mathcal{E}_\theta) = -\dfrac{\partial B_z}{\partial t}, \\ \mathcal{E}_\theta(0,\zeta,t) = \mathcal{E}_\theta(R,\zeta,t) = 0. \end{cases} \tag{11}$$

With this model, the expression of the Lorentz force $\mathbf{F} = (F_r, F_\theta, F_z)$ is given by

$$\begin{cases} F_r = q(\mathcal{E}_r + v_\theta B_z + v_\zeta E_r/c), \\ F_\theta = q(\mathcal{E}_\theta - v_r B_z), \\ F_z = q(E_z + v_r E_r/c). \end{cases}$$

In the right-hand sides of (11) also appear the charge density ρ and the current density $\mathbf{J} = (J_r, J_\theta, J_\zeta = \rho c - J_z)$, obtained respectively as the zero and the first moments of the distribution function f solution to the Vlasov equation. Details about the model and its derivation can be found in [1].

As a result, the model problem (1) we will study later on is the variational formulation deduced from (11), and its P_1 and P_2 finite element approximations (2), (see details in the Appendix).

Then, we have implemented a P_1 and P_2 finite element approximation, by using the FreeFem++ package [18], and a particle approximation of the paraxial Vlasov equation. The coupling between these two approaches, performed using assignment and interpolation procedures adapted to the P_1 and P_2 finite element respectively, required us to built specific algorithms, (see also details in the Appendix).

The last step to construct the database of numerical approximations, consists in collecting all the numerical results computed by these two families of finite elements at each time step (see Remark 3.1). Note also that, since the two methods have not the same degrees of freedom, we chose to keep in the database the common data, that is the quantities computed at the vertices of the mesh.

4. Data mining methods and probabilistic model for discretization error analysis

In this section, exploiting the database described above, we show how to use data mining methods in combination with a probabilistic model to characterize the regions of space and time where the two finite element approximation schemes P_1 and P_2 are numerically equivalent. As stated above, the database we consider is composed by data computed at each time step and for each node of the mesh. More precisely, a given row of the database is informed by the approximations of all of the physical unknowns computed at a given time step and mesh node.

Then, the set of variables which corresponds to the columns of the database we considered is listed below:

$$v_r^{(i)}, v_\theta^{(i)}, v_\zeta^{(i)}, p_r^{(i)}, p_\theta^{(i)}, p_\zeta^{(i)}, E_r^{(i)}, E_z^{(i)}, B_z^{(i)}, \mathcal{E}_r^{(i)}, \mathcal{E}_\theta^{(i)}, \rho^{(i)}, J_r^{(i)}, J_\theta^{(i)}, J_\zeta^{(i)}, F_r^{(i)}, F_\theta^{(i)}, F_z^{(i)} \quad (i = 1, 2), \tag{12}$$

where the exponent i specifies that the approximations are computed by the P_i Lagrange finite element, the quantities $(p_r^{(i)}, p_\theta^{(i)}, p_\zeta^{(i)}) = \gamma\, m\, (v_r^{(i)}, v_\theta^{(i)}, v_\zeta^{(i)})$ denoting the impulsions. Considering all the 100 time steps t_n and the 1250 space nodes (r_j, ζ_k) of our simulations, the database we treated was composed by 125 000 rows and by the 36 variables listed in (12).

4.1. Data mining principles and objectives of exploration

Data Mining is an activity of information extraction, whose goal is to discover hidden or *a priori* unknown facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques

and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results.

Recall that our objective is to appreciate the equivalence of accuracy between the P_1 and P_2 approximate problem solutions. To this end, we have to determine if there locally exist rows in the database, characterized by their (r_j, ζ_k, t_n), such that the P_1 method would bring a better approximation than the P_2 finite element or, at least, an equivalent one. In the absence of the exact solution (the u of Section 2), we cannot determine what "better approximation" means. For this reason, we concentrate on identifying the subgroups in the database such that the numerical approximations computed by the P_1 and P_2 finite elements are with "*the same order*", according to relation (10). Note that even this notion of "*the same order*" will require to be defined.

Our starting point will be the longitudinal component of the magnetic field B_z. Other variables of (12) can be equivalently considered, and the possible interference between variables will be partly discussed in Section 5. To identify the approximations of "*the same order*", we will only keep in the database the rows where the electromagnetic field is theoretically non zero. Looking for Eq. (11) – where only derivations with respect to r (and not ζ) are involved – we readily see that it corresponds to the rows of the database parameterized by (ζ_k, t_n) such that:

$$\int_0^R \rho(r, \zeta_k, t_n) dr \neq 0.$$

By applying this rule, we extract from the database the corresponding rows and we got a dataset, (called in the rest of the paper the database), made of 27 864 rows to be explored, for identifying, if there exist, rows such that:

$$|B_z^{(1)}(r_j, \zeta_k, t_n) - B_z(r_j, \zeta_k, t_n)| \simeq |B_z^{(2)}(r_j, \zeta_k, t_n) - B_z(r_j, \zeta_k, t_n)|. \tag{13}$$

Relation (13) describes the situation where local behavior (namely, for a given (r_j, ζ_k, t_n)) of approximations $B_z^{(1)}$ and $B_z^{(2)}$ are equivalent. It gives an alternative point of view that differs from the Bramble–Hilbert theorem which rather evaluates a global error, that can be written in our case:

$$\left\| B_z^{(1)} - B_z \right\|_{H^1(\Omega)} \leq \gamma_1 h, \quad \text{and} \quad \left\| B_z^{(2)} - B_z \right\|_{H^1(\Omega)} \leq \gamma_2 h^2, \tag{14}$$

where γ_i, $(i = 1, 2)$, are two given but *u*nknown constants that can modify the actual quality of approximations. In particular, such situation would show that global estimate (14) does not prevent any local estimate such as (13) to occur.

Indeed, the situation described by (13) means that, in a certain sense (will be soon specified), $B_z^{(1)}$ and $B_z^{(2)}$ have the "*same numerical order*". Again, since B_z is not known, we will identify (13) by exploring the subgroups in the database such that $B_z^{(1)}$ and $B_z^{(2)}$ are of "*the same order*".

To identify such situations, let us first define the notion of "*same numerical order*". For this purpose, we introduce a threshold α and a new qualitative binomial variable called P_1 *vs.* P_2 as follows:

Let N denotes the total number of rows in the dataset, then:

$$\forall l = 1, \quad N : (P_1 \ vs. \ P_2)_l \equiv \begin{vmatrix} \textit{Same Order}, & \text{if } |B_{z,l}^{(2)} - B_{z,l}^{(1)}| \leq \alpha \max_{j=1,N} |B_{z,j}^{(2)} - B_{z,j}^{(1)}|, \\ \textit{Different Order}, & \text{if not}, \end{vmatrix} \tag{15}$$

where the threshold α denotes a percent of the maximum of the absolute difference between $B_z^{(1)}$ and $B_z^{(2)}$ found in all the dataset, $B_{z,j}^{(1)}$ and $B_{z,j}^{(2)}$ the evaluations of B_z by the finite elements P_1 and P_2 corresponding to the jth row on the dataset.

