

# An exact framework for uncertainty quantification in Monte Carlo simulation

P Saracco<sup>1</sup> and M G Pia<sup>2</sup>

National Institute for Nuclear Physics (I.N.F.N.)

Via Dodecaneso, 33

16146 Genova (Italy)

E-mail: <sup>1</sup> [paolo.saracco@ge.infn.it](mailto:paolo.saracco@ge.infn.it), <sup>2</sup> [Maria.Grazia.Pia@cern.ch](mailto:Maria.Grazia.Pia@cern.ch)

**Abstract.** In the context of Monte Carlo (MC) simulation of particle transport Uncertainty Quantification (UQ) addresses the issue of predicting non statistical errors affecting the physical results, i.e. errors deriving mainly from uncertainties in the physics data and/or in the model they embed. In the case of a single uncertainty a simple analytical relation exists among its the Probability Density Function (PDF) and the corresponding PDF for the output of the simulation: this allows a complete statistical analysis of the results of the simulation. We examine the extension of this result to the multi-variate case, when more than one of the physical input parameters are affected by uncertainties: a typical scenario is the prediction of the dependence of the simulation on input cross section tabulations.

## 1. Introduction and problem definition

Uncertainty Quantification (UQ) may refer to a wide variety of specific problems pertaining different scientific areas: related studies involve many methodologies or techniques and even terminologies[1]-[10], so that it is necessary to better define this problem in a more precise way: we are interested in studying how uncertainties of the input data needed for physics simulation transfer into the output. Even within this more restricted domain the problem seems quite complex because the meaning of *input data needed for physics simulation* is rather undetermined: it may refer to those data needed to specify the experimental setup to be simulated - like, e.g., geometrical data or to the composition of the materials - to some conditions externally applied to the system - like temperature, pressure or electromagnetic fields - or to physical data - cross sections or, often, physical models - needed to describe particles transport. All these data are generally affected by uncertainties, which necessarily imply uncertainties in the output of the simulation. So from a responsible MC user's perspective *input data* should refer both to those that are under his direct experimental control and to (many) other data, whose values and uncertainties come from different experiments: it is remarkable that the latter are often not considered. In this contribution we summarize previous results[11, 12] together with some new ones.

To give our analysis a well established mathematical ground we make a fundamental assumption: to be able to disentangle input data uncertainty from the process of simulation itself. In other words we assume that the process of simulation relies on the knowledge of the values of some parameters  $x_1, \dots, x_N$ , with their associate uncertainties, and that these values cannot vary during the simulation (or that these values do not depend on the simulation



itself): this assumption enables to assume *a priori* the knowledge of the Probability Distribution Function (PDF)  $f(x_1, \dots, x_N)$  for the input parameters  $x_1, \dots, x_N$ , that can then be assumed as stochastic parameters.

This is our main hypothesis: its validity in the generic case must be assessed by users case by case.

We just mention two typical cases under which this hypothesis can be not valid or questionable: the first concerns the local temperature of some experimental apparatus we want to simulate; to some extent it can be assumed as fixed by external conditions with their associate uncertainties, external temperature and/or cooling system of the apparatus if it exists, but local significant temperature variations can be present depending on the local rate of energy deposition due to the transported particles. This last quantity is really tracked by the simulation itself, but its consequence in terms of local temperature variation is not normally taken into account automatically by the MC code: the relation between energy deposition and local temperature variations can be taken into account by coupling the MC simulation with some thermo-mechanical code or model, but even if this can be done, not all MC codes are able to use different cross sections data (cross sections data at different temperatures) in the same simulation. The second case concerns the density of some isotopes, which can vary due to activation/decay processes.

