



CONTRIBUTION TO THE MATHEMATICAL AND NUMERICAL ANALYSIS OF UNCERTAIN SYSTEMS OF CONSERVATION LAWS AND OF THE LINEAR AND NONLINEAR BOLTZMANN EQUATION

Gaël Poëtte

► To cite this version:

Gaël Poëtte. CONTRIBUTION TO THE MATHEMATICAL AND NUMERICAL ANALYSIS OF UNCERTAIN SYSTEMS OF CONSERVATION LAWS AND OF THE LINEAR AND NONLINEAR BOLTZMANN EQUATION. Numerical Analysis [cs.NA]. Université de Bordeaux 1, 2019. tel-02288678

HAL Id: tel-02288678

<https://hal.science/tel-02288678>

Submitted on 15 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**HABILITATION À DIRIGER DES RECHERCHES
UNIVERSITÉ BORDEAUX**
**ÉCOLE DOCTORALE DE
MATHÉMATIQUES ET INFORMATIQUE**

Spécialité : MATHÉMATIQUES APPLIQUÉES

Présentée par
Gaël Poëtte
Intitulée

**CONTRIBUTION À L'ANALYSE MATHÉMATIQUE ET
NUMÉRIQUE
DES SYSTÈMES DE LOIS DE CONSERVATION
INCERTAINS
ET DE L'ÉQUATION DE BOLTZMANN LINÉAIRE ET
NON-LINÉAIRE**

Laboratoire de rattachement
Institut Mathématiques de Bordeaux
de l'Université de Bordeaux

Organisme d'accueil
CEA CESTA DAM
F-33114 Le Barp, France

Soutenue le 10/09/2019 devant le jury composé de:

Paola CINNELLA	Président
Martin FRANK	Rapporteur
Andrea ZOIA	Rapporteur
Rodolphe TURPAULT	Rapporteur
Bruno DESPRÉS	Examinateur
Pietro CONGEDO	Examinateur
Daniel VANDERHAEGEN	Examinateur

**CONTRIBUTION À L'ANALYSE MATHÉMATIQUE ET
NUMÉRIQUE
DES SYSTÈMES DE LOIS DE CONSERVATION
INCERTAINS
ET DE L'ÉQUATION DE BOLTZMANN LINÉAIRE ET
NON-LINÉAIRE**

Gaël Poëtte

**CONTRIBUTION TO THE MATHEMATICAL AND
NUMERICAL ANALYSIS
OF UNCERTAIN SYSTEMS OF CONSERVATION LAWS
AND OF THE LINEAR AND NONLINEAR BOLTZMANN
EQUATION**

Gaël Poëtte

*Maybe skateboarding is a crime afterall...
(#shlagboarding is definitely not!)*

Remerciements

Je tiens tout d'abord à remercier mon jury: merci à Paola d'avoir accepter de présider celui-ci, merci aux rapporteurs Rodolphe, Martin et Andrea d'avoir eu le courage de relire ce manuscrit (pendant vos vacances d'été!). Merci à Daniel qui, même s'il n'était pas rapporteur, a également fait cet effort (et merci pour le bouquin!). Merci à Bruno et Pietro d'avoir accepté d'être examinateurs mais surtout pour les collaborations passées et futures (Bruno, je te paies une bière à la première occasion!).

Merci à Laetitia pour ta présence et son soutien. Tu m'auras supporté le temps d'une thèse et d'une HDR (et avant et entre les deux même)... Je pense que tu es prête pour la phase suivante (kilos en plus, bière à la main, 8.5 sous les pieds, en peignoir mauve dans une mini-rampe de ma confection). Merci à Donatella et Dante pour n'avoir pas manquer une occasion de me faire lever le nez du mémoire pendant la rédaction...

Merci à ma madre et à mon frangin. Madre, je suis pressé que tu nous rejoignes. On s'éclate bien ici avec le frangin et les kids. On se fait des sessions de skate de ouf et on discute méthodes numériques.

Merci à Parrain, Tati et madre encore... Promis, c'est la dernière fois que je vous traîne à une soutenance relou avec que des symboles chelou et qui vous fait pleurer en plus.

Merci aux gens de mon ancien labo Hervé, Xavier, Adrien, David, Philippe, Frédéric, Stéphane, Mohamed, Christophe (encore merci pour le Babuel-Perissac!), Cédric, Gilles (d'une manière ou d'une autre, vous avez fait partie de l'aventure HDR, de par nos discussions et nos études), ainsi que les copains de l'ancien service (je ne vais pas tous vous citer mais mention particulière à Benoît, Marie-Pierre et Thierry parce que sans vous trois et vos conseils de gainage et de gym, je crois que j'aurais encore mal au dos). Merci aux gens de mon nouveau labo Isabelle, Alexia, David, Nicolas, Pierre, Frédéric, Marc, Ludovic (ceux qui étaient à ma soutenance se sont sans doute rendu compte qu'ils m'ont plus aidé qu'ils ne l'auraient pensé!) mais également de mon nouveau service (encore une fois, je ne vais pas tous vous citer mais le cœur y est). Pardon à Paul que je vais pas mal torturer pendant ses prochaines années.

Merci aux fous, Maxime, Arnaud, Julie, Jennifer, Mathieu, Sylvine, Sophie & Sophie, David. C'était quand même pas mal nos soirées dans les bars... Maintenant on est tous à gauche à droite dans des contrées plus ou moins bizarres pour certains à nous occuper de nos progénitures...

Merci aux copains du skatepark et de la rue, de Metz et d'ailleurs, aux shlags et aux dinlows (#IrideForMat, #IrideForRobin), à François (#shlagboarding), au frangin (encore toi, t'es partout!), à Étienne (nan! Xanti!) et à Jordan (la meilleure surprise sur un skatepark girondin!) et à tous ceux que j'ai croisé avant, Valentin, Vincent, Guillaume. Parce skate+potes+binouzes=♡ reste l'équation la plus simple à résoudre.

Bon j'arrête ici, sinon je ne mettrai jamais le document en ligne. Merci à Tati Do.

Contents

I General Introduction	2
1 Introduction	3
1.1 The (quadratic) Boltzmann equation and two of its limits	4
1.1.1 One Hydrodynamic limit of Boltzmann equation	6
1.1.2 The Linear Boltzmann equation limit	9
1.1.3 Deterministic resolution schemes to solve (1.22) and (1.30)	11
1.1.4 Stochastic resolution schemes to solve (1.22) and (1.30)	12
1.2 V&V: the role of numerical analysis, UQ and HPC	13
1.2.1 Verification & Validation (V&V)	13
1.2.2 Numerical analysis: the main tool for verification	14
1.2.3 Uncertainty Quantification (UQ): the main tool for validation	14
1.2.4 Numerical/uncertainty analysis as tools for V&V and the role of HPC	15
1.3 Few words on the content and style of this document	19
1.3.1 Content and	19
1.3.2 ... Style	20
1.3.3 Few words on the notations and the presentation tricks	20
II Uncertainty quantification for hyperbolic systems of conservation laws	23
2 Physical Motivations and toy problem (fil rouge)	24
3 Polynomial Chaos as an alternative to Monte-Carlo methods for UQ	30
3.1 Wiener's Homogeneous Chaos [295] and Cameron-Martin's theorem [55]	31
3.1.1 On Stone-Weierstrass' approximation theorem	32
3.1.2 On Wiener's Homogeneous Chaos [295]	32
3.1.3 On Cameron and Martin's theorem [55]	34
3.2 Polynomial Chaos for uncertainty quantification (UQ)	36
3.2.1 Transformation of a gaussian random variable into a uniform one	37
3.2.2 Mapping of a uniform random variable into an Arcsinus and a Binomial one	38
3.3 Introduction of generalized Polynomial Chaos (gPC) for UQ	40
3.4 The construction of the gPC basis	42
3.4.1 Inner product defined by an arbitrary probability measure	43
3.4.2 Moments of a probability measure and Hankel determinants	43
3.4.3 Christoffel's formulae, Jacobi's matrix and construction procedures	45
3.4.4 Taking into account discrete/categorical input variables with gPC	47
3.5 Curse of dimensionality and Gibbs phenomenon	47
3.5.1 Curse of dimensionality	47
3.5.2 Sensitivity to the Gibbs phenomenon	48
3.6 Summary for generalized Polynomial Chaos	49

4 Intrusive application of gPC for systems of conservation laws	51
4.1 Intrusive application of gPC	52
4.1.1 The P -truncated gPC reduced model: a P_n -like closure	52
4.1.2 Roe solver for the P -truncated intrusive gPC reduced model	55
4.1.3 Application to the 'fil rouge' problem of chapter 2	56
4.2 A step-by-step study of intrusive gPC for systems of conservation laws	57
4.2.1 The particular case of a scalar conservation law	57
4.2.2 Possible loss of wellposedness for non-scalar systems of conservation laws	60
4.2.3 A closure ensuring wellposedness for general systems of conservation laws	62
4.2.4 Application to the 'fil rouge' problem of chapter 2	67
4.3 Summary for intrusive gPC and the entropy closure reduced models	71
5 Non-Intrusive application of gPC for systems of conservation laws	73
5.1 Non-intrusive application of gPC	74
5.2 Choice of the experimental design (the most common ones for UQ)	75
5.2.1 The Monte-Carlo (MC) integration method	76
5.2.2 Low discrepancy sequences/Quasi Monte-Carlo	77
5.2.3 Gauss quadrature rules	78
5.2.4 Clenshaw-Curtis (CC) quadrature rule	82
5.2.5 MC vs. Quasi MC vs. Gauss vs. etc.	83
5.3 Integration vs. Regression vs. Collocation vs. Kriging	84
5.3.1 Regression-gPC approximations	84
5.3.2 Collocation-gPC approximation	90
5.3.3 Kriging-gPC approximations	92
5.4 Few other applications of gPC	98
5.4.1 Application to the 'fil rouge' problem of chapter 2	98
5.4.2 Integration vs. Regression vs. Collocation vs. Kriging vs. discontinuity	99
5.5 Summary for non-intrusive gPC for systems of conservation laws	102
6 The non-intrusive iterative gPC (i-gPC) approach	104
6.1 The main idea behind iterative-gPC (i-gPC)	104
6.1.1 A particular change of variable $Z(X)$ ensuring a gain	107
6.1.2 Description of the i-gPC approximation algorithm	107
6.1.3 Weak contraction of the i-gPC approximation	108
6.2 Application of i-gPC on two simple test-problems	109
6.2.1 Discontinuous output random variable	109
6.2.2 Smooth output random variable	110
6.3 Numerical analysis of i-gPC in finite integration context (stopping criterion)	112
6.3.1 Convergence behaviour of i-gPC under finite numerical integration accuracy	114
6.3.2 Strategy for adaptive approximation truncation	115
6.4 Few other applications of i-gPC	118
6.4.1 Application to the 'fil rouge' configuration	118
6.4.2 Integration vs. Regression vs. Collocation vs. Kriging vs. i-gPC	119
6.5 Summary for non-intrusive gPC and i-gPC approximations	120
7 Non intrusive gPC for Direct Numerical Simulation (DNS) acceleration	123
7.1 Perturbation reduced models as a limit of the gPC one	124
7.1.1 Perturbation reduced model of a system of conservation laws	124
7.1.2 gPC reduced model of a system of conservation laws	126
7.1.3 The perturbative reduced model as a limit of the gPC one	126
7.2 Direct Numerical Simulation (DNS) acceleration via gPC	129
7.2.1 The shock tube experiments and their initial conditions	130
7.2.2 Stochastic dimension reduction for the initial Uncertain Interface Position	132
7.2.3 The Multimaterial 2D Euler system	133
7.2.4 Observable of interest, Simulations and Comparisons with Experimental Results	134
7.3 Conclusion for the gPC application to chaotic flows	138

8 Toward an application of Cameron-Martin's theorem (not only its special case)	139
8.1 An attempt to apply theorem 3.3: an i-gPC decomposition of the residue	140
8.1.1 Analysis of theorem 3.3 and comparison to theorem 3.4	140
8.1.2 i-gPC decomposition of the residue in an infinite integration accuracy context	141
8.1.3 i-gPC decomposition of the residue in a finite integration accuracy context	143
8.2 Numerical Applications of the i-gPC decomposition of residue method	145
8.2.1 Some (hydrodynamically motivated) 1D test-problems	145
8.2.2 Some (well-known in the literature) multidimensional test-cases	156
8.3 Summary for the i-gPC decomposition of the residue algorithm	160
III Monte-Carlo schemes for the (non)linear Boltzmann equation	161
9 Monte-Carlo methods for the linear Boltzmann equation	162
9.1 General Methodology for the construction of an MC scheme	165
9.2 The analog (Adjoint) MC scheme (mimics physics)	166
9.2.1 Expectation form over the analog set of random variables	166
9.2.2 Construction of the analog MC scheme	168
9.3 The semi-analog (Adjoint) MC scheme (implicit capture)	171
9.3.1 Expectation form over the semi-analog set of random variables	171
9.3.2 Construction of the semi-analog MC scheme	172
9.4 The non-analog (Adjoint) MC scheme	174
9.4.1 Expectation form over the non-analog set of random variables	174
9.4.2 Construction of the Adjoint non-analog MC scheme	175
9.5 Direct formulation and direct set of random variables	177
9.5.1 Adjoint and direct formulations of the same transport equation	177
9.5.2 Direct Integral formulation for the non-analog scheme	178
9.5.3 Construction of the direct non-analog MC scheme	178
9.6 Common approximations to simplify the samplings and resolutions	181
9.6.1 The interaction time τ of probability measure $f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds$	181
9.6.2 The energy and angle correlated samplings $\mathbf{V}' = V'W'$	185
9.6.3 The modification of the weight of the particle $w_p(t)$	187
9.7 Variance and moments of the MC schemes	188
9.7.1 Asymptotic variance of the analog scheme (full_analog and multiplicity)	188
9.7.2 Asymptotic variance of the semi-analog scheme	191
9.7.3 Asymptotic variance of the non-analog scheme	192
9.7.4 Comparisons of the standard deviations of the MC schemes (homogeneous)	193
9.8 A general canvas for developing MC schemes	196
9.8.1 Sampling the initial MC particle population	196
9.8.2 A general skeleton in order to develop each scheme in the same platform	204
9.9 Taking into account a source term	207
9.9.1 Application of Duhammel's principle: source sampling (direct)	208
9.9.2 Quasi-Static method for the transport equation with source term	210
9.10 Taking into account an acceleration term in MC resolution schemes	213
9.10.1 An MC resolution with curved trajectories in the comobile frame	214
9.10.2 An MC resolution with straight trajectories in a new frame	215
9.11 The Uncertain Linear Boltzmann equation	222
9.11.1 Non-intrusive resolution of the uncertain linear Boltzmann equation	223
9.11.2 A gPC-intrusive Monte-Carlo scheme for the uncertain linear Boltzmann equation	225
9.11.3 Summary	230
9.12 Application of gPC for MC accelerations for the linear Boltzmann equation	231
9.12.1 Variance reduction, AP scheme, same problems, different denominations	234
9.12.2 Application of gPC to accelerate MC integration	234
9.12.3 Acceleration by gPC of the MC resolution of the linear Boltzmann equation	247
9.12.4 Summary	253