In such a way, the variable P_1 *vs.* P_2, especially its value "*Same Order*", will allow us to detect and to characterize situations where relation (13) holds, if any.

4.2. Determination of α by a probabilistic model

At first glance, the threshold α, which defines the two classes of the binomial variable P_1 *vs.* P_2 in (15), could be chosen in a heuristic way, using some *a* priori informations about the model and the data. Our purpose here is to propose a more rigorous way to estimate it, by using a probabilistic model. This model is based on a sampling-like property observed on the numerical results we got, which constitute the variables of our database.

Table 1
Kolmogorov–Smirnov test results.

Number of picking time steps	Sampling size	Null hypothesis	p-value	Decision
10	3000	$B_z^{(1)}$ distributions identical	0.022	Rejected
		$B_z^{(2)}$ distributions identical	0.000	Rejected
7	4248	$B_z^{(1)}$ distributions identical	0.000	Rejected
		$B_z^{(2)}$ distributions identical	0.000	Rejected
6	4680	$B_z^{(1)}$ distributions identical	0.368	***Accepted***
		$B_z^{(2)}$ distributions identical	0.000	Rejected
5	5808	$B_z^{(1)}$ distributions identical	0.431	***Accepted***
		$B_z^{(2)}$ distributions identical	0.064	***Accepted***

4.2.1. Equivalent systematic sampling

The well-known stability condition for the classical Particle-In-Cell method [6,19], as for the one presented above Section 3 (see [4]), implies that any given particle must not travel across more than one element of the mesh, here a triangle, during a given time step. Hence, if we look for the number of time steps needed for the particles to travel across one triangle during the simulation time, we find that even the faster particles need at least ten time steps to travel across the smallest triangles.

In other words, this property means that the variations of the electromagnetic field will be certainly "significative" after ten time steps. As a consequence, the B_z distribution should be searched statistically equivalent between the whole population of approximations and a sampling obtained by picking a minimum number of time steps that must be at most equal to ten time steps, on all the hundred time steps which defined the database.

More concretely, having not any nice parametric features of the distribution $B_z^{(1)}$ and $B_z^{(2)}$, such normality, we processed the non parametric Kolmogorov–Smirnov test [14], (with a standard p-value equal to 0.05), to identify which systematic sampling will be statistically equivalent to the whole population regarding the distributions of $B_z^{(1)}$ and $B_z^{(2)}$ computed by the P_1 and P_2 finite elements.

Having implementing this test under *IBM SPSS Statistics*, we found that the optimal systematic sampling, that is the biggest one, is built by picking the data each five time steps, which corresponds to 5808 rows of the 27 864 rows of the database, (see Table 1).

Indeed, one can observe in Table 1 that from ten time steps to six time steps, the *null hypothesis* which supposes that the B_z distributions are identical between the sampling and the rest of the database is rejected. But, when one consider one time steps on five, the *null hypothesis* is validated according to a critical p-value equal to 0.05.

This is in agreement with the meaning of the p-value: If the p-value is very small, usually less than or equal to a threshold value previously chosen called the significance level (traditionally 5%), it suggests that the observed data are inconsistent with the assumption that the *null hypothesis* is true, and thus that hypothesis must be rejected and the other hypothesis accepted as true [29].

In statistical terms, it means that if we consider in the database a new dataset constituted by a randomness systematic sampling (for more details, see for example [7]) where only one in five time steps is retained, the longitudinal magnetic field B_z distribution must be statistically equivalent between the whole database and the corresponding dataset sampling.

Here, in our case, we retain the "Every 5 time steps" systematic sampling which is motivated by the above statistical property we detailed.

4.2.2. The probabilistic model

Let us remind that our aim is to determine the parameter α, that will define the *S*ame Order of the binomial variable P_1 *vs.* P_2. For this purpose, we introduce the randomness variable Y which denotes the absolute difference between the numerical estimates processed by the P_1 and P_2 finite elements associated to the longitudinal magnetic component B_z, i.e. $Y = |B_z^{(2)} - B_z^{(1)}|$.

For a given α, let p_α denotes the uniform probability of picking any individual l, $(l = 1, N)$, in the database to be in the "*Same Order*" category. Then

$$\forall l = 1, N : \quad p_\alpha \equiv \text{Prob}\{Y_l < \alpha \max_{j=1,N} Y_j\}, \tag{16}$$

where α is the threshold we introduced in (15), Y_j the trace of Y on any element number j in the database which is composed by N elements, (N is equal to the 27 864 points (r_j, ζ_k, t_m) which are present in the database). For a given α, p_α is measured on the whole population of the database as the percent of rows which are qualified "*Same Order*".

Let us now consider a systematic sampling of n elements, ($n < N$, where $n = 5808$ according to the statistical results of Section 4.2.1), and the randomness Bernoulli variable X_i, $(i = 1, n)$, defined by:

$$X_i = \begin{vmatrix} 1 & \text{if } Y_i < \alpha \max\limits_{j=1,N} Y_j, \\ 0 & \text{if not,} \end{vmatrix} \tag{17}$$

where Y_i denotes the trace of Y on any element i, $(i = 1, n)$, which belongs to the sampling.

Moreover, we have:

$$\text{Prob}\{X_i = 1\} = \text{Prob}\{Y_i < \alpha \max_{j=1,N} Y_j\} = p_\alpha. \tag{18}$$

We also introduce the randomness variable X which allows us to count the number of all the individuals in the sampling which are qualified "*Same Order*":

$$X = \sum_{i=1}^{n} X_i. \tag{19}$$

Then, X follows a binomial law of parameters n and p_α, usually denoted $X \hookrightarrow \mathcal{B}(n, p_\alpha)$, whose expected value μ_X and standard deviation σ_X are given by:

$$\mu_X = np_\alpha \quad \text{and} \quad \sigma_X^2 = np_\alpha(1 - p_\alpha). \tag{20}$$

Finally, we introduce the frequency of "*Same Order*" in the sampling which corresponds to the randomness variable X/n.

Our purpose, as we got in the Section 4.2.1 by statistical arguments, is to guarantee that X/n measured on the sampling does not diverge "too much" from p_α evaluated on the total population.

To specify this property, let ϵ be a small and positive parameter which belongs to the interval [0, 1]. Then, our aim is to ensure that:

$$\text{Prob}\left\{p_\alpha - \epsilon p_\alpha \le \frac{X}{n} \le p_\alpha + \epsilon p_\alpha\right\} \ge S, \tag{21}$$

where S denotes the confidence threshold usually chosen as 95%.

One can transform inequality (21) to bring up a centered and reduced randomness variable as follows:

$$\text{Prob}\left\{\left|\frac{X - \mu_X}{\sigma_X}\right| \le \epsilon \sqrt{\frac{np_\alpha}{(1 - p_\alpha)}}\right\} \ge S. \tag{22}$$

Therefore, we are in position to approximate the probability of the binomial law in Eq. (22) by the normal law:

$$\text{Prob}\left\{\left|\frac{X - \mu_X}{\sigma_X}\right| \le \epsilon \sqrt{\frac{np_\alpha}{(1 - p_\alpha)}}\right\} \approx 2 \int_0^{\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt. \tag{23}$$

In these conditions, the determination of p_α is obtained by:

$$2 \int_0^{\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt \ge 0.95, \tag{24}$$

Table 2
Marginal distribution of the qualitative variable P_1 *vs.* P_2 when $\alpha^* = 0.62\%$.

