Mathematical Modelling and Analysis
Volume 22 Number 1, January 2017, 106–120
https://doi.org/10.3846/13926292.2017.1272499
© Vilnius Gediminas Technical University, 2017
Publisher: Taylor&Francis and VGTU
http://www.tandfonline.com/TMMA
ISSN: 1392-6292
eISSN: 1648-3510

Probabilistic Approach to Characterize Quantitative Uncertainty in Numerical Approximations

Joel Chaskalovic[a] **and Franck Assous**[b]

[a] *D'Alembert, University Pierre and Marie Curie*
Paris, France
[b] *Department of Mathematics, Ariel University*
40700 Ariel, Israel
E-mail(*corresp.*): franckassous55@gmail.com
E-mail: joel.chaskalovic@upmc.fr

Received 28 February, 2016; revised 9 December, 2016; published online January 5, 2017

Abstract. This paper proposes a statistical and probabilistic approach to compare and analyze the errors of two different approximation methods. We introduce the principle of numerical uncertainty in such a process, and we illustrate it by considering the discretization difference between two different approximation orders, e.g., first and second order Lagrangian finite element. Then, we derive a probabilistic approach to define and to qualify equivalent results. We illustrate our approach on a model problem on which we built the two above mentioned finite element approximations. We consider some variables as physical "predictors", and we characterize how they influence the odds of the approximation methods to be locally "same order accurate".

Keywords: probabilistic models, data mining, quantitative uncertainty, finite elements, Big Data.

AMS Subject Classification: 65P25; 77C10.

1 Introduction

With the development of super-computers, numerical methods produce today a huge quantity of numerical results, referred as big data, particularly for unsteady and three dimensional problems, in which approximations of vector fields or tensor components are computed. It seems thus interesting to look at new tools to analyze the corresponding simulated big data. Probabilistic and statistical methods could be (part of) such tools, that would help to characterize and compare quantitative uncertainty in approximate results of partial differential equations. In previous papers [1, 2], we have started to apply data mining techniques to scientific computing. In particular, data mining was used there as a tool to compare and evaluate *a posteriori* different asymptotic models. We relied on the fact that data mining techniques have already proved to be efficient in other contexts which deal with big data, like in biology [11,16], medicine [19,20], marketing [15,17], advertising and communications [5,7]. In this article, we propose a probabilistic approach to model and compute *a posteriori* quantitative uncertainty in numerical approximations of partial differential equations. It concerns the treatment of errors involved in an approximation process, which describes a given real system by a mathematical model, solved by numerical approximation methods, and whose exploitation, *in fine*, will provide the understanding, the control and the forecast of this system. In that sense, it can be used to compare two numerical solutions which may be based on different modeling choices, discretization processes, or solver strategies. Let us illustrate our purpose on basic examples. Let us consider two mathematical models used to describe a given system, for instance the Navier-Stokes and the Stokes equations in fluid mechanics. Our aim will be to evaluate on the results the relative quality of the two models, namely to define and measure a kind of modeling error between them. Ideally, one would like to evaluate and compare the gap between each of the models and the given system. However, as the actual complete rules of the system are generally unknown, we focus on comparison between the two models. In the case where observed data are available, another application could be to determine the closeness of the solution approximation to observed data. As another example, consider two finite element approximations, says P_k and $P_{k'}$ $(0 \leq k < k')$, applied for finding approximate solutions to a given problem. Based on classical finite element estimates, and under conditions of regularity of the mesh and of the solution, the numerical results computed by the $P_{k'}$ finite element method converge faster to the exact solution than those computed by the P_k one. However, due to the presence of uncertainty which appears in the standard error estimates, through the unknown constants, situations where P_k and $P_{k'}$ finite elements lead to equivalent results are possible (see for example [3]). In [1], we introduced and discussed separately different types of errors (modeling, approximations, etc.), and we focus on one of them. In reality, however, these different types of errors co-exist, and it can be informative to investigate the possible interferences. The approach proposed in this article is based on numerical computed results, warehoused in a database, and could allows to study these interferences. Continuing with the same example above, one could define a database made of finite difference approximation of the Navier-Stokes equations, that one wants to compare to a finite element approximation of Stokes equations. The proposed method here can characterize the errors involved in the approximation process, for instance the regions of space (and time if any) that two approximations "differ" or are "comparable", in a sense we will define. Obviously, It is not realistic to assume that any simulation analyst would construct several approximation methods of the problem, and then, want to figure out where the cheaper (or more efficient) approximation was adequate. The basic idea is rather to get a quite general indication, depending on the given circumstances, on the best adapted approximation method for a given class of problems. In other words, our aim is to identify a kind of stochastic behavior in the approximation process which will justify our stochastic approach. In our view, this is part of the topic to automatically characterize the approximation errors via statistical, probabilistic and unsupervised or supervised learning techniques implemented in data mining for big data. This paper is organized as follows. In Section 2, we introduce the principle of quantitative uncertainty for a variational formulation. We then illustrate it by considering two different finite elements approximation of a given elliptic standard problem. In particular, we highlight how a quantitative uncertainty can appear in finite elements error estimates. In Section 3, we derive a probabilistic approach, that allows us to qualify by logistic regressions the equivalent results. Then, we introduce in Section 4 a model problem and we derive a linear and a quadratic finite element approximations to illustrate on the corresponding simulated data our approach. Concluding remarks are finally drawn.