10 Monte-Carlo methods for the nonlinear Boltzmann equation	254
10.1 Boltzmann equation coupled to Bateman system (neutronics)	255
10.1.1 Classical MC schemes for neutron transport	256
10.1.2 An Asymptotic Preserving MC scheme for neutron transport	261
10.1.3 Summary	266
10.2 Boltzmann equation coupled to Stefan's law (photomics)	267
10.2.1 Classical Monte-Carlo schemes for photon transport	270
10.2.2 Two Asymptotic Preserving MC schemes for photon transport	275
10.2.3 Summary	284
IV Conclusion	287
11 Conclusion	288
11.1 On Uncertainty Quantification (part II)	288
11.2 On Monte-Carlo resolution schemes (part III)	289
V Appendix	291
A Analytical resolution of the uncertain Burgers' equation	292
B Statistical hypothesis testing in a nutshell	301
B.1 A (too brief and general) presentation of statistical hypothesis testing	301
B.2 Statistical hypothesis testing and uncertainty propagation for V&V	302

Part I

General Introduction

Chapter 1

Introduction

Contents

1.1	The (quadratic) Boltzmann equation and two of its limits	4
1.1.1	One Hydrodynamic limit of Boltzmann equation	6
1.1.2	The Linear Boltzmann equation limit	9
1.1.3	Deterministic resolution schemes to solve (1.22) and (1.30)	11
1.1.4	Stochastic resolution schemes to solve (1.22) and (1.30)	12
1.2	V&V: the role of numerical analysis, UQ and HPC	13
1.2.1	Verification & Validation (V&V)	13
1.2.2	Numerical analysis: the main tool for verification	14
1.2.3	Uncertainty Quantification (UQ): the main tool for validation	14
1.2.4	Numerical/uncertainty analysis as tools for V&V and the role of HPC	15
1.3	Few words on the content and style of this document	19
1.3.1	Content and	19
1.3.2	... Style	20
1.3.3	Few words on the notations and the presentation tricks	20

This document presents my research contributions to two fields of application: uncertainty quantification for systems of conservation laws (part II), and the numerical (Monte-Carlo) resolution of the (linear and nonlinear) deterministic Boltzmann equation (part III). Basically, in part II, we present some strategies to solve stochastic partial differential equations (PDEs) with deterministic methods whereas in part III we solve deterministic PDEs with stochastic resolution schemes. At first glance, the two topics may appear different if not orthogonal. In this introductory part, we explain in which sense both subjects are parts of the same research topic. We suggest two ways to emphasize their common points:

- in section 1.1, we insist on the fact that two models are mainly studied in this document, systems of conservation laws and the (linear and nonlinear) Boltzmann equation. We first recall they are two limits of the same more general quadratic Boltzmann equation. There exists several ways to derive those limits and they are of importance in this document. For this reason, we take few pages to briefly introduce them to obtain the two main models studied in this manuscript. We also briefly go through the common resolution schemes for those two limits and put forward analogies: some well-known numerical methods applied in one field of application can be very useful in the other once some similarities noticed. It opens to new ideas, new models and new numerical methods inspired from one field to the other.
- On another hand, in section 1.2, instead of focusing on the models we are solving in parts II and III, we focus on the purposes we aim at achieving with these models in a Verification & Validation (V&V) context. V&V provides the basic bricks for someone willing to compare efficiently experimental and numerical results. We will see that in part II, we aim at quantifying probabilistically

the fluctuation/discrepancy between experimental observations and numerical results (validation). On another hand, in part III, we focus on reducing the numerical error of Monte-Carlo simulations (verification).

In brief, in this document, we deal with two models, two different goals with respect to these two models because of two different needs with respect to V&V. The common objective remains we aim at providing better simulation codes, improving their capabilities and ensuring better physical interpretations.

1.1 The (quadratic) Boltzmann equation and two of its limits

In this section, we do not aim at being exhaustive on the quadratic Boltzmann equation, its conditions for relevance [18] nor the resolution strategies for systems of conservation laws or the linear Boltzmann equation. The material is mainly taken from [207, 248, 128, 60, 18] and rearranged so that it helps putting forward that the research contributions presented in this document are linked, intertwined and first steps of an ongoing scientific project.

The (quadratic) Boltzmann equation models the properties of dilute gases by *statistically* analysing the elementary collision processes between pairs of molecules. By statistically, we mean it does not aim at characterising the positions and velocities of each particles of the gases but the probability distribution $f(\mathbf{x}, t, \mathbf{v}) \geq 0$ of having particles at position $\mathbf{x} = (x_1, x_2, x_3)^t \in \mathcal{D} \subset \mathbb{R}^3$, velocity $\mathbf{v} = (v_1, v_2, v_3)^t \in \mathbb{R}^3$ and time $t \in [0, T] \subset \mathbb{R}^+$. The unknown consequently depends on $3(\mathbf{x}) + 1(t) + 3(\mathbf{v}) = 7$ independent variables. The probability distribution f is in a 7-dimensional space and is vectorial for a mixture of M different species of non-reacting mono-atomic particles of masses $(m_i)_{i \in \{1, \dots, M\}}$

$$f(\mathbf{x}, t, \mathbf{v}) = (f_1(\mathbf{x}, t, \mathbf{v}), \dots, f_M(\mathbf{x}, t, \mathbf{v}))^t.$$

Each $(f_i)_{i \in \{1, \dots, M\}} \geq 0$ denotes the probability distribution for species i and satisfies a coupled nonlinear integro-differential equation $\forall i \in \{1, \dots, M\}$

$$\partial_t f_i(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_i(\mathbf{x}, t, \mathbf{v}) + F_i(\mathbf{x}, t, \mathbf{v}) \partial_{\mathbf{v}} f_i(\mathbf{x}, t, \mathbf{v}) = Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}). \quad (1.1)$$

In the above equation, F_i denotes the force applied to particle i and the coupling between species is made via $(Q_i)_{i \in \{1, \dots, M\}}$, the collision kernels. These kernels are given by $\forall i \in \{1, \dots, M\}$

$$Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}) = \sum_{j=1}^M Q_{i,j}(f_i, f_j)(\mathbf{x}, t, \mathbf{v}), \quad (1.2)$$

with

$$Q_{i,j}(f_i, f_j)(\mathbf{x}, t, \mathbf{v}) = \int |\mathbf{v} - \mathbf{v}_j| \sigma_{i,j}(\mathbf{v} - \mathbf{v}_j) (f_i(\mathbf{x}, t, \mathbf{v}'(\mathbf{v}, \mathbf{v}_j)) f_j(\mathbf{x}, t, \mathbf{v}'_j(\mathbf{v}, \mathbf{v}_j)) - f_i(\mathbf{x}, t, \mathbf{v}) f_j(\mathbf{x}, t, \mathbf{v}_j)) d\mathbf{v}_j. \quad (1.3)$$

They are built considering interactions resulting solely from two particles (hence the *quadratic* denomination) that are assumed to be uncorrelated prior to the collision (molecular chaos, see [18, 251, 159]). In (1.3), the term $\sigma_{i,j}$ denotes the scattering differential cross-sections between species i and j . It describes the probability of binary collisions between i and j together with the probability for a certain change of velocities, from \mathbf{v} and \mathbf{v}_j to \mathbf{v}' and \mathbf{v}'_j . The expression of the scattering kernels with respect to $\mathbf{v}', \mathbf{v}'_j$ may be too general here (one can for example explicit the relations $\mathbf{v}'_j(\mathbf{v}, \mathbf{v}_j)$ and $\mathbf{v}'(\mathbf{v}, \mathbf{v}_j)$ based on kinematic considerations, see [248, 128, 60] and [32]¹), but it is not central to illustrate our purpose. Let

¹The notations of this document are very close to the one of this paper regarding collision kernels.

us rewrite system (1.1) in a more concise form by introducing

$$\begin{aligned} f(\mathbf{x}, t, \mathbf{v}, \lambda) &= \sum_{i=1}^M f_i(\mathbf{x}, t, \mathbf{v}) \delta_i(\lambda), \\ F(\mathbf{x}, t, \mathbf{v}, \lambda) &= \sum_{i=1}^M F_i(\mathbf{x}, t, \mathbf{v}) \delta_i(\lambda), \\ Q(f, f)(\mathbf{x}, t, \mathbf{v}, \lambda) &= \sum_{i=1}^M Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}) \delta_i(\lambda). \end{aligned}$$

Hence, equation (1.1) can be rewritten as a scalar concise equation

$$\partial_t f(\mathbf{x}, t, \mathbf{v}, \lambda) + \mathbf{v} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{v}, \lambda) + F(\mathbf{x}, t, \mathbf{v}, \lambda) \partial_{\mathbf{v}} f(\mathbf{x}, t, \mathbf{v}, \lambda) = Q(f, f)(\mathbf{x}, t, \mathbf{v}, \lambda). \quad (1.4)$$

Of course, integrate with respect to λ in the above expression and we recover (1.1). Expression (1.4), equivalent to (1.1) but expressed with respect to an additional parameter λ , may appear unconventional but prepares some discussions for the construction of MC scheme² in part III. The global collision kernel $Q(f, f)$ satisfies conservation properties for

$$\begin{aligned} \text{mass: } & \iint Q(f, f)(\mathbf{x}, t, \mathbf{v}) m(\lambda) d\mathbf{v} d\lambda = \sum_{i=1}^M \int Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}) m_i d\mathbf{v} = 0, \\ \text{momentum: } & \iint Q(f, f)(\mathbf{x}, t, \mathbf{v}) m(\lambda) \mathbf{v} d\mathbf{v} d\lambda = \sum_{i=1}^M \int Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}) m_i \mathbf{v} d\mathbf{v} = 0, \quad (1.5) \\ \text{energy: } & \iint Q(f, f)(\mathbf{x}, t, \mathbf{v}) m(\lambda) \frac{1}{2} |\mathbf{v}|^2 d\mathbf{v} d\lambda = \sum_{i=1}^M \int Q_i(f_i, f)(\mathbf{x}, t, \mathbf{v}) m_i \frac{1}{2} |\mathbf{v}|^2 d\mathbf{v} = 0. \end{aligned}$$

Note that in the above expression, we have $m(\lambda) = \sum_{i=1}^M m_i \delta_i(\lambda)$. The kernel also satisfies (a multi-species version, see [44, 89]) of the H-theorem, i.e. we have³

$$\partial_t \int f(\mathbf{x}, t, \mathbf{v}) \ln(f(\mathbf{x}, t, \mathbf{v})) d\mathbf{v} + \partial_{\mathbf{x}} \int \mathbf{v} f(\mathbf{x}, t, \mathbf{v}) \ln(f(\mathbf{x}, t, \mathbf{v})) d\mathbf{v} = \int Q(f, f)(\mathbf{x}, t, \mathbf{v}) \ln(f(\mathbf{x}, t, \mathbf{v})) d\mathbf{v} \leq 0.$$

It implies that any local equilibrium is the minimum of the above Boltzmann entropy and has the form of a *local* Maxwellian measure [18]: recall \mathbf{v} is three dimensional, i.e. $\mathbf{v} = (v_1, v_2, v_3)^t$, so that $d\mathbf{v} = dv_1 dv_2 dv_3$. Then the Maxwellian measure for particle of type i is given by

$$\mathcal{M}_{\eta_i, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} = \prod_{j=1}^3 \mathcal{M}_{\eta_j, u_j, T}(\mathbf{x}, t, v_j) dv_j. \quad (1.6)$$

In (1.6), $\mathbf{u}(\mathbf{x}, t) = (u_1(\mathbf{x}, t), u_2(\mathbf{x}, t), u_3(\mathbf{x}, t))^t$ and

$$\mathcal{M}_{\eta_j, u_j, T}(\mathbf{x}, t, v_j) = \eta_j^{\frac{1}{3}}(\mathbf{x}, t) \frac{\sqrt{m_j}}{\sqrt{2\pi T(\mathbf{x}, t)}} \exp\left(-m_j \frac{|v_j - u_j(\mathbf{x}, t)|^2}{2T(\mathbf{x}, t)}\right), \quad (1.7)$$

with $j \in \{1, \dots, M\}$ and $i \in \{1, 2, 3\}$. If f satisfies the above described equilibrium property, the first

²See how expression (10.62) is built for example.

³where $f(\mathbf{x}, t, \mathbf{v}) = \int f(\mathbf{x}, t, \mathbf{v}, \lambda) d\lambda$.

three moments of f are given by

$$\begin{aligned} \iint m(\lambda) f(\mathbf{x}, t, \mathbf{v}, \lambda) d\mathbf{v} d\lambda &= \rho(\mathbf{x}, t), \quad \text{which defines the mass density,} \\ \frac{1}{\rho(\mathbf{x}, t)} \iint m(\lambda) f(\mathbf{x}, t, \mathbf{v}, \lambda) \mathbf{v} d\mathbf{v} d\lambda &= \mathbf{u}(\mathbf{x}, t), \quad \text{which defines the bulk velocity,} \quad (1.8) \\ \frac{1}{\rho(\mathbf{x}, t)} \iint m(\lambda) f(\mathbf{x}, t, \mathbf{v}, \lambda) \frac{1}{3} |\mathbf{v} - \mathbf{u}(\mathbf{x}, t)|^2 d\mathbf{v} d\lambda &= T(\mathbf{x}, t), \quad \text{which defines the temperature,} \end{aligned}$$

where $\rho(\mathbf{x}, t) = \sum_{i=1}^M m_i \eta_i(\mathbf{x}, t)$ with $\eta_i(\mathbf{x}, t) = \int f_i(\mathbf{x}, t, \mathbf{v}) d\mathbf{v}$, $\forall i \in \{1, \dots, M\}$. The problem of existence and uniqueness of solutions, stability around global equilibrium in short/long times etc. of the quadratic Boltzmann equation must come with proper initial and boundary conditions. Some questions regarding the previous considerations are still not fully resolved [128, 248, 44, 91, 131]. We keep those kind of theoretical considerations and discussions for the particular cases treated in parts II and III.