	Count	Percent
Different order	22 040	79.1%
Same order	5 824	20.9%

which can be solved using a table of standard normal distributions [21]. We get:

$$\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}} \geq 1.96 \iff p_\alpha \geq \frac{3.84}{3.84 + n\epsilon^2}. \tag{25}$$

In our case, if we consider the following numerical values of ϵ and n: $\epsilon = 5\%$, $n = 5808$, (where n is the number we found by statistical arguments by the help of the Kolmogorov–Smirnov test—see Section 4.2.1), Eq. (25) leads to the minimum value p_α^* of p_α, which corresponds to:

$$p_\alpha^* = 0.209.$$

With this result, which means that 20.9% of the total database must be qualified by the "*Same Order*" category according to the meaning of p_α defined by (16), we are in position to determine the corresponding value of α^* estimated on the whole database.

Practically, we have obtained α^* by processing successive marginal distributions under *IBM SPSS Modeler*. We found $\alpha^* = 0.62\%$, which guarantees that 20.9% of the total database is composed by "*Same Order*" elements regarding the longitudinal component B_z, (see Table 2).

4.2.3. Error estimate of the binomial law by the normal law

Since we have approximated the binomial law by the normal law in (23), we would like to specify the result (25), by taking into account the error bound for this approximation.

The corresponding error estimate is mainly due to Uspensky [33] from which one can prove the following lemma:

Lemma 4.1. *If* $npq \geq 25, \forall (x_1, x_2) \in R^2, (x_1 < x_2)$, *we have:*

$$\left|\text{Prob}\left\{x_1 \leq \frac{X-\mu_X}{\sigma_X} \leq x_2\right\} - [\phi(x_2) - \phi(x_1)]\right| \leq \frac{C}{\sqrt{npq}}, \tag{26}$$

where:

$$\phi(x) \equiv \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt \quad \textit{and} \quad C < 0.588,$$
$$X \hookrightarrow \mathcal{B}(n, p), \quad (q = 1 - p). \tag{27}$$

In our case, we have: $x_2 = -x_1 = \epsilon\sqrt{\frac{np_\alpha^*}{1-p_\alpha^*}}$. Then, we get:

$$\left|\text{Prob}\left\{\left|\frac{X-\mu_X}{\sigma_X}\right| \leq \epsilon\sqrt{\frac{np_\alpha^*}{1-p_\alpha^*}}\right\} - 2\int_0^{\epsilon\sqrt{\frac{np_\alpha^*}{1-p_\alpha^*}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt\right| \leq \frac{C}{\sqrt{np_\alpha^*(1-p_\alpha^*)}}. \tag{28}$$

We can numerically evaluate the right side of inequality (28) and we finally get:

$$\left|\text{Prob}\left\{\left|\frac{X-\mu_X}{\sigma_X}\right| \leq \epsilon\sqrt{\frac{np_\alpha^*}{1-p_\alpha^*}}\right\} - 2\int_0^{\epsilon\sqrt{\frac{np_\alpha^*}{1-p_\alpha^*}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt\right| \leq 0.019. \tag{29}$$

Table 3
Marginal distribution of the qualitative variable P_1 *vs.* P_2 when $\alpha^{**} = 0.75\%$.

	Count	Percent
Different order	21 037	75.5%
Same order	6 827	24.5%

It means that the approximation we proceeded in (23) implies for inequality (22) that:

$$\text{Prob}\left\{\left|\frac{X-\mu_X}{\sigma_X}\right| \leq \epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}}\right\} \geq 0.95 - 0.019 \approx 0.93, \tag{30}$$

which is still a convenient confidence threshold.

But, if one wants to keep a confidence level of 0.95, inequality (24) must be changed with a right hand equal to 0.97.

After some calculations, the corresponding inequality (25) becomes:

$$\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}} \geq 2.17 \iff p_\alpha \geq \frac{4.71}{4.71 + n\epsilon^2}. \tag{31}$$

Therefore, the new derived value p_α^{**} and its associated value α^{**} are then:

$$p_\alpha^{**} = 0.245 \quad \text{and} \quad \alpha^{**} = 0.75\%. \tag{32}$$

So, 24.5% of the total database is composed by "*Same Order*" elements regarding the longitudinal component B_z, (see Table 3), which guarantee a confidence level greater or equal to 95% between the value of p_α^{**} measured on the whole population and its estimation by the frequency X/n on the systematic sampling.

Remark 4.1. The two values p_α^* and p_α^{**} we found, satisfy the condition $npq \geq 25$ of Lemma 4.1, with $n = 5808$.

4.3. Logistic regression and qualification of the "Same Order" category

At this point, we are able to perform now the analysis of the "*Same Order*" category defined in (15), according to the choice of α^{**} (Eq. (32)) identified in the previous section. This choice guarantees that any systematic sample made by selecting one of five time steps will statistically be equivalent, for a confident level of 95%, to the whole database, regarding the magnetic longitudinal component B_z.

Let us begin by recalling some basic knowledge about logistic regression applied to our situation. The interested reader is referred for example to [13,17,22] or [20].

We consider the randomness binary variable Z which simply allows to code the variable P_1 *vs.* P_2 defined by (15) as follows:

$$\forall\, l = 1, \quad N : Z_l \equiv \begin{cases} 1 & \text{if } |B_{z,l}^{(2)} - B_{z,l}^{(1)}| \leq \alpha^{**} \max\limits_{j=1,N} |B_{z,j}^{(2)} - B_{z,j}^{(1)}|, \\ 0 & \text{if not,} \end{cases} \tag{33}$$

where N denotes the size of the entire population of the database, (i.e., the total numbers of rows: $N = 27\,864$).

Obviously, Z is a function of many predictors which are the discrete time t_n, the space coordinates (r_j, ζ_k) and all of the numerical approximations computed by the Particle-In-Cell finite elements method we implemented. We remark that all of this predictors are real variables.

If we denote by X one of these predictors, our interest is to model the conditional probability $p(x)$ defined by:

$$p(x) \equiv \text{Prob}\{Z = 1 | X = x\}.$$

It is not a surprise that statisticians have tackled this problem by trying to use linear regression. The most obvious idea is to let $p(x)$ be a linear function of x. With this assumption, every increment of a component of x will add or subtract

a proportional quantity to the probability. The conceptual problem here is that $p(x)$ belongs to the interval [0, 1], and linear functions are unbounded. So, linear models cannot be used on $p(x)$.