Under the quoted assumption the probability that MC simulation of some physical observable  $Y$  have a result between  $y$  and  $y + dy$  has density

$$g_{MC}(y) \simeq \int_{-\infty}^{+\infty} dx_1 \cdots dx_N f(x_1, \dots, x_N) \sqrt{\frac{N_E}{2\pi\sigma_{y_0}^2}} \exp \left[ -\frac{(y - y_0(x_1, \dots, x_N))^2}{2\sigma_{y_0}^2/N_E} \right] \quad (1)$$

for a simulation encompassing  $N_E$  events. Here  $y_0(x_1, \dots, x_N)$  is the sampled mean for the physical observable  $Y$  when the input unknowns assume values  $(x_1, \dots, x_N)$  and  $\sigma_{y_0}^2$  is its sampled variance. This result derives from the Central Limit Theorem (CLT) if  $N_E$  is sufficiently large to make  $\sigma_{y_0}^2$  independent from  $N_E$  itself. In the limit  $N_E \rightarrow \infty$

$$g(y) = \int_{-\infty}^{+\infty} dx_1 \cdots dx_N f(x_1, \dots, x_N) \delta(y - y_0(x_1, \dots, x_N)) . \quad (2)$$

Equation (2) mathematically states how uncertainties in the input data *exactly transfer* into our knowledge of the observable  $Y$  because this is an assignment of a probability for each possible outcome of this observable.

Even if we derived (2) from (1) we could as well make the opposite: (2) represents the forward propagation of uncertainty for a deterministic problem having solution  $y_0$  which, in turn, depends (parametrically, not dynamically) on some unknown input values whose PDF is given by  $f(x_1, \dots, x_N)$ . Searching the solution of such deterministic problem by MC simulation - that is we go back from (2) to (1) - results in a (gaussian) statistical blurring of the exact result  $y = y_0(x_1, \dots, x_N)$  of the underlying deterministic problem. This dual interpretation of (2) and (1) is crucial to our investigation: it is a clear indication of which is the natural goal for any UQ project - the determination of  $g(y)$  - and of its prerequisite - the knowledge of  $f(x_1, \dots, x_N)$ . Unfortunately, but quite obviously, the determination of  $g(y)$  requires also the knowledge of the parametric dependence of the exact solution  $y = y_0(x_1, \dots, x_N)$  on the input uncertainties and on their PDFs: except for very simple examples we are not able to exploit this dependance, but clearly we can use simulation to extract these required informations, as we shall see. It is relevant to stress that our attention has turned to the determination of the exact form of  $g(y)$  rather than the one of  $g_{MC}(y)$ : with this passage we obtain an important result both on the conceptual - as we have seen - and on the practical side. In fact the inherent dependency of

$g_{MC}(y)$  from the details of the simulation<sup>1</sup> makes it impossible to determine it other than by statistical sampling: this implies the necessity of running a very large number of simulations, each time sampling  $(x_1, \dots, x_N)$  out of  $f(x_1, \dots, x_N)$ . On the contrary the determination of the dependence  $y_0 = y_0(x_1, \dots, x_N)$  can be accurately obtained with fewer runs, as we shall discuss.

Uncertainty quantification directly derives from (2): in fact if we know  $g(y)$  the determination of any required statistical information of the physical output, e.g. its confidence intervals, from the unknown input parameters is straightforward. So the task of any UQ project is not the study of how uncertainties of the input data needed for physics simulation transfer into the output (of the simulation), that is the determination of  $g_{MC}(y)$ , rather the use of simulation to determine the properties of  $g(y)$  as better as possible. We note, *en passant*, that techniques we are going to develop to study  $g(y)$  are then applicable to more general contexts than simulation, for instance also when simulations are coupled to deterministic codes.

Moreover we learn that the necessary prerequisite is a precise, as far as possible, knowledge of the PDF for the input data, because it is directly transferred into the physical output probability distribution.

In most of the cases uncertainties in the data which are under the user's control are independent from uncertainties in the physical data needed for transport. So our approach enables to clearly distinguish two very different phases of UQ projects: (i) a *validation phase* and (ii) a *problem specific analysis phase*.

- (i) **The validation phase** - All physical data needed for transport should be validated: all general purpose MC codes make use of cross section tabulations as well as of physical models. All these data should be carefully analyzed giving to users at least the knowledge about their confidence intervals (better, their PDF). It should be clear that this cannot be a users's responsibility: without a previous validation phase any UQ project is meaningless. On the other side how much these data are individually relevant in any given problem cannot be a priori established.
- (ii) **The problem specific analysis phase** - Most of realistic situations under simulation may involve *a priori* hundreds of physical parameters in their full definition: they include parameters whose confidence intervals (or PDF) come from the validation phase, as well as problem specific parameters, like, e.g., geometrical or material composition parameters: for these it is user's responsibility to establish at least proper confidence intervals, better PDF. It will become clear that a full UQ process is out of human possibilities and largely meaningless. So a very detailed analysis of the problem at hand should be carried on to identify those parameters which are likely to be the most relevant in the given experimental configuration. Methods we shall expose in Section 2 can be applied to the selected set of parameters. If necessary successive iterations must be carried on other parameters.