Now, to understand the logical structure of this document, it is enough considering a mixture of only two species of particles, i.e. we have $M = 2$: we consider H -particles⁴, described by $f_H(\mathbf{x}, t, \mathbf{v}) \geq 0$, and l -ones⁵ described by $f_l(\mathbf{x}, t, \mathbf{v}) \geq 0$. In this case (1.1) resumes to

$$\left\{ \begin{array}{l} \partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) + F_H(\mathbf{x}, t, \mathbf{v}) \partial_{\mathbf{v}} f_H(\mathbf{x}, t, \mathbf{v}) \\ \qquad \qquad \qquad = Q_H(f_H, f_l)(\mathbf{x}, t, \mathbf{v}), \\ \qquad \qquad \qquad = Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}) + Q_{H,l}(f_H, f_l)(\mathbf{x}, t, \mathbf{v}), \\ \partial_t f_l(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_l(\mathbf{x}, t, \mathbf{v}) + F_l(\mathbf{x}, t, \mathbf{v}) \partial_{\mathbf{v}} f_l(\mathbf{x}, t, \mathbf{v}) \\ \qquad \qquad \qquad = Q_l(f_H, f_l)(\mathbf{x}, t, \mathbf{v}), \\ \qquad \qquad \qquad = Q_{l,l}(f_l, f_l)(\mathbf{x}, t, \mathbf{v}) + Q_{l,H}(f_l, f_H)(\mathbf{x}, t, \mathbf{v}). \end{array} \right. \quad (1.9)$$

Such kind of model is, for example, intensively used in plasma physics. The heavy particles would be ions whereas the light ones would be electrons. The forces applied depend on the electric and magnetic fields. The ion-ion collisions kernels could then model fusion reactions for example.

In the following sections, we study two asymptotical limits for system (1.9) which are central in the document.

1.1.1 One Hydrodynamic limit of Boltzmann equation

The first limit of equation (1.9) corresponds to the hydrodynamic one. The material of this section is inspired from [207, 128, 248], only slightly simplified to illustrate our purpose. To describe this first asymptotical limit, it is enough considering species H is alone, i.e. we have $f_l(\mathbf{x}, t, \mathbf{v}) = 0$. The methodology we present remains true and applicable for a mixture, see [207], the resulting model is only more complex and longer to introduce. Suppose furthermore the force applied to the particles is negligible, (1.9) simplifies to the scalar equation

$$\partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) = Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}). \quad (1.10)$$

Let us introduce

$$\left\{ \begin{array}{l} \mathbf{x} = \mathbf{x}^* \mathcal{X}, \mathbf{v} = \mathbf{v}^* \mathcal{V}, t = t^* \mathcal{T}, \\ \sigma_{H,H} = \sigma_{H,H}^* \frac{1}{\lambda_{H,H}}, \end{array} \right. \quad (1.11)$$

where the superscript * denotes a nondimensional variable. Quantity \mathcal{X} corresponds to a macroscopic observation length scale, $\lambda_{H,H}$ corresponds to the microscopic mean free path of the particles, \mathcal{V} corresponds to the bulk velocity and \mathcal{T} is the observation time scale. Let us introduce $f_H^*(\mathbf{x}^*, t^*, \mathbf{v}^*) = f_H(\mathbf{x}, t, \mathbf{v})$, then (1.10) can be rewritten

$$\frac{\mathcal{X}}{\mathcal{T} \mathcal{V}} \partial_{t^*} f_H^*(\mathbf{x}^*, t^*, \mathbf{v}^*) + \mathbf{v}^* \partial_{\mathbf{x}^*} f_H^*(\mathbf{x}^*, t^*, \mathbf{v}^*) = \frac{\mathcal{X}}{\lambda_{H,H}} Q_{H,H}^*(f_H^*, f_H^*)(\mathbf{x}^*, t^*, \mathbf{v}^*). \quad (1.12)$$

⁴H will stand for *Heavy*.

⁵l will stand for *light*.

Dropping the upperscripts $*$, we can rewrite (1.12) as

$$S_{tr}\partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) = \frac{1}{K_n} Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}). \quad (1.13)$$

In the above expression, $S_{tr} = \frac{\chi}{\mathcal{T}\mathcal{V}}$ is commonly called the Strouhal number and $K_n = \frac{\lambda_{H,H}}{\chi}$ the Knudsen number. Assume $K_n \sim 0$, then local thermodynamic equilibrium is reached almost instantaneously and $f_H(\mathbf{x}, t, \mathbf{v}) \sim \mathcal{M}_{\rho, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v})$. The state of the gas is only determined by the thermodynamic fields ρ, \mathbf{u}, T . In this document, we distinguish three ways to formally derive the different hydrodynamic limits, i.e. to identify the system of equations satisfied by the thermodynamic fields ρ, \mathbf{u}, T in the stiff regime $K_n \sim 0$:

- applying the *extended thermodynamic moment closure*, as described in [207], consists in first writing the local conservation laws

$$\begin{cases} \partial_t \int m f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} & + \partial_{\mathbf{x}} \int m \mathbf{v} f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} = 0, \\ \partial_t \int m \mathbf{v} f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} & + \partial_{\mathbf{x}} \int m \mathbf{v} \otimes \mathbf{v} f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} = 0, \\ \partial_t \int m \frac{|\mathbf{v}|^2}{2} f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} & + \partial_{\mathbf{x}} \int m \mathbf{v} \frac{|\mathbf{v}|^2}{2} f_H(\mathbf{x}, t, \mathbf{v}) d\mathbf{v} = 0. \end{cases} \quad (1.14)$$

In the above expression, we used the conservation relation satisfied by the collision kernel (1.5). The system is called a system of (hierarchical) moments: the flux of an equation is the unknown of the next one. It is not closed. The second key ingredient of the extended thermodynamic moment closure consists in assuming the solution f_H minimizes the entropy

$$s(f)(\mathbf{x}, t) = \int f(\mathbf{x}, t, \mathbf{v}) \ln(f(\mathbf{x}, t, \mathbf{v})) d\mathbf{v}. \quad (1.15)$$

Due to the H-theorem and the choice of considering only the first three moments⁶, it results in assuming $f \sim \mathcal{M}_{\rho, \mathbf{u}, T}$. The successive moments for the hydrodynamic fields, namely the macroscopic density $\rho(\mathbf{x}, t)$, the bulk velocity $\mathbf{u}(\mathbf{x}, t)$ and the temperature $T(\mathbf{x}, t)$ associated to $f_H(\mathbf{x}, t, \mathbf{v})$ as in (1.8) satisfy:

$$\begin{aligned} \int m f_H(\mathbf{x}, t, \mathbf{v}) - 1 & \quad d\mathbf{v} = \rho(\mathbf{x}, t), \\ \int m f_H(\mathbf{x}, t, \mathbf{v}) v_i & \quad d\mathbf{v} = \rho(\mathbf{x}, t) u_i(\mathbf{x}, t), \quad \forall i \in \{1, 2, 3\}, \\ \int m f_H(\mathbf{x}, t, \mathbf{v}) v_i v_j & \quad d\mathbf{v} = \rho(\mathbf{x}, t) u_i(\mathbf{x}, t) u_j(\mathbf{x}, t) + \delta_{i,j} \rho(\mathbf{x}, t) T(\mathbf{x}, t), \quad \forall i, j \in \{1, 2, 3\}, \\ \int m f_H(\mathbf{x}, t, \mathbf{v}) \frac{|\mathbf{v}|^2}{2} & \quad d\mathbf{v} = \frac{1}{2} \rho(\mathbf{x}, t) |\mathbf{u}(\mathbf{x}, t)|^2 + \frac{3}{2} \rho(\mathbf{x}, t) T(\mathbf{x}, t), \\ \int m f_H(\mathbf{x}, t, \mathbf{v}) \frac{|\mathbf{v}|^2}{2} v_i & \quad d\mathbf{v} = u_i \left(\frac{1}{2} \rho(\mathbf{x}, t) |\mathbf{u}(\mathbf{x}, t)|^2 + \frac{5}{2} \rho(\mathbf{x}, t) T(\mathbf{x}, t) \right), \quad \forall i \in \{1, 2, 3\}. \end{aligned}$$

In term of thermodynamic quantities, (1.14) can be rewritten

$$\begin{cases} \partial_t \rho(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho(\mathbf{x}, t) \mathbf{u}(\mathbf{x}, t)) = 0, \\ \partial_t \rho(\mathbf{x}, t) \mathbf{u}(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho(\mathbf{x}, t) \mathbf{u}(\mathbf{x}, t) \otimes \mathbf{u}(\mathbf{x}, t) + p(\mathbf{x}, t) I_3) = 0, \\ \partial_t \rho(\mathbf{x}, t) e(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho(\mathbf{x}, t) \mathbf{u}(\mathbf{x}, t) e(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t) p(\mathbf{x}, t)) = 0. \end{cases} \quad (1.16)$$

In (1.16), I_3 is the identity matrix of size 3, $e(\mathbf{x}, t) = \varepsilon(\mathbf{x}, t) + \frac{1}{2} |\mathbf{u}(\mathbf{x}, t)|^2$ and the closure relations are given by

$$\varepsilon(\mathbf{x}, t) = \frac{3}{2} T(\mathbf{x}, t) \text{ and } p(\mathbf{x}, t) = \rho(\mathbf{x}, t) T(\mathbf{x}, t) = (\gamma - 1) \rho(\mathbf{x}, t) \varepsilon(\mathbf{x}, t) \text{ with } \gamma = \frac{5}{3}. \quad (1.17)$$

System (1.16) together with the closure relation is called the Euler system for a perfect mono-atomic gas (relative to $\gamma = \frac{5}{3}$). System (1.16) enters the more general fields of systems of conservation laws

⁶This can be generalized to an arbitrary number of moments.

which are studied, from an uncertainty quantification point of view, in part II. The way the system is closed, applying the extended thermodynamic moment closure only up to the third moment, is also known as the M_n closure model with $n = 3$, i.e. the M_3 model. It is generalized to uncertain systems of conservation laws in part II.

- On another hand, the same hydrodynamic limit can also be recovered performing a Hilbert development [128, 248, 143]. It consists in assuming $\frac{1}{K_n} = \mathcal{O}(\frac{1}{\delta})$ together with $S_{tr} = \mathcal{O}(1)$ and expanding f_H as a formal power serie $f_H(\mathbf{x}, t, \mathbf{v}) = \sum_{k=0}^{\infty} f_H^k(\mathbf{x}, t, \mathbf{v}) \delta^k$ which, once plugged into the dimensionless equation (1.13), leads to

$$\partial_t \begin{pmatrix} 0 \\ \sum_{k=2}^{\infty} f_H^k \delta^k \end{pmatrix} + \mathbf{v} \partial_{\mathbf{x}} \begin{pmatrix} 0 \\ \sum_{k=2}^{\infty} f_H^k \delta^k \end{pmatrix} = \begin{pmatrix} Q_{H,H}(f_H^0, f_H^0) \\ \sum_{k=2}^{\infty} \sum_{i+j=k} Q_{H,H}(f_H^i, f_H^j) \delta^k \end{pmatrix}. \quad (1.18)$$

The equations governing the coefficients f_H^k are obtained equating the coefficients multiplying the successive powers of δ . Order 0 (i.e. δ^0 , first line in (1.18)) gives

$$Q_{H,H}(f_H^0, f_H^0) = 0,$$

and we recover the leading order f_H^0 is the local Maxwellian, of the form (1.6). In other words, the leading order is solution of the Euler system up to $\mathcal{O}(\delta)$

$$\begin{cases} \partial_t \rho^0(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho^0(\mathbf{x}, t) u^0(\mathbf{x}, t)) = \mathcal{O}(\delta), \\ \partial_t \rho^0(\mathbf{x}, t) u^0(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho^0(\mathbf{x}, t) u^0(\mathbf{x}, t) \otimes u^0(\mathbf{x}, t) + p^0(\mathbf{x}, t) I_3) = \mathcal{O}(\delta), \\ \partial_t \rho^0(\mathbf{x}, t) e^0(\mathbf{x}, t) & + \partial_{\mathbf{x}} (\rho^0(\mathbf{x}, t) u^0(\mathbf{x}, t) e^0(\mathbf{x}, t) + u^0(\mathbf{x}, t) p^0(\mathbf{x}, t)) = \mathcal{O}(\delta), \end{cases} \quad (1.19)$$

with the same closure relations (1.17). Naturally, the Hilbert development allows considering fluctuations around the leading order 0 by studying orders 1, 2, ... (i.e. $\delta^1, \delta^2, \dots$ and the next lines of (1.18)) etc. to derive finer corrections. The first order equation is solution of

$$\partial_t f_H^0 + \mathbf{v} \partial_{\mathbf{x}} f_H^0 = Q(f_H^0, f_H^1) + Q(f_H^1, f_H^0), \quad (1.20)$$

and leads to the Navier-Stokes system, see [128], once plugged into (1.19) and after the introduction of some $\mathcal{O}(\delta^2)$ viscous corrections. Hilbert developments are central in both parts of this document: in part II they are a useful tool to bridge the gap between perturbation models and Polynomial Chaos ones (see section 7). In part III, they are used to identify particular (stiff) regimes and make sure the (Monte-Carlo in part III) resolution scheme we build allows capturing the limit with a good accuracy. Other different expansions can be encountered in the literature: for example the Chapman-Enskog development [248, 62] is based on an expansion of general structure $f_H(\mathbf{x}, t, \mathbf{v}) = \mathcal{M}_{\rho, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v})(1 + \mathcal{O}(\delta))$ and consequently requires an *a priori* hypothesis⁷.