Therefore, the idea is to first introduce the odds of getting ($Z = 1$ if $X = x$) *vs.* ($Z = 0$ if $X = x$) defined by:

$$\text{odds} = \frac{p(x)}{1 - p(x)}. \tag{34}$$

Odds values are in $[0, +\infty[$, and then, $\ln\left(\frac{p(x)}{1-p(x)}\right)$, (also called the *logit* function of $p(x)$), has values in $]-\infty, +\infty[$.

As a consequence, it is possible to consider the following linear regression:

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x, \tag{35}$$

or equivalently by extracting $p(x)$:

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}. \tag{36}$$

Eq. (35) can easily be generalized for a set $(X_1, \ldots, X_n)$ of n predictors as follows:

$$\ln\left(\frac{p(x_1, \ldots, x_n)}{1 - p(x_1, \ldots, x_n)}\right) = \beta_0 + \beta_1 x_1, + \cdots + \beta_n x_n, \tag{37}$$

where $p(x_1, \ldots, x_n)$ denotes the conditional probability defined by:

$$p(x_1, \ldots, x_n) \equiv \text{Prob}\{Z = 1 | X_1 = x_1, \ldots, X_n = x_n\}.$$

The parameters $(\beta_1, \ldots, \beta_n)$ are usually estimated by implementing the maximum likelihood method [20].

To make as clear as possible, we continue our purpose in the case of the one dimensional case modeled by Eq. (35) without loss of generality.

A very nice property one can deduce from (35) is the interpretation of the coefficients, particularly of β_1. Indeed, let us introduce the odds ratio $o(x_0, x_1)$ defined by:

$$o(x_0, x_1) \equiv \frac{p(x_1)/(1 - p(x_1))}{p(x_0)/(1 - p(x_0))}, \tag{38}$$

where (x_0, x_1) correspond to two different values of the randomness variable X.

An odds ratio of 1 indicates that the condition or event under study is equally likely to occur in both cases when $x = x_0$ or $x = x_1$. An odds ratio greater than 1 indicates that the condition or event is more likely to occur when $x = x_1$. And an odds ratio less than 1 indicates that the condition or event is less likely to occur when $x = x_1$.

As a consequence, one can easily show for an increment of a unit of X, ($x_0 = x, x_1 = x + 1$), Eqs. (35) and (38) lead to the meaning of β_1:

$$\beta_1 = \ln(o(x, x + 1)). \tag{39}$$

In the same way, for an increment of c unities of X, one can show that:

$$o(x, x + c) = \exp(c\beta_1). \tag{40}$$

Interpretation (39) and (40) of β_1 can be obtained for each coefficient β_i of the multidimensional case (37), under condition that one keeps fixed all the other variables x_j, ($j \neq i$).

We are now in position to model the relationship between the "*Same Order*" category ($Z = 1$) and the other predictors X_i, by the help of the logistic regression.

5. Numerical results

Let us come back now to the study of difference of accuracy between $B_z^{(1)}$ and $B_z^{(2)}$. We are interested in qualifying the subgroup composed by the "*Same Order*" of the coding variable Z defined by (33).

The first elementary feature of Z confirmed our suspicion: the database contains a non negligible quantity of rows such that (13) is satisfied. Indeed, recalling that when $\alpha^{**} = 0.75\%$, $p_\alpha^{**} = 24.5\%$. It means that more than 6800 rows in the database corresponds to this case as it is shown in Table 3, see above.

So, we processed the logistic regression model under the data mining platform *IBM SPSS Modeler* by retaining as predictors variables X_i, all of the available variables listed in (12), except the variables $B_z^{(1)}$ and $B_z^{(2)}$ which directly participate in the definition of the variable Z.

The corresponding model then presents the following properties.

1. The contingency Table 4 shows that almost 75% of the category "*Same Order*" are correctly classified by the logistic regression. This result is quite satisfying if one keeps in mind the two next features of our analysis.
   - Our objectives in this study are to propose tools to identify and eventually to confirm the existence of situations where P_1 and P_2 finite elements lead to equivalent results, as defined in (10). It is not yet time to decide under which circumstances one will be able to leave the P_2 finite element for the P_1 ones, to get accurate results for a given concrete application.
   - Anyway, one has to keep in mind that daily meteorology forecasts are systematically published with an index of confidence which is between 60% and 80% which does not disturb anyone....
2. We processed a second logistic model by only retaining the time t and the space coordinates (r, ζ) as the only predictor variables. Table 5 shows that these three variables are sufficient to produce an equivalent model compared with the previous "full" one, even when quite better regarding the "*Same Order*" group.
   Moreover, the order of importance of these three predictors in the logistic regression model is plotted in Fig. 1. Then, we can notice the dominant role played by r in the model. This is in agreement with the features of solutions to paraxial Vlasov–Maxwell Eqs. (11), especially the equation of B_z, where both the differential equation and the boundary condition are basically function of the variable r.
3. We proceed now to the main part of the results one have to consider with a logistic regression. It treats on the interpretation of the parameters which are the coefficients of the equation of the category "*Same Order*" which correspond to the value $Z = 1$ of the randomness variable defined by (33).
   Then, in the present case we implemented a logistic regression on *IBM SPSS Modeler* and we got:

$$\ln\left(\frac{\text{Prob}\{Z=1|(t,r,\zeta)\}}{1-\text{Prob}\{Z=1|(t,r,\zeta)\}}\right) = 0.01176\,t + 0.01545\,r - 0.1696\,\zeta. \tag{41}$$

   As we see above in (39) and (40), the meaning of the coefficients β_i of Eq. (41) is easier to interpret when one considers the quantities $\exp(\beta_i)$. So, we summarize all the related values in Table 6 with their corresponding p-values to proceed our analysis.
   Because we consider here a logistic regression with multiple predictor variables, the general rule to get the right interpretation of the coefficients can be formulated as follows: each estimated coefficient is the expected change in the log odds of being in the "*Same Order*" class, corresponding to a unit increase in the associated predictor variable, holding the other predictor variables constant at a certain value. Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale, corresponding to a unit increase in the associated predictor variable, holding other variables at a certain value.

   - *Interpretation of the coefficient* β_t *and* $\exp(\beta_t)$.
     According to the definition (34) of the odds and thanks to (38) and (39), we can say now that the coefficient β_t of the time t in (41) is the difference in the log odds. In other words, for a one-unit increase of time, (i.e. a time step), the expected change in log odds is $\beta_t = 0.01176$.
     Can we translate this change in log odds to the change in odds? Indeed, by (39) we can say for a one-unit increase in time, we expect to see about 1.2% increase in the odds of being in the "*Same Order*" class $(\exp(0.01176) \simeq 1.012)$.
     Even when this growth increasing seems quite small, one does not forget that we have to deal with one hundred time steps in our numerical simulations. So, by considering Eq. (40), one must deal with an about 50% increase in the odds of being in the "*Same Order*" class after 35 time steps, $(\exp(35 * 0.01176) \simeq 1.5)$ and with about 224% increase in the odds of being in the "*Same Order*" class after 100 time steps, $(\exp(100 * 0.01176) \simeq 3.24)$. In other words, the more time passes, the more the odds of being in the "*Same Order*" class increases and less useful and justified is the implementation of P_2 finite elements.