2 Quantitative uncertainty and approximation process

We seek to better understand how two different approximations of the same problem could produce numerical results with comparable accuracy. When we speak about approximations, we refer to any kind of errors, that is the modeling error, the approximation error, the discretization error or interferences of the different types, as introduced in [1]. Let $\mathbf{u}$ denote a reference solution (for instance, the exact solution, if known) to the system one wants to solve. We denote by $\mathbf{u}_{approx}$ its approximation, in the sense introduced above. Our aim is to characterize in a probabilistic sense, the error define by the "distance" between $\mathbf{u}$ and $\mathbf{u}_{approx}$. It is generally measured by a well adapted norm, and depends on the considered approximations.

Remark 1. In an "ideal case", the reference solution should be the exact one. But, in most situations, this exact solution is unknown, which motivates its approximation! In these conditions, one only will be able to assess the distance between two different kinds of approximations. This approach is classic for instance in the Model Order Reduction approach, in which the unknown exact solution is replaced by a "truth approximation", computed in a "very high dimension" subspace of the space of solution. This "truth approximation" is assumed to be "very close" to the exact one [18].

Our idea consists in building the database made of $\mathbf{u}$ and $\mathbf{u}_{approx}$, or two different computed $\mathbf{u}_{approx}$, and to derive probabilistic tools to measure equivalent results. In other words, we will assume a kind of stochastic behavior in our data, which will justify our stochastic approach. Moreover, this approach can obviously be applied to more than two approximations. To be more concrete, let us illustrate the quantitative uncertainty in the case of two different finite elements approximations. For this kind of error, we suspect that the numerical uncertainty is mainly due to the presence of uncertainty in the unknown constants involved in the error estimates. For our purpose here, we shortly recall the main steps involved in the derivation of error estimates (for more details see [3]).

Remark 2. For the sake of simplicity, we consider in what follows an elliptic standard problem, that is not time-dependent. When looking at a time-dependent problem, after time discretization, one generally solves a sequence of stationary problems, one for each time step.

Let Ω be a regular subset of $\mathbb{R}^2$ and $\mathbf{u}$ a vector field defined from Ω to $\mathbb{R}^M$, $M \in \mathbb{N}, M \geq 1$. We introduce a Hilbert space $(\mathbf{V}, \|.\|_{\mathbf{V}})$ (product of M Hilbert spaces V_m, $(1 \leq m \leq M)$), and a bilinear, continuous, and $\mathbf{V}$-elliptic form $a(\cdot,\cdot)$ defined on $\mathbf{V}\times\mathbf{V}$. We denote by $l(\cdot)$ a linear continuous form defined on $\mathbf{V}$, and by $\mathbf{V}_h$ a given finite dimension subset of $\mathbf{V}$. We consider a standard abstract elliptic variational formulation and its approximation, defined as follows:

$$\begin{cases} \text{Find } \mathbf{u} \in \mathbf{V} \text{ solution to:} \\ a(\mathbf{u},\mathbf{v}) = l(\mathbf{v}), \forall \mathbf{v} \in \mathbf{V}, \end{cases} \qquad \begin{cases} \text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ solution to:} \\ a(\mathbf{u}_h,\mathbf{v}_h) = l(\mathbf{v}_h), \forall \mathbf{v}_h \in \mathbf{V}_h. \end{cases}$$

We aim to illustrate how two different finite element approximations could produce equivalent (in a sense that we will define) numerical results. For our purpose, we first introduce a finite element partition $\mathcal{M}_h$ of Ω, assumed to exactly coincide with Ω. Let $\mathbf{u}_h^{(1)}$ denote the P_1 finite element approximation of $\mathbf{u}$, we have the classical following result (see for instance [6]).

Lemma 1. *Let h denotes the mesh size of $\mathcal{M}_h$. Assume that Ω is a convex polygonal domain and suppose the exact solution $\mathbf{u}$ belongs to $\left[C^2(\Omega)\right]^M$. The approximation $\mathbf{u}_h^{(1)}$ converges to $\mathbf{u}$ when h tends to zero and we have the following global error estimate:*

$$\|\mathbf{u} - \mathbf{u}_h^{(1)}\|_{[H^1(\Omega)]^M} \leq \gamma_1 h\,.$$

Above, $\|.\|_{[H^1(\Omega)]^M}$ denotes the standard Sobolev norm of $\left[H^1(\Omega)\right]^M$ and γ_1 is an unknown constant made up of the unknown value of $\|D^2\mathbf{u}\|_{[L^\infty(\Omega)]^M}$, on the first hand, and the unknown ellipticity constant associated to the bilinear form $a(.,.)$, on the other hand (see ([3])). Similar results can be derived for $\mathbf{u}_h^{(2)}$, the P_2 finite element approximation of $\mathbf{u}$. In this case, one obtains, assuming the exact solution $\mathbf{u}$ belonging to $\left[C^3(\Omega)\right]^M$:

$$\|\mathbf{u} - \mathbf{u}_h^{(2)}\|_{[H^1(\Omega)]^M} \ \leq \ \gamma_2 h^2,$$

where γ_2 is the unknown constant for the P_2 finite element approximation analogous to γ_1. Therefore, because the presence of the two unknown constants γ_1 and γ_2, which contain numerical uncertainty, we suspect the following numerical *local* situation to take place: $\exists\, m \in \{1,\ldots,M\}, \exists\, x \in \Omega$ such that:

$$|u_m(x) - u_{h,m}^{(1)}(x)| \ \leq \ |u_m(x) - u_{h,m}^{(2)}(x)|\ ,$$

or at least

$$|u_m(x) - u_{h,m}^{(1)}(x)| \ \simeq \ |u_m(x) - u_{h,m}^{(2)}(x)|. \tag{2.1}$$