- Grad's development is another example of asymptotical expansion to derive limits of the Boltzmann equations [207]. It is based on the identification, up to a certain order, of the coefficients $(f_H^k)_{k \in \mathbb{N}}$ in the serie

$$\left(\sum_{k=0}^{\infty} f_H^k(\mathbf{x}, t, \mathbf{v}) \frac{\partial^{k_1+k_2+k_3=k}}{\partial^{k_1} v_1 \partial^{k_2} v_2 \partial^{k_3} v_3} \right) \mathcal{M}_{\rho, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v}), \quad (1.21)$$

once plugged in the Boltzmann equation. It does not rely on any assumption on the form of the distribution, it is only assumed not too far from the Maxwellian distribution: the first order f_H^1 captures fluctuations around $\mathcal{M}_{\rho, \mathbf{u}, T}$. Each derivative of the Maxwellian with respect to the velocity components (v_1, v_2, v_3)

$$H_k(v_1, v_2, v_3) = \frac{\partial^{k_1+k_2+k_3=k}}{\partial^{k_1} v_1 \partial^{k_2} v_2 \partial^{k_3} v_3} \mathcal{M}_{\rho, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v}),$$

⁷The Maxwellian form of the $\mathcal{O}(1)$ coefficient is assumed and not deduced from an analysis as for the Hilbert one.

corresponds to the k^{th} component of the three dimensional Hermite polynomials, see [5, 117, 207]. Grad's 13 moments model is based on such development of the Boltzmann equation onto the components of the (3D) Hermite basis: very similar developments are central in part II in an uncertainty quantification context, especially with the introduction of Polynomial Chaos (see remark 3.1).

Along the previous lines, we built, *via* different ways, a reduced model from the quadratic Boltzmann one. The latter is relevant in many regimes (it is intensively used to model rarefied gas, i.e. when $K_n \not\approx 0$) but more complex to solve than Euler system when hypothesis $K_n \sim 0$ applies. The aim of part II is also to present the construction of reduced models designed to capture probabilistic features of a given set of stochastic PDE.

Every presented methodologies, extended thermodynamic moment closure or Hilbert developments or Grad's model, allow deriving the hydrodynamic limit of Boltzmann equation. It can be recast in the more general form of a system of conservation laws:

$$\partial_t U(\mathbf{x}, t) + \partial_{\mathbf{x}} F(U(\mathbf{x}, t)) = 0, \quad (1.22)$$

with

$$U(\mathbf{x}, t) = \begin{pmatrix} \rho(\mathbf{x}, t) \\ \rho(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t) \\ \rho(\mathbf{x}, t)e(\mathbf{x}, t) \end{pmatrix} \quad \text{and} \quad F(U(\mathbf{x}, t)) = \begin{pmatrix} \rho(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t) \\ \rho(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t) \otimes \mathbf{u}(\mathbf{x}, t) + p(\mathbf{x}, t)I_3 \\ \rho(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t)e(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t)p(\mathbf{x}, t) \end{pmatrix}, \quad (1.23)$$

together with (1.17). Those systems are of interest in many physical fields: hydrodynamics [195, 48, 78, 79, 182, 154], continuum mechanics [160, 183, 119], plasma physics [274, 276] etc. In section 1.1.3, we go through some resolution strategies for such systems (in particular for Euler equations) but before we identify another limit of Boltzmann equation which is intensively studied in part III of this document.

1.1.2 The Linear Boltzmann equation limit

The second limit corresponds to the linear Boltzmann one. To derive it, we rely on a Hilbert development as in the previous section. Let us consider the coupled system (1.9) (with $F_H = F_l = 0$) and introduce some nondimensional variables similar to (1.11) for every species:

$$\begin{cases} \mathbf{x} = \mathbf{x}^* \mathcal{X}, t = t^* \mathcal{T}, \\ \mathbf{v} = \mathbf{v}_H^* \mathcal{V}_H, \\ \mathbf{v} = \mathbf{v}_l^* \mathcal{V}_l, \\ \sigma_{i,j} = \sigma_{i,j}^* \frac{1}{\lambda_{i,j}}, \text{ with } i, j \in \{H, l\}. \end{cases} \quad (1.24)$$

In (1.24), we introduce a bulk velocity for each kind of particles. We have (we perform the nondimensionalization and drop the upperscript *)

$$\begin{cases} \partial_t f_H(\mathbf{x}, t, \mathbf{v}_H) + \mathbf{v}_H \frac{\mathcal{V}_H \mathcal{T}}{\mathcal{X}} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}_H) = + \frac{\mathcal{V}_H \mathcal{T}}{\lambda_{H,H}} Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}_H) \\ \quad + \frac{\mathcal{V}_H \mathcal{T}}{\lambda_{H,l}} Q_{H,l}(f_H, f_l)(\mathbf{x}, t, \mathbf{v}_H), \\ \partial_t f_l(\mathbf{x}, t, \mathbf{v}_l) + \mathbf{v}_l \frac{\mathcal{V}_l \mathcal{T}}{\mathcal{X}} \partial_{\mathbf{x}} f_l(\mathbf{x}, t, \mathbf{v}_l) = + \frac{\mathcal{V}_l \mathcal{T}}{\lambda_{l,l}} Q_{l,l}(f_l, f_l)(\mathbf{x}, t, \mathbf{v}_l) \\ \quad + \frac{\mathcal{V}_l \mathcal{T}}{\lambda_{l,H}} Q_{l,H}(f_l, f_H)(\mathbf{x}, t, \mathbf{v}_l). \end{cases} \quad (1.25)$$

In (1.24), the bulk velocities of H -particles is different than the one of particles of type l . But we assume their bulk momentum $\mathcal{P} = m_H \mathcal{V}_H = m_l \mathcal{V}_l$ is the same. It allows rewriting (1.25) with respect to m_H ,

m_l and \mathcal{P}

$$\left\{ \begin{array}{lcl} \partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \frac{\mathcal{PT}}{m_H \mathcal{X}} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) & = & + \frac{\mathcal{PT}}{m_H \lambda_{H,H}} Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}) \\ & & + \frac{\mathcal{PT}}{m_H \lambda_{H,l}} Q_{H,l}(f_H, f_l)(\mathbf{x}, t, \mathbf{v}), \\ \partial_t f_l(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \frac{\mathcal{PT}}{m_l \mathcal{X}} \partial_{\mathbf{x}} f_l(\mathbf{x}, t, \mathbf{v}) & = & + \frac{\mathcal{PT}}{m_l \lambda_{l,l}} Q_{l,l}(f_l, f_l)(\mathbf{x}, t, \mathbf{v}) \\ & & + \frac{\mathcal{PT}}{m_l \lambda_{l,H}} Q_{l,H}(f_l, f_H)(\mathbf{x}, t, \mathbf{v}). \end{array} \right. \quad (1.26)$$

Now suppose

$$\begin{aligned} - \frac{\mathcal{PT}}{m_H \mathcal{X}} &= \mathcal{O}(\delta) = \frac{\mathcal{PT}}{m_l \lambda_{l,l}} = \frac{\mathcal{PT}}{m_H \lambda_{H,l}}, \\ - \frac{\mathcal{PT}}{m_l \mathcal{X}} &= \mathcal{O}(1) = \frac{\mathcal{PT}}{m_l \lambda_{l,H}} = \frac{\mathcal{PT}}{m_H \lambda_{H,H}}. \end{aligned}$$

It implicitly makes H -particles *heavy* ones in comparison to l -ones as $\frac{m_l}{m_H} = \mathcal{O}(\delta)$. Then system (1.26) becomes

$$\left\{ \begin{array}{l} \partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \delta \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) = Q_{H,H}(f_H, f_H)(\mathbf{x}, t, \mathbf{v}) + \delta Q_{H,l}(f_H, f_l)(\mathbf{x}, t, \mathbf{v}), \\ \partial_t f_l(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_l(\mathbf{x}, t, \mathbf{v}) = \delta Q_{l,l}(f_l, f_l)(\mathbf{x}, t, \mathbf{v}) + Q_{l,H}(f_l, f_H)(\mathbf{x}, t, \mathbf{v}). \end{array} \right. \quad (1.27a)$$

Let us identify the leading orders in (1.27):

- the $\mathcal{O}(\delta^0)$ order for (1.27a) ensures $\partial_t f_H^0 = Q_{H,H}(f_H^0, f_H^0)$. It is in agreement with heavy H -particles being not very inertial so that their flux only depends on higher orders $(f_H^i)_{i>0}$. Besides, l -particles do not affect the dynamic of the H -ones.
- The $\mathcal{O}(\delta^0)$ order for (1.27b) yields

$$\partial_t f_l^0(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_l^0(\mathbf{x}, t, \mathbf{v}) = Q_{l,H}(f_l^0, f_H^0)(\mathbf{x}, t, \mathbf{v}). \quad (1.28)$$

- The zeroth order term shows the regime neglects $l-l$ collisions (no term $Q(f_l^0, f_l^0)$ in (1.28)). The first order one (i.e. the factor of $\mathcal{O}(\delta^1)$) shows the regime ensures a relaxation toward the local Maxwellian distribution for f_l as $Q_{l,l}(f_l^0, f_l^0) = 0$ when δ becomes non-negligible (for late times for example).

Now assume we initially have $f_H^0(\mathbf{x}, 0, \mathbf{v}) = \mathcal{M}_{\eta_H, \mathbf{u}, T}(\mathbf{v})$, $\forall \mathbf{x} \in \mathcal{D}$. Then $Q(f_H^0, f_H^0)(\mathbf{x}, 0, \mathbf{v}) = 0$ and $f_H^0(\mathbf{x}, t, \mathbf{v}) = \mathcal{M}_{\eta_H, \mathbf{u}, T}(\mathbf{v})$, $\forall \mathbf{x} \in \mathcal{D}, t \in [0, T]$. For such dense particles H and initial condition, f_H^0 behaves as a *background* Maxwellian distribution and (1.25) degenerates toward a scalar equation given by

$$\begin{aligned} \partial_t f_l^0(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_l^0(\mathbf{x}, t, \mathbf{v}) &= Q_{l,H}(f_l^0, \mathcal{M}_{\eta_H, \mathbf{u}, T})(\mathbf{x}, t, \mathbf{v}), \\ &= \int \underbrace{|\mathbf{v} - \mathbf{v}_H| \sigma_{l,H}(\mathbf{v} - \mathbf{v}_H) \mathcal{M}_{\eta_H, \mathbf{u}, T}(\mathbf{v}'_H(\mathbf{v}, \mathbf{v}_H)) f_l^0(\mathbf{x}, t, \mathbf{v}'(\mathbf{v}, \mathbf{v}_H))}_{|\mathbf{v}| \sigma_s(\mathbf{x}, t, \mathbf{v}_H, \mathbf{v})} d\mathbf{v}_H \\ &\quad - \int \underbrace{|\mathbf{v} - \mathbf{v}_H| \sigma_{l,H}(\mathbf{v} - \mathbf{v}_H) \mathcal{M}_{\eta_H, \mathbf{u}, T}(\mathbf{v}_H) d\mathbf{v}_H}_{|\mathbf{v}| \sigma_t(\mathbf{x}, t, \mathbf{v})} f_l^0(\mathbf{x}, t, \mathbf{v}). \end{aligned} \quad (1.29)$$

We then obtain, after some simple rearrangements, the linear Boltzmann equation more commonly written

$$\partial_t f_l(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_l(\mathbf{x}, t, \mathbf{v}) + v \sigma_t(\mathbf{x}, t, \mathbf{v}) f_l(\mathbf{x}, t, \mathbf{v}) = \int v \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') f_l(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}'. \quad (1.30)$$

Above, v denotes the norm of the velocity $v = |\mathbf{v}| \in \mathbb{R}^+$ of the particles and $\omega = \frac{\mathbf{v}}{|\mathbf{v}|} \in \mathbb{S}^2$ denotes their direction. Equation (1.30), together with some nonlinear couplings, are central in part III. It is commonly used to model the transport of neutrons or photons for examples (see the references of part

III). We focus on its resolution with a stochastic (Monte-Carlo) resolution strategy.

In the two next sections (1.1.3) and (1.1.4), we briefly go through few resolution strategies to numerically approximate solutions of the two presented models (1.22) and (1.30). The aim is not to present an exhaustive list and analysis of common resolution schemes for the two presented limits. But some of them, or their main principle, are hinted at in the document *via* references. We think it can ease the understandability of the manuscript to give some quick illustrations here. They may be of interest for the reader unfamiliar with some of the notions therein and willing to have a quick glance at them. In the next section, we briefly recall the main principle of deterministic and stochastic resolution strategies to solve the two limits of interest (1.22) and (1.30).

1.1.3 Deterministic resolution schemes to solve (1.22) and (1.30)

Let us begin with the numerical resolution of the linear Boltzmann equation (1.30): the first reflex to solve (1.30) would be to introduce a grid with respect to variables \mathbf{x}, t and \mathbf{v} . Assume $\mathbf{x} \in \mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$, a mesh of non-overlapping cells of size $\Delta\mathbf{x}$ for the spatial discretisation, $t \in [0, T] = \bigcup_{n=1}^{N_t} [t^n, t^{n+1}]$ and $\mathbf{v} \in \mathbb{R}^3 = \bigcup_{i=1}^{N_v} \mathcal{F}_i$ a regular mesh of non-overlapping cells of size $\Delta\mathbf{v}$ in the phase domain⁸. For example, a *Finite Volume* (FV) scheme [106] can be built to solve (1.30) by integrating the equation on some control volume $V_{i,j} = \mathcal{D}_i \times \mathcal{F}_j$ and introducing

$$u_{i,j}^n = \frac{1}{|V_{i,j}|} \int_{V_{i,j}} u(\mathbf{x}, t^n, \mathbf{v}) d\mathbf{x} d\mathbf{v},$$

with $|V_{i,j}| = \Delta\mathbf{x}\Delta\mathbf{v}$. If we now integrate (1.30) on every control volumes $(\mathcal{D}_i \times \mathcal{F}_j)_{i \in \{1, \dots, N_x\}, j \in \{1, \dots, N_v\}}$ and with respect to time in $[t^n, t^{n+1}]$ we have

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} + \bar{v}_j \frac{1}{\Delta\mathbf{x}} \int_{\partial\mathcal{D}_i} u_j^*(y) dy + \bar{v}_j \sigma_{t,i}^j u_{i,j}^n = \sum_{l=1}^{N_v} \bar{v}_j \sigma_{s,i}^{j,l} w_l u_{i,l}^n. \quad (1.31)$$

In (1.31), we

- globally explicited the terms,
- assumed constant velocities $\bar{v}_j \approx v \in \mathcal{F}_j, \forall j \in \{1, \dots, N_v\}$,
- and introduced a more general notation for $w_l = \Delta\mathbf{v}$.