Table 4
Contingency table for a complete regression logistic model.

Observed category	Predicted category		
	Different order	Same order	Correct percent
Different order	12 597	8 438	59.9%
Same order	5 307	15 742	74.8%

Table 5
Classification table for a regression logistic model based on the predictors (t, r, ζ).

Observed category	Predicted category		
	Different order	Same order	Correct percent
Different order	11 248	9 787	53.5%
Same order	5 267	15 788	75.5%

Remark that this behavior was not *a* priori expected. Indeed, nothing in Eq. (11) of B_z, which is basically an integration in the radial variable r, would let us to envisage this result. In fact, the time dependency of B_z is implicit and nonlinear, since it only depends on the time variation of the right-hand side $J_\theta(r, \zeta, t)$. This variation is quite complex and is related to the coupling of the paraxial Maxwell and Vlasov equation, $J_\theta(r, \zeta, t)$ depending on the velocity v_θ, that depends on the Lorentz force F, that depends on the fields solution to the paraxial model. Remark also that, as often noted [25], data mining are powerful for investigate nonlinear dependency.

- *Interpretation of the coefficient* β_r *and* $\exp(\beta_r)$.

  In the same way, for a one-unit increase of r, the expected change in log odds is $\beta_r = 0.01545$. This change in log odds corresponds to an equivalent change in odds of about 1.55% increase in the odds of being in the "*Same Order*" class, $(\exp(0.01545) \simeq 1.01556)$. As r belongs to the interval [0, 120] in our simulations, we found that, when $r = 26$, one must deal with about 50% increase in the odds of being in the "*Same Order*" class, (since $\exp(26 * 0.01545) \simeq 1.5$). Finally, when $r = 120$, the odds of being in the "*Same Order*" ratio is about 6.38 times more, (again, $\exp(120 * 0.01545) \simeq 6.38$) compared with $r = 0$; the situation "*Same Order*" is most likely when $r = 120$ rather than $r = 0$.

  In other words, the closer you get to the tube wall, the more equivalent are P_1 and P_2 approximations expected to be. This can be explained by the presence of the vanishing integral boundary condition on B_z which obliged the solution to vanish at $r = R$.

- *Interpretation of the coefficient* β_ζ *and* $\exp(\beta_\zeta)$.

  As we have $\beta_\zeta = -0.1696$, the presence of the sign "−" allows us conclude that the more ζ grows, the more the odds of being in the "*Same Order*" class decreases. Indeed, this feature can be quantified by the help of $\exp(-0.1696) \simeq 0.8440$ which means that for a one-unit increase of ζ, we expect to see about 15.6% decrease in the odds of being in the "*Same Order*" class. As a consequence, after 4 units of ζ one must deal with about 50% decrease in the odds of being in the "*Same Order*" class, $(\exp(-4 * 0.1696) \simeq 0.5)$. Moreover, as the maximum value of ζ is 15 in the mesh we implemented, $\exp(-15 * 0.1696) \simeq 0.08$. This implies that at the end of the bunch, the situation "*Same Order*" is least likely, (the odds of being in the "*Same Order*" class decreases about 92%), rather than at the beginning of the bunch where $\zeta = 0$.

  Here again, as for the interpretation of the coefficient β_t, this behavior was not expected even after a close look of the equations, and it probably related to the nonlinear coupling of the solved system.

- Finally, we also notice that Eq. (41) does not include a constant β_0. Indeed, we choose not to include a constant in the equation model of the logistic regression because its interpretation would concern the case where $t = r = \zeta = 0$ which corresponds to a singular situation for which we have not found an interesting interpretation as we did for the other above coefficients.

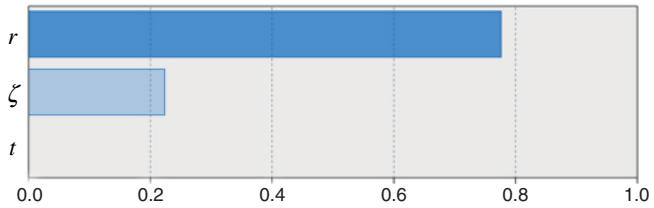

Fig. 1. Logistic regression—importance of the predictors.

Table 6
Logistic regression—parameters estimation.

Predictor	$\exp(\beta_i)$	p-value for $\exp(\beta_i)$
t	1.012	0.000
r	1.016	0.000
ζ	0.844	0.000

6. Conclusion

In this paper, we have proposed a new approach that combines data mining techniques and probabilistic methods to evaluate error estimate in approximations of partial differential equations. In a first part, we showed that classical finite element error estimates contain a part of uncertainty, due to the presence of unknowns constants. This can influence and even damage the precision of numerical results computed by these methods.

In a second part, we introduced a model problem: a quasi-static elliptic approximation of the Vlasov–Maxwell equations. We then constructed a database made of all the numerical results obtained by a P_1 and P_2 finite element Particle-In-Cell method.

Afterward, based on data mining techniques, on the first hand, and statistic and probabilistic approach, on the other hand, we compare the numerical approximations obtained by the P_1 and P_2 finite elements. Especially, we characterized the influence of predictors on the odds of being in the "*Same Order*" category between.

Future developments of this work could concern the interaction between more than two predictors, for instance by introducing in the logistic regression equation, nonlinear terms like tr, $t\zeta$ and $r\zeta$. This would allow us to specify and refine the results presented in this article.

Another important extension could consist in investigating the ability to well model the second class of the randomness variable Z, which describes the "*Different Order*" class of P_1 and P_2 finite elements approximations.

Except the relatively modest level of accuracy we got for this class in the model detailed here (either 53% of rows were well classified in the last model—see Table 5—to about 60% for the "full" model—see Table 4), one should be able to conclude about the "best" approximation between P_1 and P_2 finite elements.

Another very exciting development will be also to adapt this analysis by using experimental data considered as a reference or exact solution.

Appendix

A variational formulation

This appendix is devoted to the variational formulation and its approximation of the problem (11). Recall first that our quasi-static model is written in a frame which moves along the optical axis at the speed of light, the bunch of particles is evolving slowly in that frame. As a consequence, the computational domain Ω is defined as a simple rectangular domain in variables (r, ζ), $0 \leq r \leq R, 0 \leq \zeta \leq Z$.

In addition, as the regularity of the fields is not an issue for our study, we will assume that they are smooth enough, for instance that they belong to an H^1-style standard Sobolev space. For the sake of simplicity, we will denote by V the space of the fields and of the test functions, regardless of the boundary conditions they satisfy.

Let v denote a test function of V. Multiplying the first equation of (11) by v, and integrating over the domain $(0, R) \times (0, Z)$ yields for the electric component E_r:

Find $E_r \in V$ such that

$$\int_0^Z \int_0^R \frac{\partial}{\partial r}(r\, E_r)\, v\, dr d\zeta = \frac{1}{\varepsilon_0} \int_0^Z \int_0^R \rho\, v\, r dr d\zeta, \quad \forall v \in V, \tag{42}$$

together with the boundary condition $E_r(0, \zeta, t) = 0$. Similarly, we get readily for the E_z component:

Find $E_z \in V$ such that

$$\int_0^Z \int_0^R \frac{\partial E_z}{\partial r}\, v\, dr d\zeta = \frac{1}{c} \int_0^Z \int_0^R \frac{\partial E_r}{\partial t}\, v\, dr d\zeta, \quad \forall v \in V, \tag{43}$$

together with the perfect conductor boundary condition $E_z(R, \zeta, t) = 0$.