## 2. The path to UQ

Once we have selected a set of parameters on which perform a UQ (or sensitivity) analysis we can proceed on the basis of (2) and (1): in most of the cases two further assumptions hold, namely (i) variations of the input parameters are independent and (ii)  $y_0$  is linear in each of the parameters in their range of variability. We will discuss later these assumptions. If this is true (2) simplifies greatly to

$$g(y) = \int_{-\infty}^{+\infty} dx_1 \cdots dx_N f_1(x_1) \cdots f_N(x_N) \delta \left[ y - \bar{y}_0 - \sum_{k=1}^N a_k(x - \bar{x}_k) \right] \quad (3)$$

<sup>1</sup> From (1) the dependence on the number of generated events  $N_E$  is explicit, but obviously  $g_{MC}(y)$  depends also on the specific way MC simulation is implemented.

where  $f_j(x_j)$  and  $\bar{x}_j$  can be assumed as known from validation phase or from problem specific analysis phase. Problem defined by (3) is a completely defined mathematical problem once we are able to determine  $\bar{y}_0$  and  $a_j = \left. \frac{\partial y_0(x_1, \dots, x_N)}{\partial x_j} \right|_{x_k = \bar{x}_k}$ : it is obvious that we can obtain a statistical estimate of these required values with a predetermined accuracy by means of  $N + 1$  simulations each run with a different set of input parameters, namely  $\bar{x}_1, \dots, \bar{x}_N$  and  $\bar{x}_1, \dots, \bar{x}_j + \Delta, \dots, \bar{x}_N$ , so that indicating with  $y_{MC}(x_1, \dots, x_N)$  the output of the simulation with a given set of values for the input parameters

$$\begin{aligned}\bar{y}_0 &= y_{MC}(\bar{x}_1, \dots, \bar{x}_N) \\ a_j &= [y_{MC}(\bar{x}_1, \dots, \bar{x}_j + \Delta, \dots, \bar{x}_N) - y_{MC}(\bar{x}_1, \dots, \bar{x}_N)] / \Delta.\end{aligned}$$

Obviously these values are affected by statistical errors of the simulation that are of the order of  $\sigma_{y_0}/\sqrt{N_E^2}$ : then a rough estimate of the number of events needed in each of the simulations comes from

$$\sigma_{y_0}/\sqrt{N_E} \ll |y_{MC}(\bar{x}_1, \dots, \bar{x}_j + \Delta, \dots, \bar{x}_N) - y_{MC}(\bar{x}_1, \dots, \bar{x}_N)| \quad (4)$$

It should be realized that the use of (3) entails a large reduction in the computer time required with respect to any attempt to directly determine  $g_{MC}(y)$ , a task that would require thousands of MC runs.

From (3) it is easy to extract  $\langle y \rangle$  and  $\langle y^2 \rangle$ ; if for instance  $\bar{x}_k = \langle x_k \rangle$ , often a convenient choice

$$\begin{aligned}\langle y \rangle &= \bar{y}_0 \\ \langle y^2 \rangle &= \bar{y}_0^2 + \sum_{k=1}^N a_k^2 \langle (x_k - \bar{x}_k)^2 \rangle = \bar{y}_0^2 + \sum_{k=1}^N a_k^2 \sigma_k^2\end{aligned}$$

so that  $\sigma_y^2 = \sum_{k=1}^N a_k^2 \sigma_k^2$ . This quantity can be assumed as a first estimate of the uncertainty, but this is really correct only if  $g(y)$  is - at least approximately - a normal distribution: this is certainly true if the  $f_j$  are normal distributions as well, but not in other cases. So we are left with a classical problem in probability theory, namely the determination of the distribution for the weighted sum of  $N$  independent stochastic variables obeying (a priori) arbitrary distributions.