Inequations (2.1) mean that *locally*, i.e. for some x in Ω, P_1 finite elements might be either more accurate than P_2 finite elements, or at least equivalent, regarding the component u_m of the exact solution $\mathbf{u}$. We also remark that $x \in \Omega$, and $m \in \{1,\ldots,M\}$ are not necessarily unique. Similar uncertainties can be identified for the other different kinds of errors between $\mathbf{u}$ and $\mathbf{u}_{approx}$, or between two different $\mathbf{u}_{approx}$. In the following, our purpose will be to characterize local features such that two approximations could numerically be equivalent. To this end, we will consider stochastic approach motivated by this kind of stochastic behavior in the approximation process. Moreover, one relies on the fact that data mining techniques are appropriate to explore a given database to identify, if any, and to characterize homogenous subgroups, which corresponds to *local* properties of the database.

3 Quantification of equivalent approximation results

From now on, we will assume that the reference solution is unknown, and we will expose our approach to characterize two different approximations (not necessary finite element ones), that we will denote $\mathbf{u}_h^{(1)}$ and $\mathbf{u}_h^{(2)}$. However, it can be written in the same way between the reference solution and an approximate solution, if the former is known.

3.1 A probabilistic approach to measure equivalent results

For a given value m, $(1 \leq m \leq M)$, we consider the m-th component of the two different approximations $\mathbf{u}_h^{(1)}$ and $\mathbf{u}_h^{(2)}$ one wants to investigate, that we denote by $u_h^{(k)}, (k=1,2)$. Again, $u_h^{(1)}$ or $u_h^{(2)}$ can denote the m-th component of the reference solution, for instance in the context of Model Order Reduction method. We first build a database which consists of all the M components of the approximations $\mathbf{u}_h^{(k)}, (k=1,2)$ computed by the two approximation methods at common degree of freedom. Typically, this can be computed values at the vertices of a common mesh, like $\mathcal{M}_h$ in the illustration above. We denote by N the total number of rows in the database. Then, we introduce the Bernoulli random variable X_{u_h} whose trace on the N observations in the database is defined by:

$$\forall l = 1,\ldots,N : (X_{u_h})_l \equiv \left| \begin{array}{ll} 1, & \text{if } |u_{h,l}^{(2)} - u_{h,l}^{(1)}| \leq \alpha \max\limits_{j=1,N} |u_{h,j}^{(2)} - u_{h,j}^{(1)}|, \\ 0, & \text{if not,} \end{array} \right. \tag{3.1}$$

where $u_{h,l}^{(k)}, (k=1,2)$, denotes the trace of the approximation $u_h^{(k)}$ at a given common degree of freedom, and α is threshold to be defined.

The variable X_{u_h} defines to what extend the approximations $u_h^{(2)}$ and $u_h^{(1)}$ are assumed to be equivalent. This definition strongly depends on the threshold α. To evaluate this threshold, we introduce n, the size of the equivalent systematic sampling, defined such that $u_h^{(2)}$ and $u_h^{(1)}$ do not "differ significantly" between the sampling and the database. To determine the value of n, one can apply the non parametric Kolmogorov-Smirnov test [9], (with a standard p-value equal to 0.05), since no features are known regarding the shape of the distributions of $u_h^{(2)}$ and $u_h^{(1)}$. Practically, to get the optimal value of n, one has to successively process the above statistical tests in order to achieve the statistical significance, parameterized by the p-value. Therefore, for a given value of α, we denote p_α the uniform probability of picking an index l $(1 \le l \le N)$ from the database, such that the approximations $u_{h,l}^{(2)}$ and $u_{h,l}^{(1)}$ will have the same numerical order. In other words, we have

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

where the random variable Y is defined by $Y = |u_h^{(2)} - u_h^{(1)}|$, Y_j being the trace of Y on any element j in the database. Let us now consider a systematic sampling of n elements, $(n < N)$, and the individual Bernoulli random variables $X_i, (i = 1, \dots, n)$, defined by:

$$X_i = \left| \begin{array}{ll} 1, & \text{if } \ Y_i < \alpha \max\limits_{1 \le j \le N} Y_j, \\ 0, & \text{if not,} \end{array} \right.$$

where Y_i denotes the trace of Y on any element $i, (1 \le i \le n)$, which belongs to the sampling. We have:

$$\forall i \in \{1, \dots, n\} : Prob\{X_i = 1\} = Prob\{Y_i < \alpha \max_{1 \le j \le N} Y_j\} = p_\alpha. \tag{3.2}$$

We also introduce the random variable X which allows us to count the number of all the individuals in the sampling which are qualified *"Equivalent Results"*:

$$X = \sum_{1 \le i \le n} X_i. \tag{3.3}$$

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: $\mu_X = np_\alpha$ and $\sigma_X^2 = np_\alpha(1 - p_\alpha)$. Finally, we introduce the frequency of *"Equivalent Results"* in the sampling which corresponds to the random variable X/n. Our purpose is now to guarantee that X/n measured on the sampling does not diverge "too much" from p_α evaluated on the whole population. This can be expressed by

$$Prob\,\{|X/n - p_\alpha| \le \epsilon p_\alpha\} \ge S, \tag{3.4}$$

for S a given confidence level, and ϵ a parameter, $\epsilon \in [0, 1]$. We prove the following result