To close the discrete system (1.31), it remains to define the discretisation of the spatial operator (i.e. $(u_j^*(y))_{j \in \{1, \dots, N_v\}}$) on the boundaries of the spatial counterpart of the control volume (commonly called the flux). Independently of the discretisation choice for this spatial operator, the coupling in the v -direction is done *via* the discretised collision kernel⁹, i.e. *via* $(\sigma_{s,i}^{j,l})_{i,j,l \in \{1, \dots, N_x\} \times \{1, \dots, N_v\} \times \{1, \dots, N_v\}}$ on the right hand side of (1.31). We do not aim at identifying a particular scheme here¹⁰. We just want to insist on the fact that, independently of the previous discretisation choices, the size of the mesh can be important for fine resolutions: for a first order discretisation, in dimension 6, increasing the accuracy of a factor 2 on a mesh of size $N_x^3 \times N_v^3$ implies having a factor

- $\frac{(2N_x)^3 \times (2N_v)^3}{N_x^3 \times N_v^3} = 64$ on the memory consumption .
- $\frac{2\Delta t \times (2N_x)^3 \times (2N_v)^3}{\Delta t \times N_x^3 \times N_v^3} = 128$ on the computational time (due to the cfl condition demanding Δt to be proportional to $\Delta\mathbf{x}$).

Note that if the control volumes in the velocity direction are chosen as quadrature rules (the weights $(w_l \neq \Delta\mathbf{v})_{l \in \{1, \dots, N_v\}}$ are not necessarily uniform anymore), such scheme is referred as a *discrete ordinate* one or a S_n discretisation [16, 92, 61, 116, 56] where n is the number of points of the quadrature (n

⁸with proper cut-off or weighting at infinity, this is not the purpose of the discussion here.

⁹i.e. if $\sigma_s = 0$, each equation for $j \in \{1, \dots, N_v\}$ are independent.

¹⁰We refer to [16, 92, 61] and the references therein for this purpose.

replace N_v as a discretisation parameter). Such strategy is applied in practice to ensure, for the same number of control volumes in the v -direction, a better accuracy in the integration of the collision kernel.

To avoid an exponential increase of computational effort with the needed accuracy, it is for example possible to introduce a different *basis of approximation* for the velocity dimension and a discretisation parameter n of a different kind. A P_n discretisation implies formally developing the solution $u(\mathbf{x}, t, \mathbf{v}) \approx \sum_{k=0}^n u_k(\mathbf{x}, t) P_k(\mathbf{v})$ of (1.30) on the Legendre polynomial basis $(P_k)_{k \in \mathbb{N}}$ instead of considering a sum of weighted Dirac masses as for the S_n one. The construction of the P_n model consists in plugging the previous development in (1.30) and performing a Galerkin projection (i.e. use the orthogonality of the polynomial basis for a given scalar product). We obtain a *reduced model* of unknown $U(\mathbf{x}, t) = (u_0(\mathbf{x}, t), \dots, u_n(\mathbf{x}, t))^t$ depending only on $\mathbf{x} \in \mathcal{D}$ and $t \in [0, T]$, see [281, 141] for example. The idea is to trade dimension (only \mathbf{x}, t to treat instead of $\mathbf{x}, t, \mathbf{v}$) to the size of the system to solve (U is now of size $n + 1$ whereas u was of size 1). The previous development implicitly introduces the closure hypothesis $u_k(\mathbf{x}, t) = 0, \forall k > n$. A different way to close the system consists in applying the material of section 1.1.1 with extended thermodynamic of moments: it consists in the same Galerkin projection together with the assumption that the distribution f_l minimizes entropy (1.15) given the constraints on the moments $(u_k)_{k \in \{0, \dots, n\}}$. References [226, 133, 279, 228, 94] present an interesting and detailed example for phototherapy applications. Such closures are commonly called M_n models. They are intensively studied in part II in an uncertainty quantification for systems of conservation laws' context. The three methodologies (S_n, P_n, M_n) confer to the reduced model the structure of a hyperbolic system of conservation laws,

$$\partial_t U(\mathbf{x}, t) + A \partial_x U(\mathbf{x}, t) + \Sigma_t U(\mathbf{x}, t) = \Sigma_s U(\mathbf{x}, t),$$

where A, Σ_t, Σ_s are particular matrices, see [281, 141]. In fact, such class of models are commonly called Friedrich's systems, see [205] and the reference therein. In other words it can be rewritten under the same general form as (1.22) for the Euler system. The same resolution strategies can consequently be applied. Such kind of developments and their properties (S_n, P_n and M_n) are intensively studied in part II of this document, in an uncertainty quantification context.

In this document, when dealing with hyperbolic systems of conservation laws, Finite Volume (FV) schemes are applied. Some are built in part II of this document for particular systems of conservation laws. Their construction corresponds to a significant amount of work/time for the resolution strategies of part II but they are not necessarily detailed here. On another hand, they are fully described in [236, 232, 237, 243].

1.1.4 Stochastic resolution schemes to solve (1.22) and (1.30)

Stochastic resolution methods imply the sampling of random variables (or vectors) in opposition to deterministic ones which do not. The term *Monte-Carlo* usually denotes such strategies. The description of Monte-Carlo schemes for the resolution of the deterministic linear Boltzmann equation is the object of part III. They rely on rewriting the integro-differential equation (1.30) as an integral one (and as an expectation over a set of identified random variables afterward) and demand linearity. The subject is intensively studied in part III. Despite the nonlinearity of system (1.22), Monte-Carlo methods can be applied: the methodology is commonly denoted a *kinetic scheme* and rely on linearizing Boltzmann equation (1.28) in the vicinity of equilibrium. For Euler system, this resumes to using the Maxwellian distribution directly in an *artificial* collision kernel to rewrite:

$$\partial_t f_H(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} f_H(\mathbf{x}, t, \mathbf{v}) = - \frac{f_H(\mathbf{x}, t, \mathbf{v}) - \mathcal{M}_{\rho, \mathbf{u}, T}(\mathbf{x}, t, \mathbf{v})}{\tau}. \quad (1.32)$$

Equation (1.32) is called the BGK model, see [193, 22, 27] for example, and introduces an additional parameter τ (echoing δ of the Hilbert development), a relaxation time controlling at which speed $f_H \sim \mathcal{M}_{\rho, \mathbf{u}, T}$. The more $\tau \rightarrow 0$, the faster $f_H \rightarrow \mathcal{M}_{\rho, \mathbf{u}, T}$ and the more its three first moments coincide with the solution of the Euler system. Of course, as such, (1.32) is not linear as $\rho(f_H), \mathbf{u}(f_H), T(f_H)$. It needs an additional discretisation hypothesis (for example $\rho, u, T \approx \rho^n, u^n, T^n$ on time step $[t^n, t^{n+1}]$). Once linearized, the material of chapter 9 of part III can be applied. For a nonlinear problem, the choice of the linearisation is central in practice and can confer interesting (or bad) properties to the numerical

resolution: chapter 10 of part III is strewn with (good and bad) discretisation examples. The chapter is devoted to the study and analysis of MC solvers for coupled nonlinear systems involving the linear Boltzmann equation (1.30) and presents my original contribution in this field.

1.2 V&V: the role of numerical analysis, UQ and HPC

In the previous paragraphs, the two parts of the manuscript were presented *via* two models, systems of conservation laws and the (non)linear Boltzmann equation. Another way to present those parts is *via* the objectives and stakes with respect to these models. The study of Euler system, in this document, is related to uncertainty analysis. More precisely, it is related to uncertainty propagation, presented in chapter 2 and appendix A. On the other hand, the design of Monte-Carlo schemes for the (non)linear Boltzmann equation is related to numerical analysis. Both disciplines are tools for Verification & Validation (V&V), see [13]. The fact the two types of models are related to two different fields of analysis also expresses the fact that the needs are different with respect to V&V. This is emphasized in the following sections.

1.2.1 Verification & Validation (V&V)

Verification & Validation is of particular importance at the CEA DAM: in an intensive simulation context to guaranty the performances of devices without being anymore able to perform scale 1 : 1 experiments. More generally, it is of importance for any industrial willing to make effective use of simulations and avoid the multiplication¹¹ of dangerous or polluting or costly experiments. The general sketch (see figure

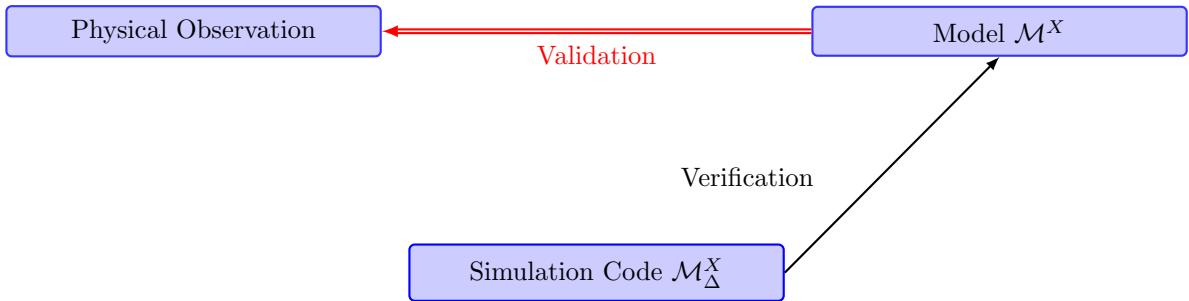


Figure 1.1: General sketch for Verification & Validation [13].

1.1) sums up the main steps of a V&V framework, its finer bricks will be detailed in the next sections. Every study aims at validating or invalidating the hypothesis

$$\text{hypothesis: model } \mathcal{M}^X \text{ is relevant to represent my physical observation.} \quad (1.33)$$

The framework depicted in figure 1.1 intends to be general but typically, in part II of this document, model \mathcal{M}^X denotes a system of conservation laws of general term (recall (1.22))

$$\mathcal{M}^X(U(\mathbf{x}, t)) = 0 \iff \partial_t U(\mathbf{x}, t) + \partial_{\mathbf{x}} F(U(\mathbf{x}, t)) = 0. \quad (1.34)$$

In part III on the other hand, we have (recall (1.30))

$$\begin{aligned} \mathcal{M}^X(u(\mathbf{x}, t, \mathbf{v})) &= 0 \\ \iff \partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + v \sigma_t(\mathbf{x}, t, \mathbf{v}) u(\mathbf{x}, t, \mathbf{v}) - \int v \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}' &= 0. \end{aligned} \quad (1.35)$$

The concise writing \mathcal{M}^X in order to denote the model is convenient but may hide some relevant properties of the solution of the model. For example, see part II, systems of conservation laws are known to develop

¹¹Note that the aim is to reduce wisely the number of experiments. Some will remain mandatory.

discontinuous solutions in finite times [81, 260, 81, 261]. The (lack of) regularity of the solution may be hidden by the too simple *black-box* notation \mathcal{M}^X . This is of importance as many numerical methods strongly depend on some smoothness assumptions (see mainly chapters 4 and 5) which may not hold. For this reason in the next parts, care will be taken to recall the whole set of PDEs of interest. Notation \mathcal{M}^X will in fact only apply in this section 1.2, convenient to denote indifferently the two models of interest in this manuscript.

Now, as presented in figure 1.1, model \mathcal{M}^X is intermediary between the physical observation of interest and the simulation code to approximate it. The question is *how can we validate or invalidate hypothesis (1.33) having only access to an approximation \mathcal{M}_Δ^X of the model \mathcal{M}^X ?* In the next sections, we briefly present how we can make effective use of the two main tools for V&V, numerical analysis and uncertainty analysis, to answer the above question.

1.2.2 Numerical analysis: the main tool for verification

Model \mathcal{M}^X can not, in general, be solved analytically. We need to rely on an approximation of its unknowns. Its accuracy is here controlled by Δ . In part II, we rely on finite volume schemes with

$$\Delta = \Delta_x = \max_{i \in \{1, \dots, N_x\}} |\mathcal{D}_i|.$$

Numerical parameter Δ_x is the maximum size of N_x non-overlapping cells in a grid tesselating the simulation domain $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$. In part III, we rely on Monte-Carlo schemes such that

$$\Delta = \frac{1}{\sqrt{N_{MC}}}.$$

Numerical parameter N_{MC} is the number of MC particles¹² of the simulation. Notation Δ can even denote a vector of numerical parameters controlling the accuracy of \mathcal{M}_Δ^X with respect to \mathcal{M}^X . For example, suppose model \mathcal{M}^X couples a system of conservation laws and the linear Boltzmann equation¹³. Suppose furthermore that the former is solved thanks to a finite volume scheme and the latter by a Monte-Carlo scheme: then $\Delta = (\Delta_x, \frac{1}{\sqrt{N_{MC}}})$. In this case (and with well built discretisation schemes), both discretisation parameters must go to zero in order to obtain a converging approximation. Multiple discretisation parameter methods will be intensively discussed in this manuscript¹⁴. Making sure Δ ensures a converging behaviour of the solution as $\Delta \rightarrow 0$ is the purpose of numerical analysis, main tool for verification. Numerical analysis will better be put in the V&V context in section 1.2.4.

1.2.3 Uncertainty Quantification (UQ): the main tool for validation

Now, model \mathcal{M}^X , independently of any approximation parameter Δ , also depends on a vector of physical¹⁵ parameters X . Vector X is typically related to some quantities involved in the closure/constitutive relations¹⁶ of the model or some fluctuations in the initial¹⁷ or the boundary¹⁸ conditions.