Remark A.1. One can choose to handle the boundary condition by using an integration by parts formula regarding r. We get:

$$-\int_0^Z \int_0^R E_z \frac{\partial v}{\partial r}\, dr d\zeta - \int_0^Z E_z v(r = 0, \zeta, t) d\zeta = \frac{1}{c} \int_0^Z \int_0^R \frac{\partial E_r}{\partial t}\, v\, dr d\zeta, \quad \forall v \in V.$$

The difficulty to compute the magnetic component B_z is related to its boundary condition which has an integral form. To overcome this difficulty, we introduce the variable $\mathcal{B}_z(r, \zeta, t)$ defined by

$$\mathcal{B}_z(r, \zeta, t) := \int_0^r B_z(s, \zeta, t)\, s\, ds, \tag{44}$$

so that

$$\frac{\partial \mathcal{B}_z}{\partial r} = r\, B_z(r) \quad \text{and} \quad \frac{\partial^2 \mathcal{B}_z}{\partial r^2} = r \frac{\partial B_z}{\partial r} + B_z.$$

Hence, the third equation of (11) is written as

$$\frac{\partial^2 \mathcal{B}_z}{\partial r^2} - B_z = \mu_0 r J_\theta,$$

whereas the integral boundary condition gives:

$$\int_0^R B_z(s, \zeta, t)\, r\, dr = \mathcal{B}_z(R, \zeta, t) = 0.$$

Hence, one equivalently replaces the third equation of (11) by:

$$\begin{cases} r \dfrac{\partial^2 \mathcal{B}_z}{\partial r^2} - \dfrac{\partial \mathcal{B}_z}{\partial r} = \mu_0 r^2 J_\theta, \\ \mathcal{B}_z(R, \zeta, t) = 0, \\ \dfrac{\partial \mathcal{B}_z}{\partial r}(0, \zeta, t) = 0. \end{cases} \tag{45}$$

Then the unknown $B_z(r, \zeta, t)$ will be found from $\mathcal{B}_z(r, \zeta, t)$ by derivation. Note that there is no difficulty to compute $B_z(0, \zeta, t)$. Indeed, one easily deduces from the equations above in $\mathcal{B}_z$ that $B_z(0, \zeta, t) = 0$.

One can simply derive a variational formulation for $\mathcal{B}_z(r, \zeta, t)$. Multiplying (45) by a test function v, integrating over the domain $(0, R) \times (0, Z)$, and using an integration by parts in the second order term in r yields:

Find $\mathcal{B}_z \in V$ such that:

$$\begin{aligned} &-\int_0^Z \int_0^R \frac{\partial \mathcal{B}_z}{\partial r} \frac{\partial v}{\partial r}\, r\, dr d\zeta + R \int_0^Z \frac{\partial \mathcal{B}_z}{\partial r}\, v(R, \zeta, t) d\zeta - \int_0^Z \int_0^R \frac{\partial \mathcal{B}_z}{\partial r}\, v\, dr d\zeta \\ &= \mu_0 \int_0^Z \int_0^R r^2 J_\theta\, v\, dr d\zeta, \quad \forall v \in V. \end{aligned} \tag{46}$$

The advantage of this formulation being that only $\mathcal{U} := \frac{\partial \mathcal{B}_z}{\partial r}$ is involved and can be chosen as a new unknown. Moreover, the boundary condition $\frac{\partial \mathcal{B}_z}{\partial r}(0, \zeta, t) = 0$ is handled (in a weak way) through the integration by parts.

Remark A.2. Solving a variational formulation in $\mathcal{U}$ leaves an indetermination in the computation of $\mathcal{B}_z$. Indeed, $\mathcal{B}_z$ is *a priori* determined up to an additive constant. Using the boundary condition $\mathcal{B}_z(R, \zeta, t) = 0$ allows us to uniquely determine $\mathcal{B}_z$, for instance by choosing $\mathcal{B}_z(r, \zeta, t) = \int_R^r \mathcal{U}(r, \zeta, t)$.

The variational formulations for pseudo-fields $\mathcal{E}_r$ and $\mathcal{E}_\theta$ are straightforward and similar to the one of E_r, only the right-hand sides being different. For the sake of completeness, we write done in this appendix. We have

*F*ind $\mathcal{E}_r \in V$ such that

$$\int_0^Z \int_0^R \frac{\partial}{\partial r}(r\,\mathcal{E}_r)\, v\, dr d\zeta = \mu_0 c \int_0^Z \int_0^R J_\zeta\, v\, r dr d\zeta - \frac{1}{c}\int_0^Z \int_0^R \frac{\partial E_z}{\partial t}\, v\, r dr d\zeta, \quad \forall v \in V, \tag{47}$$

together with the boundary condition $\mathcal{E}_r(0, \zeta, t) = 0$.

*F*ind $\mathcal{E}_\theta \in V$ such that

$$\int_0^Z \int_0^R \frac{\partial}{\partial r}(r\,\mathcal{E}_\theta)\, v\, dr d\zeta = -\int_0^Z \int_0^R \frac{\partial B_z}{\partial t}\, v\, r dr d\zeta, \quad \forall v \in V. \tag{48}$$

Remark A.3. One can choose to handle the boundary condition on $\mathcal{E}_\theta$ by using an integration by parts formula regarding r, which gives the variational formulation:

$$\int_0^Z \int_0^R r\,\mathcal{E}_\theta \frac{\partial v}{\partial r}\, dr d\zeta = \int_0^Z \int_0^R \frac{\partial B_z}{\partial t}\, v\, r dr d\zeta, \quad \forall v \in V. \tag{49}$$

From these variational formulations, one derives the finite element conforming P_1 and P_2 approximations in an efficient way by using the FreeFem++ package [18]. The time discretization is performed with a classical finite difference scheme. As mentioned above, time discretization is not an issue here. Indeed, since the time derivative only appears in the right-hand sides of the variational formulations, there is no necessity to satisfy any stability condition, in the spirit of CFL condition.

A particle approximation

The paraxial Vlasov equation is numerically solved by means of a particle method [6]. One approximate the function $rf(\mathbf{X}, \mathbf{P}, t)$ at any time t by a linear combination of delta distributions in the phase space $(\mathbf{X}, \mathbf{P})$, namely:

$$rf(\mathbf{X}, \mathbf{P}, t) = \sum_k w_k\, \delta(\mathbf{X} - \mathbf{X}_k(t))\delta(\mathbf{P} - \mathbf{P}_k(t)), \tag{50}$$

where w_k denotes the constant weight of the particle k. Its position in the phase space $\mathbf{X}_k = (r, \zeta)$ and $\mathbf{P}_k = (p_r, p_\theta, p_z)$ is solution to the differential system:

$$\begin{cases} \dfrac{dr}{dt} = \dfrac{1}{\gamma m} p_r, & \dfrac{d\zeta}{dt} = c - \dfrac{1}{\gamma m} p_z, \\ \dfrac{dp_r}{dt} = \dfrac{1}{\gamma m r} p_\theta^2 + F_r, & \dfrac{dp_\theta}{dt} = -\dfrac{1}{\gamma m r} p_r p_\theta + F_\theta, \quad \dfrac{dp_z}{dt} = F_z, \end{cases} \tag{51}$$

together with initial conditions.