Theorem 1. *Let X be the binomial variable $\mathcal{B}(n, p_\alpha)$ defined by (3.3) such that $np_\alpha(1 - p_\alpha) \ge 25$, with p_α defined by (3.2). Let $(\epsilon, S') \in [0, 1]^2$ and p_α^* the smallest value of p_α solution to:*

$$2 \int_0^{\epsilon \sqrt{\frac{np_\alpha}{(1-p_\alpha)}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt \ge S'. \tag{3.5}$$

Then,

$$Prob \left\{ \left| \frac{X^* - \mu_{X^*}}{\sigma_{X^*}} \right| \le \epsilon \sqrt{\frac{np_\alpha^*}{(1 - p_\alpha^*)}} \right\} \ge S' - \frac{C}{\sqrt{np_\alpha^*(1 - p_\alpha^*)}}, \tag{3.6}$$

where X^* denotes the binomial variable $\mathcal{B}$ (n, p_α^*) and C is a positive constant lower than 0.588.

Proof. Let X be the binomial variable $\mathcal{B}\ (n, p_\alpha)$. Under suitable conditions that will be explained later, we have:

$$Prob\left\{\left|\frac{X-\mu_X}{\sigma_X}\right| \le \epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}}\right\} \approx 2\int_0^{\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt. \tag{3.7}$$

Hence, taking into account inequality (3.5) and its associated smallest value p^*, we have:

$$Prob\left\{\left|\frac{X^*-\mu_{X^*}}{\sigma_{X^*}}\right| \le \epsilon\sqrt{\frac{np^*_\alpha}{(1-p^*_\alpha)}}\right\} \ge S'. \tag{3.8}$$

Given that (3.7) is an approximation, the corresponding error estimate is mainly due to Uspensky [21]. In our case, by directly applying his result, one can prove that:

$$\left|Prob\left\{\left|\frac{X^*-\mu_{X^*}}{\sigma_{X^*}}\right| \le \epsilon\sqrt{\frac{np^*_\alpha}{1-p^*_\alpha}}\right\} - 2\int_0^{\epsilon\sqrt{\frac{np^*_\alpha}{1-p^*_\alpha}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt\right| \le \frac{C}{\sqrt{np^*_\alpha(1-p^*_\alpha)}}, \tag{3.9}$$

where $C < 0.588$ [21]. So, using (3.9) in the approximation (3.7) implies that (3.6) holds from (3.8). □

Remarks:

1. The condition $np_\alpha(1-p_\alpha) \ge 25$ is the one mentioned in Uspensky [21], which allows us to approximate the binomial law by the normal one, with the corresponding error estimate (3.9).

2. For a given value of the confident level S', inequation (3.5) can be solved to get p^*_α, using a table of standard normal distributions [14].

3. With elementary transformations, one can get our objective control given by (3.4) from (3.6), substituting p^*_α to p_α and setting:

$$S = S' - C/\sqrt{np^*_\alpha(1-p^*_\alpha)}.$$

4. Given the value p^*_α related to a given confident level S' solution to (3.6), the associated value α^* is processed by marginal distributions on the whole database. This guarantees that the value p^*_α fits the corresponding percent of rows which are qualified *"Equivalent Results"*.

3.2 Logistic regression and local qualification of equivalent results

We introduce the logistic regression [8, 13] to analyze and qualify the *"Equivalent Results"* class $\{X_{u_h} = 1\}$ defined in (3.1), according to the value of the parameter α^* defined above. In Subsection 3.1, we have considered the case where one of the M components of the approximation $\mathbf{u}_h$ satisfies inequality (2.1). Now, we take into account that the variable X_{u_h} is a function of all other available predictors (i.e. the other independent variables used to predict X_{u_h}), that is the space coordinates and the $M-1$ other approximations computed by the two methods one aims to compare. Denoting by Z_i $(1 \leq i \leq I)$ these I predictors, we aim to model the conditional probability $p(z_1, z_2, \ldots, z_I)$ defined by:

$$p(z_1, \ldots, z_I) \equiv Prob\,\{X_{u_h} = 1 | Z_1 = z_1, \ldots, Z_I = z_I\}.$$

We introduce now the odds $\omega(z_1, \ldots, z_I)$ of getting:

$$\Big\{X_{u_h}{=}1 \text{ if } (Z_1{=}z_1, \ldots, Z_I{=}z_I)\Big\} versus \Big\{X_{u_h}{=}0 \text{ if } (Z_1 = z_1, \ldots, Z_I = z_I)\Big\}$$

for given values $(z_1, \ldots, z_I)$ of the predictors $(Z_1, \ldots, Z_I)$, defined by:

$$\omega(z_1, \ldots, z_I) = \frac{p(z_1, \ldots, z_I)}{1 - p(z_1, \ldots, z_I)}. \tag{3.10}$$

Odds values are in $[0, +\infty[$ which allows us to consider the following linear regression, denoting by β_i $(0 \leq i \leq I)$ the coefficients of the regression

$$\ln\left(\frac{p(z_1, \ldots, z_I)}{1 - p(z_1, \ldots, z_I)}\right) = \beta_0 + \beta_1\, z_1 + \cdots + \beta_I z_I \equiv \beta_0 + \boldsymbol{\beta}.\mathbf{z}\,, \tag{3.11}$$

where, for simplicity, we introduced the vectors $\boldsymbol{\beta}$ and $\mathbf{z}$ that belong to $\mathbb{R}^I$. A very nice property deduced from (3.11) is the interpretation of the coefficients β_i. Indeed, let us introduce the odds ratio $o(z_i^{(0)}, z_i^{(1)})$ relatively to one of the predictors $Z_i, (i = 1, \ldots, I)$ defined by:

$$o(z_i^{(0)}, z_i^{(1)}) \equiv \frac{P\big(z_i^{(1)}\big)}{P\big(z_i^{(0)}\big)}, \tag{3.12}$$

$$P\big(z_i^{(k)}\big) = \frac{p(z_1, ..., z_{i-1}, z_i^{(k)}, z_{i+1}, ..., z_I)}{1 - p(z_1, ..., z_{i-1}, z_i^{(k)}, z_{i+1}, ..., z_I)}, \quad k = 0, 1.$$

where $z_i^{(0)}$ and $z_i^{(1)}$ correspond to two different values of the random predictor variable z_i. One can easily show the following lemma:

Lemma 2. *Equations (3.11) and (3.12) lead to:* $\beta_i = \ln(o(z_i + 1, z_i))$ *and* $o(z_i + c, z_i) = \exp(c\beta_i)$*,* $\forall c \in \mathbb{R}$ *and* $\forall i = 1, \ldots, I$*.*

In the next section this lemma will allow us to model and quantify the relationship between the *"Equivalent Results"* category $(X_{u_h} = 1)$ and the other predictors Z_i.

We now proceed to the evaluation of the coefficient β_0 and the vector $\boldsymbol{\beta} \in \mathbb{R}^I$ which determine the "linear" regression (3.11). We assume that, for each training independent data-point indexed by $k, (k = 1, \ldots, n)$, we have a vector of features $\mathbf{z}^{(k)} = (z_1^{(k)}, \ldots, z_I^{(k)})$ which belongs to $\mathbb{R}^I$, and an observed class $y^{(k)}$. The probability of that class was either $p(\mathbf{z}^{(k)})$, if $y^{(k)} = 1$, or $1 - p(\mathbf{z}^{(k)})$, if $y^{(k)} = 0$. The corresponding likelihood function $L(\beta_0, \boldsymbol{\beta})$ is then defined by:

$$L(\beta_0, \boldsymbol{\beta}) \equiv \prod_{k=1}^{n} p(\mathbf{z}^{(k)})^{y^{(k)}} \big(1 - p(\mathbf{z}^{(k)})\big)^{1 - y^{(k)}}. \tag{3.13}$$

So we have the following result:

Lemma 3. *Let $\widehat{\beta_i}, (i = 0, \ldots, I)$, the estimators of the parameters β_i introduced in (3.11) which maximize the likelihood function $L(\beta_0, \boldsymbol{\beta})$ defined in (3.13). Then, $\widehat{\beta_i}$ are solution to:*

$$\begin{cases} \sum\limits_{k=1}^{n} y^{(k)} z_q^{(k)} - \sum\limits_{k=1}^{n} \dfrac{z_q^{(k)}\, e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}}{1 + e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}} \; = \; 0, \quad \forall q = 1, \ldots, I, \\ \sum\limits_{k=1}^{n} y^{(k)} - \sum\limits_{k=1}^{n} \dfrac{e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}}{1 + e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}} = 0. \end{cases} \tag{3.14}$$

Proof. The log-likelihood turning products into sums, this yields, using (3)

$$\begin{aligned} \ln\Big(L(\beta_0, \boldsymbol{\beta})\Big) \quad &= \quad \sum_{k=1}^{n} y^{(k)} \ln\Big(p(\mathbf{z}^{(k)})\Big) + \sum_{k=1}^{n} (1 - y^{(k)}) \ln\Big(1 - p(\mathbf{z}^{(k)})\Big) \\ &= \quad \sum_{k=1}^{n} y^{(k)} \ln\left(\frac{p(\mathbf{z}^{(k)})}{1 - p(\mathbf{z}^{(k)})}\right) + \sum_{k=1}^{n} \ln\Big(1 - p(\mathbf{z}^{(k)})\Big) \\ &= \quad \sum_{k=1}^{n} y^{(k)} (\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}) + \sum_{k=1}^{n} \ln\Big(1 - p(\mathbf{z}^{(k)})\Big), \end{aligned} \tag{3.15}$$

where we also used (3.11). Moreover, from (3.11), we can isolate $p(\mathbf{z}^{(k)})$ as

$$p(\mathbf{z}^{(k)}) = \frac{e^{\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}}}, \quad \ln\big(1 - p(\mathbf{z}^{(k)}\big) \; = \; -\ln\big(1 + e^{\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}}\big).$$

Finally, the log-likelihood (3.15) can be written in its final expression:

$$\ln\Big(L(\beta_0, \boldsymbol{\beta})\Big) \; = \; \sum_{k=1}^{n} y^{(k)} (\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}) - \sum_{k=1}^{n} \ln\big(1 + e^{\beta_0 + \mathbf{z}^{(k)}.\boldsymbol{\beta}}\big).$$

So, the necessary conditions to maximize the likelihood function $L(\beta_0, \boldsymbol{\beta})$ are obtained by canceling each of its first order partial derivative with respect to $\beta_q, (q = 0, \ldots, I)$. Then, we get the following system:

$$\begin{cases} \dfrac{\partial \ln\Big(L(\widehat{\beta_0}, \widehat{\boldsymbol{\beta}})\Big)}{\partial \widehat{\beta_q}} = \sum\limits_{k=1}^{n} y^{(k)} z_q^{(k)} - \sum\limits_{k=1}^{n} \dfrac{z_q^{(k)}\, e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}}{1 + e^{\widehat{\beta_0}+z^{(k)}.\widehat{\boldsymbol{\beta}}}} = 0, \quad \forall q = 1, \ldots, I, \\ \dfrac{\partial \ln\Big(L(\widehat{\beta_0}, \widehat{\boldsymbol{\beta}})\Big)}{\partial \widehat{\beta_0}} = \sum\limits_{k=1}^{n} y^{(k)} - \sum\limits_{k=1}^{n} \dfrac{e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}}{1 + e^{\widehat{\beta_0}+\mathbf{z}^{(k)}.\widehat{\boldsymbol{\beta}}}} = 0. \end{cases} \tag{3.16}$$