In some cases the calculation of  $g(y)$  can be carried on exactly. We quoted the case when all the  $f_j$ s are normal distributions: this is a specific example of a general class of distributions which *characteristic functions* - the Fourier transform of the given distributions - are closed under product: these are the so-called **stable distributions**. In probability theory, a random variable is said to be stable (or to have a stable distribution) if it has the property that a linear combination of two independent copies of the variable has the same distribution, up to location and scale parameters. The stable distribution family is also sometimes referred to as the *Lévy* [13]  $\alpha$ -stable distribution. Distributions belonging to this family [14, 15] have characteristic functions of the form

$$\phi(q; \mu, c, \alpha, \beta) = \exp[it\mu - |ct|^\alpha (1 - i\beta \text{sgn}(t)) \Phi] \quad (5)$$

where  $\Phi = \tan(\pi\alpha/2)$  if  $\alpha \neq 1$  and  $\Phi = -2 \log |t|/2$  if  $\alpha = 1$ . This is a 4 parameters ( $-\infty < \mu < \infty, 0 < c < \infty, 0 < \alpha \leq 2$  and  $-1 \leq \beta \leq 1$ ) family which is closed under product for fixed  $\alpha$ : the analytic form of the corresponding PDF is known only for some special

<sup>2</sup> We can naturally assume  $\sigma_{y_0}$  as constant for small variations  $\Delta$  of the parameters.

values of the parameters. Among these we mention the normal distribution ( $\alpha = 2$ ,  $c = \sigma/\sqrt{2}$ ), the Cauchy distribution ( $\alpha = 1$ ,  $\beta = 0$ ), and the Lévy distribution ( $\alpha = 1/2$ ,  $\beta = 1$ ). All these distributions apart the normal one are *heavy tailed*, that is their behavior for large  $x$  is  $|x|^{-1-\alpha}$ . In the limit  $\alpha \rightarrow 0$  or  $c \rightarrow 0$  they approach a Dirac delta.

The most common case of applicability of these properties to UQ is when we have to analyze a certain number of input parameters affected by different *statistical errors*.

Another useful case in practice is when we have a certain number of input parameters with flat distributions: this is the case when input parameters are given in the form  $\bar{x}_j \pm \Delta x_j/2$ . In this case  $g(x)$  is given by a generalization of the *Irwin-Hall distribution* [16, 17] recently revisited [18].

Unfortunately all these results apply only to cases when all the input unknowns have the same distributions with different parameters: for instance they can be all normal distributions, with different means and variances, or all flat distributions, with different means and experimental errors, and so on. It should be clear that in a generic UQ problem this is a strong limitation, because we can have the case of input parameters obeying different distributions as well: this problem is clearly not soluble in full generality so we must work on some approximation scheme to the search for  $g(y)$ .

### 3. A more general result

We recently succeeded in finding an exact analytical expression for the weighted sum of  $N$  independent stochastic variables obeying arbitrary polynomial distributions supported on bounded intervals, whose proof can be found in [19]. Here we simply quote our central result and we show how this can be useful in determining  $g(y)$  with some predetermined error: an arbitrary distribution  $f(x)$  supported on a bounded interval  $a \leq x \leq b$  can always be approximated with a predetermined accuracy by a sequence of polynomials each defined on different sub-intervals of  $[a, b]$ : the most simple case of such procedure is to subdivide the interval  $[a, b]$  in  $k$  sub-intervals  $x_0 = a, x_1, \dots, x_k = b$  and to approximate the given distribution with sequence of segments joining the values  $f(x_0), \dots, f(x_k)$ . More accurate results can be obtained with a lower number of subintervals by using splines. Accuracy of an approximation is defined with respect to some given norm, measuring the *distance* between the exact result and its approximation; in our case, to the purpose of studying the propagation of error, it is convenient to make the choice

$$||f|| = \sup_{[a, b]} |f(x)| \quad (6)$$

so that the distance between two function is the maximum of their absolute differences. We assert without proof that if  $f$  is continuous (or if it has at most a finite number of discontinuities) it is always possible to find a subdivision of  $[a, b]$  such that the distance between  $f$  and its polynomial approximation  $f_{\text{app}}$  is bounded by a predefined  $\varepsilon > 0$ <sup>3</sup>. Then if we have  $N$  such distributions, each approximated by this procedure, the maximum error of the distribution of the sum  $x_1 + \dots + x_N$  is clearly  $N\varepsilon$ : for a weighted sum  $a_1x_1 + \dots + a_Nx_N$  the bound on the error turns out to be  $\varepsilon \sum_j |a_j|$ . This means that with this procedure we can obtain the

required distribution  $g(y)$  with any predetermined error provided we are able to make exactly the convolution<sup>4</sup> of generic polynomial forms defined over different intervals.