The corresponding particle charge and current densities ρ and $\mathbf{J}$ are obtained by introducing the particle approximation (50) in the following equations:

$$\rho = q \int f d\mathbf{P}, \qquad \mathbf{J} = q \int \mathbf{V}(\mathbf{P}) f d\mathbf{P}. \tag{52}$$

Then, we get:

$$r\rho(\mathbf{X},t) = q\sum_k w_k \delta(\mathbf{X}-\mathbf{X}_k(t)), \tag{53}$$

and

$$r\mathbf{J}(\mathbf{X},t) = q\sum_k w_k \mathbf{V}_k(t)\delta(\mathbf{X}-\mathbf{X}_k(t)). \tag{54}$$

Such expressions, built at the particle positions, cannot be used in this form for solving field equations. Indeed, a P_1 finite element approximation requires values of ρ and $\mathbf{J}$ at the vertices of the triangular mesh (denoted ρ^M and $\mathbf{J}^M$), whereas a P_2 approximation needs these values at the vertices and at the middle of the edges of the triangular mesh (also denoted ρ^M and $\mathbf{J}^M$). Following the classical procedure [6,19], we introduce the assignment and interpolation procedures.

Let us denote by $\{\mathbf{a}_i = (r_i,\zeta_i)\}$ the set of mesh vertices where ρ^M and $\mathbf{J}^M$ are needed, namely the vertices of the triangular mesh (P_1 finite element) or the vertices and the middle of the edges of the triangular mesh (P_2 finite element). Let χ_i be the finite element basis functions (for P_1 or P_2) and let $\mathbb{M}$ be the mass matrix with entries $\mathbb{M}_{i,j} = \int_\Omega \chi_i\,\chi_j r dr d\zeta$. We define the values of the charge and the current densities at the nodes of the mesh by:

$$\sum_j \mathbb{M}_{i,j}\,\rho^M(\mathbf{a}_j,t) = \int_\Omega \rho(\mathbf{X},t)\,\chi_i\,r\,dr\,d\zeta = q\sum_{k\in\mathcal{K}_{\mathbf{a}_i}} w_k\,\chi_i(\mathbf{X}_k(t)), \tag{55}$$

and

$$\sum_j \mathbb{M}_{i,j}\,\mathbf{J}^M(\mathbf{a}_i,t) = \int_\Omega \mathbf{J}(\mathbf{X},t)\,\chi_i\,r\,dr\,d\zeta = q\sum_{k\in\mathcal{K}_{\mathbf{a}_i}} w_k\,\mathbf{V}_k(t)\,\chi_i(\mathbf{X}_k(t)), \tag{56}$$

$\mathcal{K}_{\mathbf{a}_i}$ being the set of particle located in the element which have $\mathbf{a}_i$ as a vertex.

Since $\sum_i \chi_i \equiv 1$, the assignment procedure (55)–(56) preserves the total charge and the total current of particles in the sense that

$$\sum_i \left(\sum_j \mathbb{M}_{i,j}\,\rho^M(\mathbf{a}_j,t)\right) = \int_\Omega \rho(\mathbf{X},t)\,r\,dr\,d\zeta = q\sum_k w_k, \tag{57}$$

and

$$\sum_i \left(\sum_j \mathbb{M}_{i,j}\,\mathbf{J}^M(\mathbf{a}_i,t)\right) = \int_\Omega \mathbf{J}(\mathbf{X},t)\,r\,dr\,d\zeta = q\sum_k w_k\mathbf{V}_k(t). \tag{58}$$

Remark A.4. In the case of P_1 finite element, one often performs a mass lumping (or diagonalization) of the mass matrix $\mathbb{M}$. The assignment procedure (55)–(56) is expressed as:

$$\rho^M(\mathbf{a}_i,t) = \frac{1}{V_i}\int_\Omega \rho(\mathbf{X},t)\,\chi_i\,r\,dr\,d\zeta = \frac{q}{V_i}\sum_{k\in\mathcal{K}_{\mathbf{a}_i}} w_k\,\chi_i(\mathbf{X}_k(t)), \tag{59}$$

and

$$\mathbf{J}^M(\mathbf{a}_i,t) = \frac{1}{V_i}\int_\Omega \mathbf{J}(\mathbf{X},t)\,\chi_i\,r\,dr\,d\zeta = \frac{q}{V_i}\sum_{k\in\mathcal{K}_{\mathbf{a}_i}} w_k\,\mathbf{V}_k(t)\,\chi_i(\mathbf{X}_k(t)), \tag{60}$$

where $V_i = s\int_\Omega \chi_i\,r\,dr\,d\zeta$ is the volume associated to the node $\mathbf{a}_i$, or equivalently the diagonal term $\mathbb{M}_{i,i}$ appearing in the lumped mass matrix.

According to the general approach [6,19,8], time discretization of system (51) is built from a leapfrog scheme, which is a second-order centered finite-difference scheme. The particle positions are defined at time t_n and the particle

momenta are computed at time $t_{n+1/2}$. Eqs. (51) which correspond to the momentum $\mathbf{P}_k$, are approximated by

$$\begin{cases} \dfrac{1}{\Delta t}\left(p_r^{n+\frac{1}{2}} - p_r^{n-\frac{1}{2}}\right) = \dfrac{1}{\gamma^n m r^n} p_\theta^{n+\frac{1}{2}} p_\theta^{n-\frac{1}{2}} + F_r^n, \\ \dfrac{1}{\Delta t}\left(p_\theta^{n+\frac{1}{2}} - p_\theta^{n-\frac{1}{2}}\right) = -\dfrac{1}{2\gamma^n m r^n}\left(p_r^{n+\frac{1}{2}} p_\theta^{n-\frac{1}{2}} + p_r^{n-\frac{1}{2}} p_\theta^{n+\frac{1}{2}}\right) + F_\theta^n, \\ \dfrac{1}{\Delta t}\left(p_z^{n+\frac{1}{2}} - p_z^{n-\frac{1}{2}}\right) = F_z^n, \end{cases} \tag{61}$$

where $(F_r^n, F_\theta^n, F_z^n)$ is a numerical approximation at the time t_n of the Lorentz force $\mathbf{F} = (F_r, F_\theta, F_z)$ given by

$$\begin{cases} F_r = q(\mathcal{E}_r + v_\theta B_z + v_\zeta E_r/c), \\ F_\theta = q(\mathcal{E}_\theta - v_r B_z), \\ F_z = q(E_z + v_r E_r/c) \end{cases} \tag{62}$$

the computation of which requiring the knowledge of the electromagnetic fields. Since they are determined by finite element methods, an interpolation procedure is necessary to recover the values of the fields at the particle locations. For this purpose, we use an interpolation procedure, similar to the one proposed in [5], where the fields are computed by a finite difference method.