Equations (3.16) can not be exactly solved but however are approximated by numerical schemes as Newton's method [10]. □

In the next section, we will consider a model problem and we will illustrate our approach by considering two different finite element approximations, as in Section 2. To better understand the quantitative process of approximation error, we have summarized below the workflow suggested by our method. For a given mathematical model described by a system of (partial) differential equations, we have to

1. construct the variational formulations,
2. choose the numerical approximations to be compared,
3. store the database made of the numerical approximations,
4. determine the sampling size n obtained by a goodness-of-fit test processed on the two approximations,
5. determine first the optimal p_{α}^{*} solution to (3.7), then the α_{*} obtained by marginal distributions on the whole database,
6. qualify equivalent approximations results processed by a logistic regression,
7. identify predictors which significantly increase or decrease the odds of being in the "Same Order" approximation.

4 Application to a model problem

In this section, we introduce a model problem, and, as in the beginning of the article, we apply our approach to a linear P_1 and a quadratic P_2 finite element approximations. For our purpose, we consider a quasi-static elliptic approximation of the Vlasov-Maxwell equations in a relativistic case [2]. This models the propagation of an electron bunch in an hollow cylindrical tube [4]. The Vlasov equation is approximated by a particle method whereas the electromagnetic field is discretized by P_1 and P_2 finite elements, using the FreeFem++ package [12]. We notice that as the right hand side of the quasi-static Maxwell equation is explicitly time dependent, the resulting problem consists of a sequence of elliptic problems solved at each time step. As a consequence, the database of approximations will include all these time steps.

Remark 3. This model, derived from plasma simulations, has the advantage to be "rich enough" to produce big data stored in 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.

In the sequel, we will illustrate the probabilistic framework we propose by characterizing the numerical uncertainty in the error estimates as described in (2.1), regarding the longitudinal component of the magnetic field denoted H_z. We denote by $H_z^{(1)}$ and $H_z^{(2)}$ the P_1 and P_2 finite element approximations of the reference solution H_z. More precisely, as exposed in Section 2, one is interested to investigate the situation characterized by

$$\left|H_z^{(1)}(r_{j^*},\zeta_{k^*},t_{n^*})-H_z(r_{j^*},\zeta_{k^*},t_{n^*})\right| \simeq \left|H_z^{(2)}(r_{j^*},\zeta_{k^*},t_{n^*})-H_z(r_{j^*},\zeta_{k^*},t_{n^*})\right|$$

for a discrete time t_{n^*} and a space node (r_{j^*},ζ_{k^*}) of a mesh $\mathcal{M}_h$, introduced for these finite element approximations. As described before, we construct the database made of the P_1 and P_2 computed solutions. It is composed by 125000 rows and by the 36 variables, as we considered 100 time steps t_n, 1250 space nodes (r_j,ζ_k), and at all, 36 physical variables in our simulations. The main results we have obtained are summarized below:

1. *Size of the sampling*: As explained in Subsection 3.1, without statistical features of the components $H_z^{(1)}$ and $H_z^{(2)}$, we processed the Kolmogorov-Smirnov test. Consequently, the equivalent sampling we consider is composed by 5800 rows of the database.

2. *Determination of α^* by the probabilistic model*: To apply Theorem 1 in our case, we choose $\epsilon = 5\%$ and $S' = 97\%$ to obtain the confident level S equals to the standard value which is 95%. Indeed, if $S' = 97\%$, one can show that (3.5) is satisfied when
$$\epsilon\sqrt{\frac{np_\alpha}{(1-p_\alpha)}} \geq 2.17 \iff p_\alpha \geq \frac{4.71}{4.71+n\epsilon^2}.$$
So, if $n = 5800$ and $\epsilon = 5\%$, we get $p_\alpha \geq p^* \simeq 0.245$ and its corresponding value $\alpha^* \simeq 0.75$. As a consequence, the right-hand side of the inequality (3.6) approximatively equals to 95%.

3. *Logistic regression and qualification of Equivalent Results for $H_z^{(1)}$ and $H_z^{(2)}$*: Given the previous value of α^*, we processed under the Data Mining platform *IBM SPSS Modeler* the logistic regression to qualify the dependency between the *"Equivalent Results"* category for $H_z^{(1)}$ and $H_z^{(2)}$ (named *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* in the sequel), namely the value of the corresponding random variable $X_{u_h} = 1$ defined in (3.1), with the time t and the spacial coordinates (r, ζ), as potential predictors. The corresponding model we have obtained presents the following properties:

- The equation of the logistic regression we found is described by:
$$\ln\left(\frac{Prob\{X_{u_h}=1|(t,r,\zeta)\}}{1{-}Prob\{X_{u_h}{=}1|(t,r,\zeta)\}}\right) = 0.01176t{+}0.01545r{-}0.1696\zeta, \quad (4.1)$$

where the three estimators $(\widehat{\beta_t}, \widehat{\beta_r}, \widehat{\beta_\zeta}) = (0.1176, 0.01545, -01696)$ of the coefficients β_t, β_r and β_ζ have been estimated by maximizing the corresponding likelihood function coupled with Newton's scheme, as explained in Subsection 3.2. As we saw above in Lemma 2, the meaning of the coefficient $\widehat{\beta_i}$ of (4.1) is easy to interpret by the help of $\exp(\widehat{\beta_i})$. 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 *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* 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.