Parameter X affects the model hence its solution. By convention in the literature, when one aims at performing an uncertainty analysis, the dependence with respect to X of the solution of the model is made explicit. In other words in part II, our unknown is $U(\mathbf{x}, t, X)$ solution of $\mathcal{M}^X = \mathcal{M}(U(\mathbf{x}, t, X))$.

The variability range of X affects considerably the solution. In sketch 1.2, the fluctuations of X are characterised probabilistically. This is to reproduce something we experimentally observe: two

¹²A 'MC particle' will be defined in part III.

¹³This is the case for example in radiative hydrodynamics, see [48, 178, 59, 203].

¹⁴See for example the discussions related to figure 4.2 (right) in part II or figure 9.3 in part III. See also section 10.2.1 in which a flaw in a Monte-Carlo scheme for photonics is put forward. The solver has two competing discretisation parameters $\Delta = (\Delta_x, \Delta_t)$: taking $\Delta_x \rightarrow 0$ and $\Delta_t \rightarrow 0$ does not necessarily ensure $\mathcal{M}_\Delta^X \rightarrow \mathcal{M}^X$. A new MC scheme is then introduced in section 10.2.2, avoiding the previous problematic behaviour.

¹⁵The term *physical* here is mainly used in opposition to the term *numerical* of the previous paragraph.

¹⁶For systems of conservation laws in part II, X can parameter the flux F . It can refer to some parameters in the equation of state p (such as the heat capacity ratio γ for a perfect gas in hydrodynamics), or in the constitutive laws of materials (such as the Gruneisen coefficient in continuum mechanics) etc. For the linear Boltzmann equation, X can parameter the cross-sections/opacities $(\sigma_\alpha)_{\alpha \in \{s, t\}}$.

¹⁷The exact ambient temperature in the experimental chamber for example.

¹⁸The exact incoming velocity of a fluid in a wind tunnel for example [230, 229].

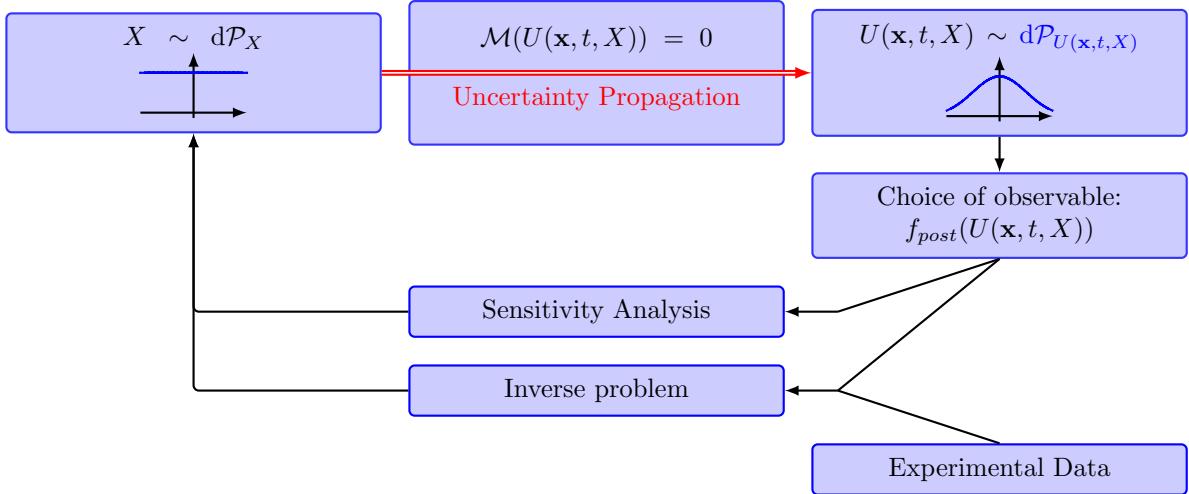


Figure 1.2: General sketch for uncertainty analysis [271].

independent identical experiments do not necessarily lead to exactly the same results. Notation $X \sim dP_X$ should be read: X follows the distribution having probability measure dP_X . In this document, the choice has been made to rely on probability theory to model the uncertainty. Some alternatives exist, see for example [86, 88, 87], but discussions about the relevance of these recent frameworks is beyond the scope of this manuscript. Besides, probability theory also plays an important role in part III. Parameter X being characterised by random variables, the solution $U(\mathbf{x}, t, X)$ is a random process (see appendix A). The uncertainty propagation step aims at characterising probabilistically $U(\mathbf{x}, t, X)$, i.e. finding $dP_{U(\mathbf{x}, t, X)}$ such that $U(\mathbf{x}, t, X) \sim dP_{U(\mathbf{x}, t, X)}$. The propagation step is mandatory but often only a first step toward sensitivity analysis or calibration etc. (see sketch 1.2). In this manuscript, we will focus on the propagation step, central to carry out any kind of uncertainty study, even if (Bayesian) calibration [31, 30] or sensitivity analysis [138, 241] were tackled in my publications.

1.2.4 Numerical/uncertainty analysis as tools for V&V and the role of HPC

In order to better understand how numerical and uncertainty analysis can help answer the question of paragraph 1.2.1 and validating or invalidating hypothesis (1.33), let us perform some very simple calculations¹⁹.

Assume we observe, during an experiment, a quantity U_r : the subscript r is for *reality*. Now, during that experiment, there are some parameters X which are hard to control and for which we only have a finite accuracy. As a consequence, we only have access to U_{exp} , a noisy approximation of U_r such that

$$U_r = U_{\text{exp}} + \delta_X. \quad (1.36)$$

The term δ_X quantifies *probabilistically* the discrepancy between the observation and reality²⁰. On the other hand, we want to recover the same observation U_r from a simulation code approximating model \mathcal{M}^X . We consequently also have

$$U_r = U_{\mathcal{M}^X} + \delta_{\mathcal{M}^X}. \quad (1.37)$$

¹⁹Many thanks to Marc Sancandi for his help to build this simple but relevant example.

²⁰Note that U_r is not necessarily a random variable. For example, U_{exp} and δ_X can be random variables but such that their sum is deterministic, equals to U_r : it is like having random variables X and $Y = 1 - X$ such that $X + Y$ is deterministic with $X + Y = 1$.

The term $\delta_{\mathcal{M}^X}$ characterises a flaw in the model²¹ which, we hope, is small.

Besides, we never have access to $U_{\mathcal{M}^X}$ but rather to $U_{\mathcal{M}_\Delta^X}$ such that $U_{\mathcal{M}^X} = U_{\mathcal{M}_\Delta^X} + \delta_\Delta$ where δ_Δ quantifies a numerical error²². In other words, (1.37) is equal to

$$U_r = U_{\mathcal{M}_\Delta^X} + \delta_\Delta + \delta_{\mathcal{M}^X}. \quad (1.38)$$

When comparing experimental results and simulation ones, we typically subtract (1.36) to (1.38) to get

$$U_{\text{exp}} - U_{\mathcal{M}_\Delta^X} = \underbrace{\delta_X - \delta_{\mathcal{M}^X} - \delta_\Delta}_{\delta_0}. \quad (1.39)$$

With the introduction of δ_0 , we want to put forward the fact that the discrepancy between the approximated model and the observations is a sum of three terms. To validate²³ or invalidate hypothesis (1.33), we consequently need to be able to extract $\delta_{\mathcal{M}^X}$ from δ_0 . This step can be very tricky. Indeed, there may exist some misleading situations for which

$$\delta_0 \approx 0 = \delta_\Delta + \delta_{\mathcal{M}^X} - \delta_X \text{ but } \delta_{\mathcal{M}^X} = \delta_X - \delta_\Delta \neq 0!$$

Such situation can lead to believe a model is relevant whereas *it is not*, the experiment and/or the numerical errors cancelling the model discrepancy. We clearly want to avoid such kind of situations. In the following, we present how numerical analysis and uncertainty analysis represent ways to avoid such error compensations.

Expression (1.39) has three unknowns for only one equation. In a sense, *numerical analysis* and *uncertainty analysis* are ways to provide the two remaining equations mandatory to eliminate unknowns δ_Δ and δ_X in (1.39).

Let us begin with δ_Δ : ideally, the numerical scheme to approximate \mathcal{M}^X converges so that for Δ small enough $\delta_\Delta = \mathcal{O}(\Delta^\zeta)$ with $\zeta < 0$. The exponent ζ is commonly called the order of the converging scheme, see [154, 74]. It is prescribed by the scheme definition and *must not be evaluated*. The previous notation is equivalent to

$$\text{for } \Delta \sim 0, \exists C \in \mathbb{R} \text{ such that } |C| < \infty \text{ and } \mathcal{O}(\Delta^\zeta) = C\Delta^\zeta. \quad (1.40)$$

When Δ is small enough to ensure $\delta_\Delta = C\Delta^\zeta$, the approximation is said to be in the asymptotic regime. In this case, this implies C can be estimated and δ_Δ quantified. One can theoretically find Δ such that δ_Δ is under a certain threshold and such that it can be considered negligible in comparison to the two other terms

$$\delta_\Delta = \mathcal{O}(\Delta^\zeta) = C\Delta^\zeta \ll \min(|\delta_{\mathcal{M}^X}|, |\delta_X|). \quad (1.41)$$

In order to make sure (1.41) occurs, one must play with Δ . Decreasing Δ comes with higher computational costs. It can occur that the choice of parameters Δ ensuring (1.41) leads to too costly simulations. In order to make sure (1.41) occurs with equivalent restitution times²⁴, there are several levers, commonly used in the literature, within my publications and throughout this document:

1. the first lever is High Performance Computing (HPC). Suppose one has access to a converging numerical scheme but Δ can not be made small enough to get (1.41) (for example because the computations are too time consuming or there is not enough memory available etc.). One first solution would be to get a new more powerful (faster or with more computational units) computer. It is supposed to ensure, with the same solver, a better accuracy by taking smaller Δ with, hopefully,

²¹For example, it stands for a missing operator in the set of PDEs: can we use Euler system as a model or shall we use Navier-Stokes equations (part II)? A diffusion equation or a transport one (part III)? Shall we neglect the effect of external forces F on our particles (part III)?

²²Note that δ_Δ can be deterministic (if from a deterministic scheme) or stochastic (if from a stochastic scheme, see the previous section).

²³Validating hypothesis (1.33) would imply verifying that $\delta_{\mathcal{M}^X} \approx 0$.

²⁴If you are not patient enough to wait for more powerful computers.

reasonable restitution times. For those not patient enough to wait for the next generation of machine, better scalable parallel strategies can be studied and implemented. In [99] for example, the strong and weak scalabilities of several parallel strategies for a Monte-Carlo scheme for neutronics are studied. The present document is also strewn with HPC discussions.

2. The second lever is the order ζ of the scheme. Relying on high-order schemes [154, 74], i.e. such that $\zeta < 1$ or even $\zeta \ll 1$, ensures improving the quality of the solution for constant Δ and with the same computer. Of course, this strategy implies developing new accurate and converging schemes together with their practical verification. Publications [236, 232, 243] intensively rely on such strategy.
3. A third possibility consists in working on the constant C : reducing it also allows obtaining a better accuracy for a constant grid Δ and with the same computer. It is often called an *acceleration technic* or a *variance reductions method* in the context of a Monte-Carlo resolution. This solution is applied more frequently than one could think: for example, in section 5.3.3 and [240], we show that Kriging can be understood as a method aiming at *reducing the constant* multiplying the convergence rate of a Lagrange interpolation. The design of *asymptotic-preserving* schemes also enters this same category and are tackled in my publications [243, 3] and in this document (see remark 9.1, sections 7 and 9.12 and chapter 10).

Of course, there are some intrications between the three above points. The design of numerical methods already integrates parallel considerations. The design of new architectures takes into account the type of calculations needed etc. With the above lines, we briefly presented the main elements of solution to make sure we can eliminate δ_Δ in equation (1.39).

Once (1.41) ensured, (1.39) becomes

$$U_{\text{exp}} - U_{\mathcal{M}_\Delta^X} = \delta_0 = \underbrace{\delta_\Delta}_{\mathcal{O}(\Delta^\zeta) \ll 1} + \delta_{\mathcal{M}^X} - \delta_X = \delta_{\mathcal{M}^X} - \delta_X. \quad (1.42)$$

It only remains to be able to differentiate the model discrepancy $\delta_{\mathcal{M}^X}$ from the noisy term δ_X . This is where uncertainty analysis plays a role. In order to isolate $\delta_{\mathcal{M}^X}$, one can

1. either try to reduce $\delta_X \ll 1$ so that only remains $\delta_{\mathcal{M}^X}$.
2. Or, if δ_X is not easily reducible, try to characterise δ_X in order to be able to eliminate it from δ_0 .

The first solution may imply relying on several (N_{exp}) independent experimental results (U_{exp}^i) $_{i \in \{1, \dots, N_{\text{exp}}\}}$ of noises (δ_X^i) $_{i \in \{1, \dots, N_{\text{exp}}\}}$. The Guide for the expression of Uncertainty in Measurement (GUM, see [112]) provides recommendations to characterise statistically each realisation (δ_X^i) $_{i \in \{1, \dots, N_{\text{exp}}\}}$. We consequently have

$$U_r - U_{\text{exp}} = U_r - \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} U_{\text{exp}}^i = \delta_X = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} \delta_X^i \sim \mathcal{L}(0, \sigma_{\text{exp}}). \quad (1.43)$$

The above expression puts forward the fact that δ_X follows an unidentified centered²⁵ distribution $\mathcal{L}(0, \sigma_{\text{exp}})$ of variance σ_{exp}^2 . If furthermore the experiments are independent and $N_{\text{exp}} \rightarrow \infty$, the central limit theorem²⁶ ensures

$$U_r - U_{\text{exp}} = U_r - \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} U_{\text{exp}}^i = \delta_X = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} \delta_X^i \sim \mathcal{L}(0, \sigma_{\text{exp}}) \xrightarrow{N_{\text{exp}} \rightarrow \infty} \mathcal{G}\left(0, \frac{\sigma_{\text{exp}}}{\sqrt{N_{\text{exp}}}}\right). \quad (1.44)$$

In (1.44), the noise δ_X is asymptotically gaussian²⁷. By *asymptotically*, we mean that the error will be gaussian *only if enough experiments are carried out*. Now, in order to reduce δ_X , one can

²⁵The mean is zero.