Thanks to the linearity of the convolution this is equivalent to the ability of performing the convolution of  $N$  monomial forms with arbitrary exponents supported on different bounded

<sup>3</sup> This is a consequence of Weierstrass approximation theorem.

<sup>4</sup> We remind the reader that the (weighted) sum of  $N$  stochastic variables is expressed by the convolution of their (rescaled) distributions.

intervals: this is exactly what we proved in [19]. The convolution of  $N$  monomials  $x_1^{p_1}, \dots, x_N^{p_N}$  that are different from zero only over the intervals  $-c_1 \leq x_1 \leq c_1, \dots, -c_N \leq x_N \leq c_N$  is given by

$$I(x; \vec{c}; \vec{p}; N) = \frac{\prod_{k=1}^N (\max(p_k, 1))!}{2 \left( \sum_{k=1}^N (p_k + 1) - 1 \right)!} \prod_{k=1}^N \hat{Q}(p_k, \lambda_k) \quad (7)$$

$$\times \sum_{m \in \mathbb{M}_N(x, \vec{c}, \vec{\lambda})} (-1)^{C(m)} m_{\sum_{k=1}^N (p_k + 1) - 1} \text{sgn}(m|_{\vec{\lambda}=1}) \Big|_{\vec{\lambda}=1}$$

where  $\hat{O}(q, \lambda) = \sum_{s=0}^q \frac{(-1)^s}{s!} \frac{\partial^s}{\partial \lambda^s}$  and  $\mathbb{M}_N(x, \vec{c}, \vec{\lambda}) = \{m(x) = x \pm c_1 \lambda_1 \pm c_2 \lambda_2 \dots \pm c_N \lambda_N\}$ .

Equation (7) appears as a generalization of the result [18]. We remark that calculation of  $g(y)$  along the quoted guidelines can be tedious and often long, so we are planning to release a specific dedicated software to automate the process.

Anyway the outlined procedure **is always able to give a definite answer on the expected PDF**  $g(y)$  with a predefined accuracy: it is then possible to extract confidence intervals for the required physical quantity, or any other statistical information we can need.

Two sources of errors must be accurately followed up: the first concerns the (statistical) errors in the values required for the calculation ( $\bar{y}_0$  and the  $a_j = \left. \frac{\partial y_0}{\partial x_j} \right|_{x_k = \bar{x}_k}$ ) that, as we told, can be extracted from accurate simulations; the second concerns propagation of errors in the procedure of approximating individual input PDFs with polynomial forms over finite intervals.

In the most generic case some of the input PDF can be defined over infinite intervals - for instance they may be normal distributions: if this is the case we can follow the described procedure provided we *a priori* define some confidence for the whole calculation; that is to say that a normal distribution can be approximated by a distribution over a finite interval with a given confidence. So the worst case, for which we cannot easily apply the described procedure, is the case when some of the input PDF are heavy tailed.

#### 4. Conclusions and outlook

Results presented in the previous section provide a well defined mathematical framework to obtain the probability density function for any observable  $Y$  depending on some given set of input physical parameters and on their associate uncertainties.

We made three assumptions about which some comment is useful.

First we assumed that the values of the input parameters, and implicitly their probability distributions, do not vary along the simulation: on the basis of the discussion we made about equations (1) and (2) this condition should be more properly reformulated as: the input parameters and their probability distributions do not depend on the dynamical evolution of the system. It should be clear that this condition can, at least in principle, be given up: this possibility depends on our ability to build a more complete dynamical description of the system under examination including also the possible evolution of these dynamically variable parameters: to remain within the examples we made, local temperatures can be derived by some thermomechanical or thermohydraulic model which must be coupled to particle transport, or isotope concentrations can be obtained by coupling transport to some solver for Bateman's

equations; even if this is in principle possible, in practice it depends on a major reformulation of the existing simulation codes.