(a) *Interpretation of the coefficient $\widehat{\beta_t}$ and exp($\widehat{\beta_t}$)*. According to the definition (3.10) of the odds and thanks to (3.12) and Lemma 2, we can say now that the coefficient $\widehat{\beta_t}$ of the time t in (4.1) is the difference in the log odds. In other words, for a one-unit increase in time, (i.e. a time step), the expected change in log odds is $\widehat{\beta_t} = 0.01176$. Can we translate this change in log odds to the change in odds? Indeed, by Lemma 2 we can say that, 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 if 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 Lemma 2, one must deal with an about 50% increase in the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* class after 35 time steps, ($exp(35 * 0.01176) \simeq 1.5$) and with about 224% increase in the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* 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 *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* class increases. Consequently, less useful and justified is the implementation of P_2 finite elements.

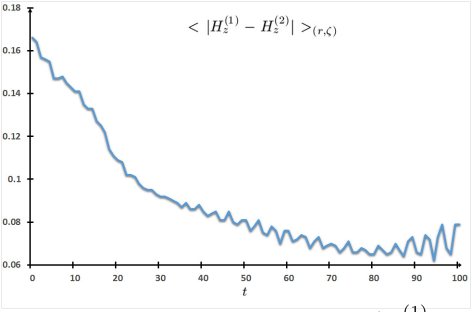

Figure 1. Time dependency of the average $< |H_z^{(1)} - H_z^{(2)}| >_{(r,\zeta)}$

This behavior can be illustrated below on Figure 1, where we plot $< |H_z^{(1)} - H_z^{(2)}| >_{(r,\zeta)}$, the average of $|H_z^{(1)} - H_z^{(2)}|$ computed over all r and ζ. As one can see, the more the time passes, the more the trend of this average decreases: this corresponds to equivalent numerical results between the P_1 and P_2 finite elements approximations of the magnetic component H_z.

(b) *Interpretation of the coefficient $\widehat{\beta_r}$ and $exp(\widehat{\beta_r})$.* In the same way, for a one-unit increase of r, the expected change in log odds is $\widehat{\beta_r} = 0.01545$. This change in log odds corresponds to an equivalent change in odds of about 1.55% increase in the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* 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 *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* class, (since $exp(26 * 0.01545) \simeq 1.5$). Finally, when $r = 120 := R$ (corresponding to the tube wall), the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* ratio is about 6.38 times more, (again, $exp(120 * 0.01545) \simeq 6.38$) compared with $r = 0$; the situation *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* is most likely when $r = R$ 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 H_z which constraints the solution to vanish at $r = R$. Here again, we illustrate this behavior on Figure 2 where the average $< |H_z^{(1)} - H_z^{(2)}| >_{(t,\zeta)}$ is depicted as a function of all the time and ζ. As one can see, this value strongly decreases as a function of r, corresponding to equivalent numerical results between the P_1 and P_2 approximations of the magnetic component H_z .

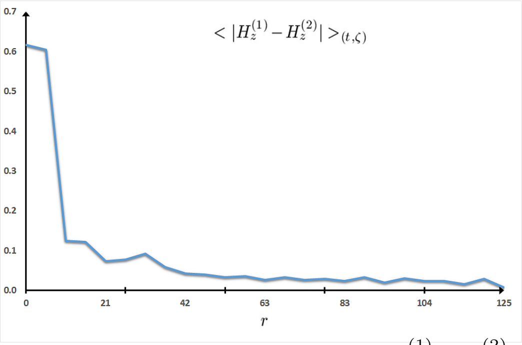

Figure 2. r-dependency of the average $< |H_z^{(1)} - H_z^{(2)}| >_{(t,\zeta)}$

(c) *Interpretation of the coefficient* $\widehat{\beta}_\zeta$ *and* $exp(\widehat{\beta}_\zeta)$. As we have $\widehat{\beta}_\zeta = -0.1696$, the presence of the sign "-" allows us conclude that the more ζ grows, the more the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* 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 *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* class. As a consequence, after 4 units of ζ one must deal with about 50% decrease in the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* 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 *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* is least likely, (the odds of being in the *"$H_z^{(1)} - H_z^{(2)}$ Equivalent Results"* class decreases about 92%), rather than at the beginning of the bunch, where $\zeta = 0$. Here again, as for the interpretation of the coefficient $\widehat{\beta}_t$, this behavior was not expected even after a close look of the equations, and is probably related to the non linear coupling of the solved system. Again, we plot on Figure 3 the average $< |H_z^{(1)} - H_z^{(2)}| >_{(r,t)}$, computed over all the time and r values. As one can see, when ζ takes high values at the end of the bunch, the curve increases which shows that the numerical approximations of H_z cannot be considered anymore equivalent.