The last step consists in computing the particle position solution to (51), that are obtained by solving the following discretized system

$$\begin{cases} \dfrac{1}{\Delta t}(r^{n+1} - r^n) = \dfrac{1}{\gamma^{n+\frac{1}{2}} m} p_r^{n+\frac{1}{2}}, \\ \dfrac{1}{\Delta t}(\zeta^{n+1} - \zeta^n) = c - \dfrac{1}{\gamma^{n+\frac{1}{2}} m} p_z^{n+\frac{1}{2}}, \end{cases} \tag{63}$$

where $\gamma^{n+\frac{1}{2}}$ is computed with $\gamma^{n+\frac{1}{2}} = \left(1 + \frac{|\mathbf{P}^{n+\frac{1}{2}}|^2}{(mc)^2}\right)^{\frac{1}{2}}$. The final complete time advance algorithm has the same structure as the one described in [5], where Maxwell equations were approached by a finite-difference method. For more details, we refer the interested reader to this reference.

References

[1] F. Assous, J. Chaskalovic, Data mining techniques for scientific computing: Application to asymptotic paraxial approximations to model ultra-relativistic particles, J. Comput. Phys. 230 (2011) 4811–4827.

[2] F. Assous, J. Chaskalovic, Error estimate evaluation in numerical approximations of partial differential equations: A pilot study using data mining methods, C.R. Méc. 341 (2013) 304–313.

[3] F. Assous, J. Chaskalovic, Indeterminate constants in numerical approximations of PDEs: A pilot study using data mining techniques, J. Comput. Appl. Math. 270 (2014) 462–470.

[4] F. Assous, F. Tsipis, A PIC method for solving a paraxial model of highly relativistic beams, J. Comput. Appl. Math. 227 (1) (2008) 136–146.

[5] F. Assous, F. Tsipis, Numerical paraxial approximation for highly relativistic beams, Comput. Phys. Comm. 180 (2009) 1086–1097.

[6] C.K. Birdsall, A.B. Langdon, Plasmas Physics Via Computer Simulation, Mac. Graw-Hill, New York, 1985.

[7] K. Black, Business Statistics For Contemporary Decision Making, seventh ed., Wiley, India, USA, 2004.

[8] J.P. Boris, Proc. Fourth. Conf. Numerical Simulation of Plasmas, Naval Res. Lab., Washington D.C, 1970, p. 3.

[9] J. Chaskalovic, A new approach in media/marketing databases explorations for application in e-business, National congress of IREP, Paris, 1999.

[10] J. Chaskalovic, Mathematical and Numerical Methods For Partial Differential Equations, Springer Verlag, Switzerland, 2014.

[11] J. Chaskalovic, A. Vanheuverzwyn, Innovation in estimations: A reliable approach for radio audience indicators, 0, 195–210—Netherlands, in: Proc. Esomar, WM^3, Dublin, 2007.

[12] P.G. Ciarlet, Basic error estimates for elliptic problems, in: P.G. Ciarlet, J.-L. Lions (Eds.), Handbook of Numerical Analysis, Vol. II, North Holland, 1991, pp. 17–351.

[13] J. Cohen, P. Cohen, S.G. West, L.S. Aiken, Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences, third ed., Lawrence Erlbaum Associates Inc. Publishers, New Jersey, 2002.

[14] G.W. Corder, D.I. Foreman, Nonparametric Statistics For Non-Statisticians: A Step-By-Step Approach, Wiley, New Jersey, 2009.

[15] P. Degond, P.-A. Raviart, On the paraxial approximation of the stationary Vlasov-Maxwell, Math. Models Methods Appl. Sci. 3 (4) (1993) 513–562.

[16] Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, No. AIAA-G-077-1998, American Institute of Aeronautics and Astronautics, Reston, VA, 1998.
[17] F.E. Harrell, Regression Modeling Strategies, Springer, New York, 2001.
[18] F. Hecht, FreeFem++, Numerical mathematics and scientific computation 3.7, Laboratoire J. L. Lions, Université Pierre et Marie Curie, 2010. http://www.freefem.org/ff++/.
[19] R.W. Hockney, J.W. Eastwood, Computer Simulation Using Particles, Adam Hilger imprint by IOP Publishing Ltd, 1988.
[20] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., Wiley, New Jersey, 2000.
[21] H.P. Hsu, Schaum's outline of probability, Random Variables, and Random Processes, second edition, USA, 2010.
[22] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction To Statistical Learning, Springer, New York, 2013.
[23] O. Kulski, J. Chaskalovic, M. Plachot, J.M. Mayenga, A. Chouraqui, F. Abirached, A.M. Serkine, J. Belaisch-Allart, Explicative factors for prognostics IIU: exploration on 2089 cycles done with statistical and data mining tools, 9th meeting of the French Federation of the Reproduction Studies, Palais des Congrés—Paris, 2004.
[24] G. Laval, S. Mas-Gallic, P.A. Raviart, Paraxial approximation of ultrarelativistic intense beams, Numer. Math. 69 (1) (1994) 33–60.
[25] R. Lefébure, G. Venturi, Data Mining—Gestion De La Relation Client, Eyrolles, France, 2001.
[26] M.A. Mostrom, DI. Mitrovich, DI.R. Welch, The ARCTIC charged particle beam propagation code, J. Comput. Phys. 128 (2) (1996) 489–497.
[27] X.L. Nguyên, J. Chaskalovic, D. Rakotonanahary, B. Fleury, Insomnia symptoms and CPAP compliance in OSAS patients: A descriptive study using data mining methods, Sleep Med. (2010).
[28] X.L. Nguyên, D. Rakotonanahary, J. Chaskalovic, C. Philippe, C. Hausser-Hauw, B. Lebeau, B. Fleury, Residual subjective daytime sleepiness under CPAP treatment in initially somnolent apnea patients: a pilot study using data mining methods, Sleep Med. 9 (5) (2007) 511–516.
[29] R. Nuzzo, Scientific method: Statistical errors, Nature 506 (7487) (2014) 150. http://dx.doi.org/10.1038/506150a.
[30] P. Petterson, G. Iaccarino, J. Nordström, Numerical analysis of the Burgers equation in the presence of uncertainty, J. Comput. Phys. 228 (22) (2009) 8394–8412.
[31] P.A. Raviart, E. Sonnendrücker, A hierarchy of approximate models for the Maxwell equations, Numer. Math. 73 (3) (1996) 329–372.
[32] S. Slinker, G. Joyce, J. Krall, R.F. Hubbard, ELBA—A three dimensional particle simulation code for high current beams, Proc. of the 14th Inter. Conf Numer. Simul. Plasmas, Annapolis, 1991.
[33] J.V. Uspensky, Introduction To Mathematical Probability, McGraw-Hill, New York, 1937.