The second major assumption is about the statistical independence of different input uncertainties: we currently do not see any way-out to this assumption; however this is the most common case in practice.

The third assumption is about the linearity of the response  $y_0(x_1, \dots, x_N)$  with respect to each of the  $x_k$ , see equation (3): it is clearly possible to give up this assumption by properly subdividing the region of integration in (2), as we noted earlier, for example using methods from [20], but this process introduce additional approximation errors whose propagation must be studied case by case for their effect on the precision on the knowledge of  $g(y)$ .

So we can conclude that the presented conceptual mathematical framework for UQ is really consistent and it relies in principle only on the assumption of mutual independence of the input unknowns: however its practical usability is not straightforward in its general formulation because of the error tracking process. Then we plan the release of a dedicated software tool to automate this task.

## References

- [1] Walker W E et al. 2003 Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support *Integrated Assessments* **4** 5-17
- [2] Marino S, Hogue I B, Ray C J and Kirschner D E 2008 A methodology for performing global uncertainty and sensitivity analysis in systems biology *J. Theor. Biol.* **7** 178-196
- [3] Levine R A and Berliner L M 1999 Statistical principles for climate change studies *J. Climate* **12** 564-574
- [4] Suslick S B and Schiozer D J 2004 Risk analysis applied to petroleum exploration and production: an overview *J. Petr. Sc. Eng.* **44** 1-9
- [5] Bernardini A and Tonon F 2010 *Bounding Uncertainties in Civil Engineering*, (Berlin: Springer and Verlag)
- [6] Oberkampf W L, DeLand S M, Rutherford B M, Diegert K V and Alvin K F, Error and uncertainty in modeling and simulation 2002 *Reliab. Eng. Syst. Safety* **75** 333-357
- [7] Swiler L P, Paez T L and Mayes R L 2009 Epistemic Uncertainty Quantification Tutorial *Proc. IMAC-XXVII, Conf. and Exposition on Structural Dynamics* (Orlando: Soc. Structural Mech) paper 294
- [8] Helton J C 2011 Quantification of margins and uncertainties: conceptual and computational basis *Reliab. Eng. Syst. Safety* **96** 976-1013
- [9] Lin G, Engel D W and Eslinger P W 2012 Survey and evaluate uncertainty quantification methodologies Pacific Northwest National Laboratory 20914 Report, (Richland: PNNL)
- [10] Oberkampf W L and Roy C J 2010 *Verification and Validation in Scientific Computing* Chapter 13 (Cambridge: CUP)
- [11] Saracco P, Batic M, Hoff G and Pia M G 2012 "Uncertainty Quantification (UQ) in generic MonteCarlo simulations" Nuclear Science Symposium Conference Record (NSS/MIC) IEEE
- [12] Saracco P, Batic M, Hoff G and Pia M G 2013 "Theoretical ground for the propagation of uncertainties in Monte Carlo particle transport" submitted to *IEEE Trans. Nucl. Science*
- [13] Lévy P. 1925 *Calcul de Probabilités* (Paris: Gauthier Villars)
- [14] Zolotarev V M, 1986 *One-dimensional Stable Distributions* (Providence: American Mathematical Society)
- [15] Fofack H and Nolan J P 1999 "Tail behavior, modes and other characteristics of stable distributions" *Extremes* 39-58
- [16] Irwin J O 1927 "On the frequency distributions of the means of samples from a population having any law of frequency, with special reference to Pearson's type 2" *Biometrika* **19** 225-239
- [17] Hall P 1927 "The distribution of means for samples of size N drawn from a population in which the variate takes value between 0 and 1, all such values being equally probable" *Biometrika* **19** 240-245
- [18] Bradley D M and Gupta RC 2002 "On the distribution of the sum of N non identically distributed random variables" *Ann. Inst. Statist. Math.* **54** 689-700
- [19] Saracco P and Pia M G 2013 "Parameter Uncertainty Quantification and the problem of determining the distribution of the sum of N independent stochastic variables: an exact solution for arbitrary polynomial distributions on different intervals" submitted to *Journ. Math. Phys.*
- [20] Luck R and Stevens J W 2004 "A simple numerical procedure for estimating nonlinear uncertainty propagation" *ISA Trans.* **43** 491-497