5 Conclusion

In this paper, we have proposed a new approach that combines probabilistic techniques and statistical methods, to characterize and compute *a posteriori* quantitative uncertainty in approximation methods. We have highlighted the

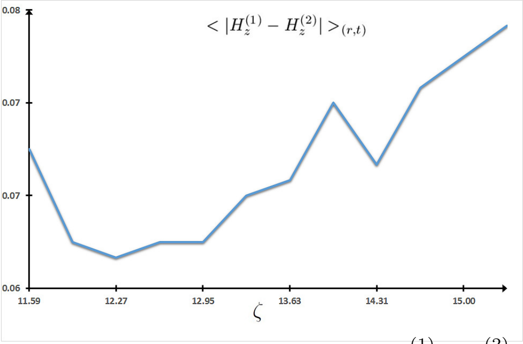

Figure 3. ζ-dependency of the average $< |H_z^{(1)} - H_z^{(2)}| >_{(r,t)}$

part of uncertainty contains in the methods, that can influence and even damage the precision of the computed numerical results.

In a second part, we have applied the approach by comparing a low order finite element method (P_1) to a high order one (P_2). We have derived a statistical and probabilistic approach to compare the two corresponding numerical approximations $\mathbf{u}_h^1$ and $\mathbf{u}_h^2$. This allows us to measure, then to qualify by logistic regression the notion of *"Equivalent Results"*. Then, we characterized the influence of predictors on the odds of being in the *"Equivalent Results"* category. Finally, we introduced, as a model problem, a quasi-static elliptic approximation of the Vlasov-Maxwell equations. We illustrate our approach by characterizing the *"Equivalent Results"* in that case. Future developments could concern the interaction between more than two predictors, for instance by introducing in the logistic regression equation, non linear terms like tr, $t\zeta$ and $r\zeta$. Another 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 two different approximations of partial differential equations. Another potential application for this work could be to determine the closeness of the solution approximation to observed data, considered as a reference or exact solution. This could be used as goodness-of-fit test for compare the model to the reality.

References

[1] F. Assous and J. Chaskalovic. Data mining techniques for scientific computing: Application to asymptotic paraxial approximations to model ultrarelativistic particles. *Journal of Computational Physics*, **230**(12):4811–4827, 2011. https://doi.org/10.1016/j.jcp.2011.03.005.

[2] F. Assous and J. Chaskalovic. Error estimate evaluation in numerical approximations of partial differential equations: A pilot study using data mining methods. *Comptes Rendus. Mécanique*, **341**(3):304–313, 2013.

[3] F. Assous and J. Chaskalovic. Indeterminate constants in numerical approximations of PDEs: A pilot study using data mining techniques. *Journal of Computational and Applied Mathematics*, **270**:462–470, 2014. https://doi.org/10.1016/j.cam.2013.12.015.

[4] C.K. Birdsall and A.B. Langdon. *Plasmas Physics Via Computer Simulation.* McGraw-Hill Book, New York, 1985.

[5] J. Chaskalovic. A new approach in media/marketing databases explorations for application in e-business. 1999.

[6] J. Chaskalovic. *Mathematical And Numerical Methods For Partial Differential Equations.* Springer Verlag, Switzerland, 2014. https://doi.org/10.1007/978-3-319-03563-5.

[7] J. Chaskalovic and A. Vanheuverzwyn. Innovation in estimations: A reliable approach for radio audience indicators. pp. 195–210, 2007.

[8] J. Cohen, P. Cohen, S.G. West and L.S. Aiken. *Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences.* Lawrence Erlbaum Associates, Mahwah, New Jersey, 2003.

[9] G.W. Corder and D.I. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.* Wiley, New Jersey,, 2009. https://doi.org/10.1002/9781118165881.

[10] M. Crouzeix and A.L. Mignot. *Analyse numérique des équations différentielles.* Masson, Differential equations, Paris, 1984.

[11] F.E. Harrell. *Regression Modeling Strategies.* Springer-Verlag, New York, 2001.

[12] F. Hecht. New development in freefem++. *Journal of Numerical Mathematics*, **20**(3-4):251–265, 2012.

[13] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression, second edition.* John Wiley & Sons, Inc., New Jersey, 2000. https://doi.org/10.1002/0471722146.

[14] H. Hsu. *Probability, Random Variables, and Random Processes, second edition.* Schaum's outlines Series, McGraw-Hill, New York, 2010.

[15] G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction To Statistical Learning.* Springer, New York, 2013.

[16] O. Kulski, J. Chaskalovic, M. Plachot, J.M. Mayenga, A. Chouraqui, F. Abirached, A.M. Serkine and J. Belaisch-Allart. Explicative factors for prognostics iiu: exploration on 2089 cycles done with statistical and data mining tools. 2004.

[17] R. Lefébure and G. Venturi. *Data Mining. Gestion De La Relation Client.* Eyrolles, France, 2001.

[18] Y. Maday, A.T. Patera and G. Turinici. Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. *Comptes Rendus Mathematique*, **335**(3):289–294, 2002. https://doi.org/10.1016/S1631-073X(02)02466-4.

[19] X.-L. Nguyên, J. Chaskalovic, D. Rakotonanahary and B. Fleury. Insomnia symptoms and CPAP compliance in OSAS patients: A descriptive study using data mining methods. *Sleep Medicine*, **11**(8):777–784, 2010. https://doi.org/10.1016/j.sleep.2010.04.008.

[20] X.-L. Nguyên, D. Rakotonanahary, J. Chaskalovic and B. Fleury. Residual subjective daytime sleepiness under CPAP treatment in initially somnolent apnea patients: a pilot study using data mining methods. *Sleep Medicine*, **9**(5):511–516, 2008. https://doi.org/10.1016/j.sleep.2007.07.016.

[21] J.V. Uspensky. *Introduction to Mathematical Probability.* McGraw-Hill Book Co., New York, 1937.