²⁶See [256], but it will be stated in the document.

²⁷The term $\mathcal{G}(0, 1)$ denotes a gaussian random variable of zero mean and variance one.

- either increase the number of experiments N_{exp} ,
- or try to reduce σ_{exp} .

Increasing N_{exp} is simple in practice but may be costly. As a consequence, δ_X of variance σ_{exp}^2 may not be gaussian. Gaussian or not, an alternative to reduce δ_X is to work on σ_{exp}^2 . Variance σ_{exp}^2 of δ_X is closely related to the performances of the detectors used during the experiments together with the more or less accurate characterisation of the conditions (modeled by X) of the experimental setting. The reduction of σ_{exp} typically comes with the identification of the components of vector X inducing the most important fluctuations of the output: this is commonly called *performing a sensitivity analysis*. It allows hierachising the main contributors of δ_X amongst the components of X . Assuming X has d independent components $X = (X_1, \dots, X_d)^t$, then σ_{exp}^2 can be decomposed [266] into

$$\sigma_{\text{exp}}^2 = \sum_{i=1}^d \mathbb{V}_i + \sum_{i,j=1}^d \mathbb{V}_{i,j} + \dots + \mathbb{V}_{1,2,\dots,d} = \sum_{s \in \mathcal{S}} \mathbb{V}_s. \quad (1.45)$$

In the above expression, $\mathcal{S} = \{\{1\}, \{2\}, \dots, \{d\}, \{1, 1\}, \{1, 2\}, \dots, \{1, 2, \dots, d\}\}$ is the set of every $2d$ combinations of variables. In (1.45), the $(\mathbb{V}_s)_{s \in \mathcal{S}}$ are the relative variances so that $(\mathbb{S}_s = \frac{\mathbb{V}_s}{\sigma_{\text{exp}}^2})_{s \in \mathcal{S}}$, the Sobol indices [266], express the percentage of variance explained by the set of variable $s \in \mathcal{S}$. Once the set of variables S having the biggest indices identified, i.e. such that $\sigma_{\text{exp}}^2 \approx \sum_{s \in S} \mathbb{V}_s$, one can decide, for example, to invest on ways to reduce the fluctuations of $(X_s)_{s \in S}$, the relevant components of X . This will lead to a reduction of σ_{exp} . In [138] and [241] for example, Sobol's indices for sensitivity analysis are estimated for (respectively) an aerothermal model and the linear Boltzmann equation.

Either ways, reducing σ_{exp} or increasing N_{exp} , aims at making sure $\delta_X \ll 1$ so that (1.42) becomes

$$U_{\text{exp}} - U_{\mathcal{M}_\Delta^X} = \delta_0 = \underbrace{\delta_\Delta}_{\mathcal{O}(\Delta^\epsilon) \ll 1} + \delta_{\mathcal{M}^X} - \delta_X = \delta_{\mathcal{M}^X} - \underbrace{\delta_X}_{\ll 1} = \delta_{\mathcal{M}^X}. \quad (1.46)$$

The modeling error $\delta_{\mathcal{M}^X}$ in δ_0 of (1.39) has been isolated, extracted, quantified. It is a deterministic quantity. We are now able to *decide whether it is small enough or whether some modeling efforts remain to be done*.

The above methodology is by far the most simple and efficient. But it also corresponds to the ideal case. Sometimes, the last analysis is not enough as δ_X can not be made arbitrary small²⁸ and we remain with

$$U_{\text{exp}} - U_{\mathcal{M}_\Delta^X} = \delta_0 = \underbrace{\delta_\Delta}_{\mathcal{O}(\Delta^\epsilon) \ll 1} + \delta_{\mathcal{M}^X} - \underbrace{\delta_X}_{\ll 1} = \delta_{\mathcal{M}^X} - \delta_X. \quad (1.47)$$

It is certainly in this situation that uncertainty quantification is the most relevant. It provides some rigorous elements of solution to answer the question of the relevance of hypothesis (1.33). Let us go through the main principles of the methodology:

- first, let us assume δ_X is not negligible but is probabilistically characterised:

$$U_r - U_{\text{exp}} = \delta_X \sim d\mathcal{P}_{\delta_X}. \quad (1.48)$$

Probability measure $d\mathcal{P}_{\delta_X}$ can, for example, be obtained applying the guidelines described in the GUM [112].

- The idea now is to propagate the fluctuations induced by the uncertain experimental setting X through model \mathcal{M}^X and to characterise probabilistically $U_{\mathcal{M}_\Delta^X}(X)$. Suppose $U_{\mathcal{M}_\Delta^X}$ denotes the mean of random variable $U_{\mathcal{M}_\Delta^X}(X)$. Then the latter can be decomposed into $U_{\mathcal{M}_\Delta^X}(X) = U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X$. The characterisation of $U_{\mathcal{M}_\Delta^X}(X)$ consequently resumes to finding the probability measure $d\mathcal{P}_{\tilde{\delta}_X}$

²⁸For example, we are not able to perform enough experiments. Or σ_{exp} remains important and reducing it would be too costly or too complex or technologically out of reach.

of $\tilde{\delta}_X$. The characterisation of $\tilde{\delta}_X$ can be done *via* an uncertainty propagation of input X through model \mathcal{M}_{Δ}^X . It can be obtained from a simulation code. Part II sums up my work on uncertainty propagation [236, 232, 84, 237, 242, 238, 241]. It is dedicated to the design and the analysis of mathematical and numerical methods to estimate probability measures such as $d\mathcal{P}_{\tilde{\delta}_X}$ and to be able to apply the described methodology. Once the propagation step performed, we have access to

$$U_r - U_{\mathcal{M}_{\Delta}^X} = \tilde{\delta}_X \sim d\mathcal{P}_{\tilde{\delta}_X}. \quad (1.49)$$

This new probability measure will be central to quantify the likelihood of hypothesis (1.33).

Suppose the uncertainty propagation performed (see part II for the how-to question). We now have two characterized distributions:

$$\begin{cases} U_r - U_{\text{exp}} &= \delta_X \sim d\mathcal{P}_{\delta_X}, \\ U_r - U_{\mathcal{M}_{\Delta}^X} &= \tilde{\delta}_X \sim d\mathcal{P}_{\tilde{\delta}_X}. \end{cases} \quad (1.50)$$

It is now easy building a new random variable $\tilde{\delta}_{\mathcal{M}^X} = U_{\text{exp}} - U_{\mathcal{M}_{\Delta}^X} = \tilde{\delta}_X - \delta_X$ which will become our *decision variable for statistical hypothesis testing* [256, 1, 250]. This is detailed in appendix B. Statistical Hypothesis Testing is a key methodology for anyone willing to compare two random variables obtained from two systems. Its main principles are recalled in appendix B with an illustration of how it can be applied in a V&V context. Based on $d\mathcal{P}_{\delta_X}$ and $d\mathcal{P}_{\tilde{\delta}_X}$, the framework allows quantifying the probability of validating falsely hypothesis (1.33). Based on this probability, we are able to *decide whether accepting hypothesis (1.33) is risky or not*.

1.3 Few words on the content and style of this document

With the above material, we aimed at giving a hint at what can be encountered in the two parts of the document. In the next sections, we detail the content of both parts, insist on the style and the presentation choices and discuss some notation tricks which hold all along the manuscript.

1.3.1 Content and ...

Let us begin by the content. As already discussed before, this document has two main parts.

Part II is dedicated to uncertainty quantification for systems of conservation laws. Uncertainty quantification can be applied *intrusively* (construction of reduced models as in the P_n, M_n examples) or *non-intrusively* (use of a black-box code) and discussion on the most relevant strategy remains an open question²⁹. I have research contributions for both, [231, 236, 239, 232, 237, 84, 241] and [238, 243, 31, 242, 233, 234, 240] respectively. In part II, I first present a toy problem involving the resolution of the uncertain Euler equations (chapter 2). It allows illustrating what is expected from an uncertainty analysis in a simplified but still challenging configuration. The problem is a *fil rouge* in the sense every presented resolution strategies, intrusive in chapter 4 and non-intrusive from chapters 5 to 8, are tested on this configuration in the same conditions. Chapter 3 corresponds to an illustrated state-of-the-art of Polynomial Chaos and its derivations from which are based the aforementioned intrusive and non-intrusive methods. The considerations of the previous sections 1.1.1 and 1.2.3 give a good hint at the questions arising in chapters 4 to 8. In part II, the uncertain system of conservation laws depend on one more (random) variable X (i.e. $U(\mathbf{x}, t, X)$ instead of $U(\mathbf{x}, t)$) and we build systems of moments intrusively (P_n, M_n -like approximations) or non-intrusively (S_n -like models) with respect to this variable X instead of \mathbf{v} . Chapters 7 and 8 are slightly different, in the sense they go beyond the *fil rouge* application.

Part III is devoted to the construction, the description and the implementation of Monte-Carlo schemes for the resolution of the linear and nonlinear Boltzmann equation. The linear Boltzmann equation is treated in chapter 9. We try to make the discussion the more complete and progressive possible. We first build the most classical MC schemes and explain how they can be compared. The

²⁹and may be compared to discussions about choosing a deterministic or a stochastic resolution scheme! Or which one of Emacs or Vi is the best...

Hilbert developments introduced in the previous section are central to understand the subtleties of the MC schemes. The *implementation* aspect takes an important place in chapter 9: it is strewn with algorithm descriptions, enriched as we take into account more and more physics (source term, acceleration etc.). The newness and the originality of the first chapter of part III mainly comes from the presentation choice (which is discussed in the next section). The new results are mainly gathered in the last sections (from 9.8 to 9.12) of chapter 9 and in chapter 10. Care has been made to make the material of these sections complementary to the published papers [241, 3, 99]. In the latter, Asymptotic Preserving MC schemes are built for two physical applications of interest implying the resolution of the nonlinear Boltzmann equation (i.e. coupled to an additional equation or set of equations).

1.3.2 ... Style

In this short section, we would like to justify the presentation choices in this manuscript: this work gathers two different disciplines. We are aware a reader may want to focus on one topic but not the other. For this reason the two parts are relatively independent. The reader interested in uncertainty quantification can easily go through the introduction and go on with part II and the reader interested in the Monte-Carlo resolution of the linear Boltzmann equation can skip part II to go directly to part III. The parts are not too strongly intertwined even if few references from one part to the other had to be done to avoid redundancies and insist on analogies³⁰. The same effort has been made to separate, within part II, the intrusive methods from the non-intrusive ones as it is commonly known and accepted in the uncertainty quantification community that some authors are hermetic to one or the other...

The styles in part II and part III are also different. This is because we think the needs in the two disciplines are different. Part II is more illustrative than part III for example. Part II describes a relatively new field and we think that an example-driven discussion helps the readability. On another hand, the Monte-Carlo resolution of the Boltzmann equation can be considered a more classical one and an original way to describe it is mandatory to avoid redundancies with the furnished literature. For this reason, part III is progressive, original (I hope) and is more implementation-driven. It aims at overcoming one difficulty I personally encountered when I started developing a Monte-Carlo method in a simulation platform: bridging the gap between the description of the MC solver and its implementation. Because where uncertainty analysts see a *black-box* method in MC resolutions, numerical analysts familiar with deterministic schemes see *black magic*³¹. Whether the part achieves the purpose of demystifying MC methods or not now depends on the reader's opinion.

In both parts, depending on the familiarity of the reader with the topic, some discussions might appear naive and simple. Those discussions are *flagged as such* and the familiar reader is invited to skip them. They are addressed to, and have been motivated by questions from, students, interns, PhD students, summer school ones or colleagues. For this reason, I also hope the document will be adapted and will benefit future interns and PhD students. Of course, this is not the main aim of the document. The first one is to present my research contributions to these two fields of applications and explain in which sense they are articulated, correlated and part of a more ambitious research project. For this, I try to make the manuscript and my publications the more complementary possible and avoid redundancies. In many aspects, I hope this document appears more mature than some of my papers. The document in general deals with very simple examples and configurations and more complex ones are tackled in my publications. The two parts of the manuscript also contain some unpublished (or not yet published, see for example chapters 5 and 8 or section 10.2). Sometimes this is due to a lack of time, sometimes because we think it did not deserve a publication but remains illustrative enough to be in this document.

1.3.3 Few words on the notations and the presentation tricks

To end this introduction, we would like to comment on some notations and presentation tricks we hope will ease the readability of the document. All along the document, we try to keep as much as possible

³⁰The reader interested in the resolution of the uncertain linear Boltzmann equation may want to go through section 9.11 in which a Monte-Carlo scheme to solve the stochastic Boltzmann equation is built.

³¹Joke made by Christian Aussourd to who I say hello!

the same notations for the same quantities. At least for the ones we consider the most important. For example,

- I intensively use footnotes. This can be quite disconcerting for some readers. In this document, the rule applied when introducing footnotes is the following: it is supposed to help the reader by introducing additional details which I consider of a lesser³² importance in the text. In other words, the reader confused by the use of footnote is encouraged to go on with the sentence as if it was not there.
- Notation u or U always denotes the main quantity, the unknown³³ we aim at approximating all along the document. This may lead to unconventional notations: u denotes a neutron density in section 10.1 for example whereas n is commonly used in the literature for the same quantity. We think this is part of a presentation trick to help the reader focus on the unknown of interest.
- My publications are in red in the document, cf. [236] for example. This trick has already been used in the previous sections.
- This document mainly discusses numerical approximation methods for *physical applications* one wants to study given a *simulation architecture*. In this context, applied mathematics are only a mean to achieve the physical purpose with the computing device at hand. High Performance Computing (HPC) is consequently as important as the physical objective. For this reason, wherever it is relevant, HPC considerations are tackled, mainly at the end of chapters or in summaries of sections.
- Notation x or \mathbf{x} always denotes a spatial variable in the spatial domain \mathcal{D} .
- Notation t denotes the time in $[0, T]$ where T denotes the final time of the simulation/numerical experiment.
- Notation \mathbf{v} or $v = |\mathbf{v}|$ always denotes the velocity variable and its norm.
- When a spatial grid is introduced, it is denoted by $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$ where the \mathcal{D}_i are N_x non-overlapping cells.
- In general X denotes a random variable or a random vector and its probability measure is denoted by $d\mathcal{P}_X$.
- Variable δ is always the small parameter in a Hilbert development.
- The more possible, every dependences of the unknowns and quantities are recalled. When they are not, we explicitly explain why we drop them.
- In the document, to characterise a random variable, we prefer dealing with probability measures instead of probability density functions. They are more general and convenient in the sense they allow dealing with discrete random variables with the same notations. Besides, we sometimes have some parameterised laws, in the sense we have some quantities which are probability measure with respect to one variable, independently of all the others: for example, by convention in this document, if $u(\mathbf{x}, t, \mathbf{v})d\mathbf{v}$ is a probability measure, then it is $\forall \mathbf{x}, t$ with respect to variable \mathbf{v} .

But despite all these efforts, the document still has some slightly abusive, but convenient, notations. They generally concern secondary quantities. To give some examples:

- in general, we consider probability measure so by convention in this document every measures sum up to 1: we consequently write $\int d\omega = 1$. The notation is not conventional but is conciser and replace both $\int \mathbf{1}_{[-1,1]}(\omega) \frac{1}{2} d\omega$ in 1D or $\int \mathbf{1}_{S^2}(\omega) \frac{1}{4\pi} d\omega$ in 2D.
- Sometimes, we also use the notation $[0, t = \Delta t]$ which means Δt and t denotes the same quantities, the end of the time step, and are both used in the chapter or the section.

³²but not of none, otherwise it would not be in the text at all.

³³ u or U as unknown.

- We also insist that in the following parts, $u(\mathbf{x}, t, \mathbf{v})$ and $u(\mathbf{x}, t)$ denote two different functions. In general, the second one is equal to the first one integrated over the velocities, i.e. $u(\mathbf{x}, t) = \int u(\mathbf{x}, t, \mathbf{v}) d\mathbf{v}$. We think this does not alter the readability of the document as care has been taken to recall, the more possible, every dependences of every quantities.

With these few lines and precisions, I hope the document will ... Be useful...

Part II

Uncertainty quantification for hyperbolic systems of conservation laws

Chapter 2

Physical Motivations and toy problem (*fil rouge*)

A ‘fil rouge’ problem to compare intrusive and non-intrusive methods for uncertainty quantification

In this chapter, we focus on limit (1.16) of the Boltzmann equation (1.1). More generally, the material of the described work can be applied to any system of conservation laws. It corresponds to a particular form of partial differential equations (PDE), whose general structure (already presented in chapter 1) is

$$\partial_t U(\mathbf{x}, t) + \partial_x f(U(\mathbf{x}, t)) = 0. \quad (2.1)$$

In (2.1), we have

- $(\mathbf{x}, t) \in \mathcal{D} \subset \mathbb{R}^3 \times [0, T]$,
- $U : \mathcal{D} \times [0, T] \longrightarrow \mathcal{D}_U \subset \mathbb{R}^d$, is the vector of unknowns,
- $f : U \in \mathbb{R}^d \longrightarrow f(U) \in \mathbb{R}^d$ is called the flux.

We nonetheless focus on the Euler system in 1D (one spatial dimension, i.e. $\mathbf{x} = x \in \mathcal{D} \subset \mathbb{R}$ here), modeling compressible gas dynamics, which corresponds to a particular choice of U and of the flux $f(U)$. It is given by the following coupled equations

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, \\ \partial_t (\rho u) + \partial_x (\rho u^2 + p) = 0, \\ \partial_t (\rho e) + \partial_x (\rho ue + pu) = 0. \end{cases} \quad (2.2)$$

For this *fil rouge* problem, we consider a perfect gas closure $p = (\gamma - 1) \left(\rho e - \rho \frac{u^2}{2} \right)$. This simplifies the discussions but the material of the next sections is not limited to such equation of states. System (2.2) must be completed with proper initial and boundary conditions, they will be dealt with later. With the above notations,

$$U(x, t) = (\rho(x, t), \rho(x, t)u(x, t), \rho(x, t)e(x, t))^t,$$

is the vector of conservative variables: ρ is the mass density of a fluid, u its velocity so that ρu is its momentum, and e is its specific total energy, so that ρe is its total energy. The variable p is the pressure of the fluid, ensuring the closure of the above system.

Let us first spend some time describing the deterministic behaviour of such system in a particular configuration of interest. It corresponds to a Sod shock tube (also called a Riemann problem). The initial condition consists in two states¹: the left one corresponds to a ‘heavy’ fluid ($\rho_L^0 = 1, p_L^0 = 1$) and the right one, to a ‘light’ fluid ($\rho_R^0 = 0.125 < \rho_L^0, p_R^0 = 0.1 < p_L^0$). Both states are separated by an interface at $x_{int} = 0.5 \in \mathcal{D} = [0, 1]$ where \mathcal{D} denotes the simulation domain. The fluids are initially

¹the superscript 0 stands for $t = 0$.

at rest ($u^0 = u_L^0 = u_R^0 = 0$). The boundary conditions are neutral ones (zero gradient) on both sides of $\mathcal{D} = [0, 1]$ but the simulation is stopped before the solution interacts with them. As soon as $t > 0$, the solution develops discontinuous behaviours²: in figure 2.1 (top right), the mass density at $t = 0.14$

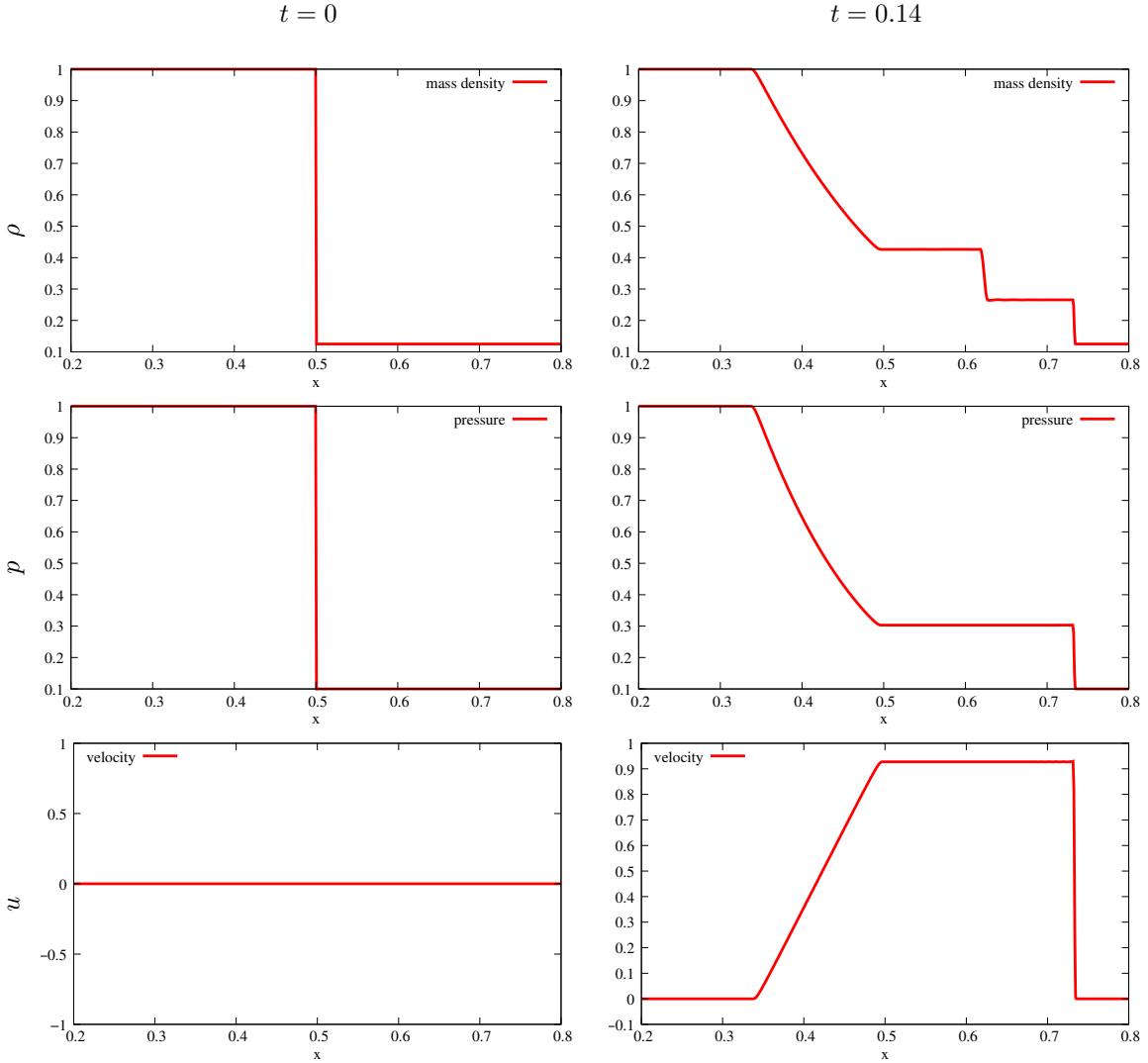


Figure 2.1: Left: Initial conditions of the Sod shock tube for ρ, p, u with respect to the spatial variable x . Right: solutions at time $t = 0.14$ for ρ, p, u with respect to the spatial variable x .

presents four constant states separated by three waves. The first wave is a smooth rarefaction fan in the heavy fluid. The second one is called either a contact discontinuity or an interface (characteristic of the initial discontinuity between the two initial states). The last one is a shock in the light fluid. The pressure and the velocity are smooth at the interface so that they only have three constant states, separated by a smooth wave: the rarefaction fan in the heavy fluid, and a discontinuous one, the shock in the light fluid.

We now suggest building a simple uncertainty quantification problem from the previous configuration. It will be solved thanks to a Monte-Carlo method. The latter remains the reference method in uncertainty quantification mainly due to its simplicity of application. We will then comment on the probabilistic properties of the above spatial profiles at time $t = 0.14$. We suppose the solution no longer only depends on (x, t) but also explicitly on a random vector $X \in \Omega \subset \mathbb{R}^Q$, where $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space

²This is proper to systems of conservation laws, not only the Euler one. Discontinuities can appear dynamically, see [81, 260, 81, 261].

[256]. The uncertainty is consequently here modeled thanks to probability theory. Alternatives exist, as possibility theory for example [86, 88, 87], but are beyond the scope of this document. The previous system rewrites

$$\partial_t U(x, t, X) + \partial_x f(U(x, t, X)) = 0,$$

where

- $(x, t, X) \in \mathcal{D} \subset \mathbb{R} \times [0, T] \times \Omega$,
- $U : \mathcal{D} \times [0, T] \times \Omega \rightarrow \mathcal{D}_U \subset \mathbb{R}^d$, is the vector of unknowns,
- $f : U \in \mathbb{R}^d \rightarrow f(U) \in \mathbb{R}^d$ is called the flux.

For Euler system, this leads to

$$U(x, t, X) = (\rho(x, t, X), \rho(x, t, X)u(x, t, X), \rho(x, t, X)e(x, t, X))^t.$$

The vector of unknowns U belonging to a probability space, the solution of the uncertainty quantification problem is a stochastic process, i.e. random variables parameterized by both x and t ³. In this sense, solving an uncertainty quantification problem corresponds to the resolution of stochastic partial differential equations (SPDE). This is emphasized in the simple example of appendix A.

In our example, the initial uncertainty is modeled by a uniform random variable $X \sim \mathcal{U}[-1, 1]$ and affects the position of the interface $x_{int}(X) = 0.5 + 0.05X$ between the two initial states. The interface position is consequently uniformly distributed in the interval [0.45, 0.55]. Figure 2.2 (top-left) shows the initial conditions for three realisations (the extremal ones $X = 0.45$ and $X = 0.55$ and the mean one $X = 0$) of the interface position. The top right picture of figure 2.2 presents the same three spatial profiles at $t = 0.14$ after applying three deterministic resolutions (runs of a simulation code as a black-box). The profiles are very close to the one obtained for $X = 0$, they only correspond to a translation on the x -axis. Still, the wave positions are affected. On the top pictures of figure 2.2, we only presented the results obtained for three chosen realisations of the interface position. In order to obtain a reference solution for this uncertainty quantification problem, we applied an MC resolution with $N_{MC} = 1000$, where N_{MC} is the number of realisations of X and of resolutions of (2.2).

Regarding the numerical resolution, the attentive reader would have probably noticed the results of figure 2.1 are analytical. On another hand, the results of figure 2.2 are numerical (see the little imperfection in the vicinity of the interfaces on figure 2.2 top-right commonly called *wall heating*). In this section, we do not aim at being exhaustive on the description the numerical scheme at use. But we insist it is of importance and must have relevant properties relative to the uncertain configuration of interest for efficiency⁴. Here, the numerical scheme needs to capture accurately both the constant states and their wave positions (chronometry): for this reason we used a 3rd order Lagrange+remap scheme, conservative, shock capturing and accurate. Its characteristics and properties are detailed in [154, 101], [232] for example.

Once the N_{MC} resolutions performed, it remains to postprocess the classical probabilistic quantities (mean, variance, moments, quantiles,...) of interest for the different variables of the system. Figure 2.2 (bottom) presents the spatial profiles of the mean and variance at $t = 0$ and $t = 0.14$ for the mass density. At $t = 0$, the mean of the mass density is smooth and linear between the two states. The linear form of the mean of the mass density between the two initial states is closely related to the distribution of the random variable modeling the initial uncertainty X : a gaussian random variable X for example would have induced an even more regular form of this same mean (it is here \mathcal{C}^0 but not \mathcal{C}^1). Note that the smoothness of the spatial profiles for $t > 0$ for the mean of the mass density has been proven in [264] for the same kind of configuration. Concerning the choice of the input random variable, we want to insist on the fact that it greatly impacts the solutions at later times. We choose X as an uniform variable but the material of this manuscript allows taking into account arbitrary random variables which are, in practice, dictated by the physics at stake.

³For example, for fixed x, t , $\rho(x, t, X)$ is a random variable.

⁴An illustration of this affirmation is given in section 9.11.1 of chapter 9.

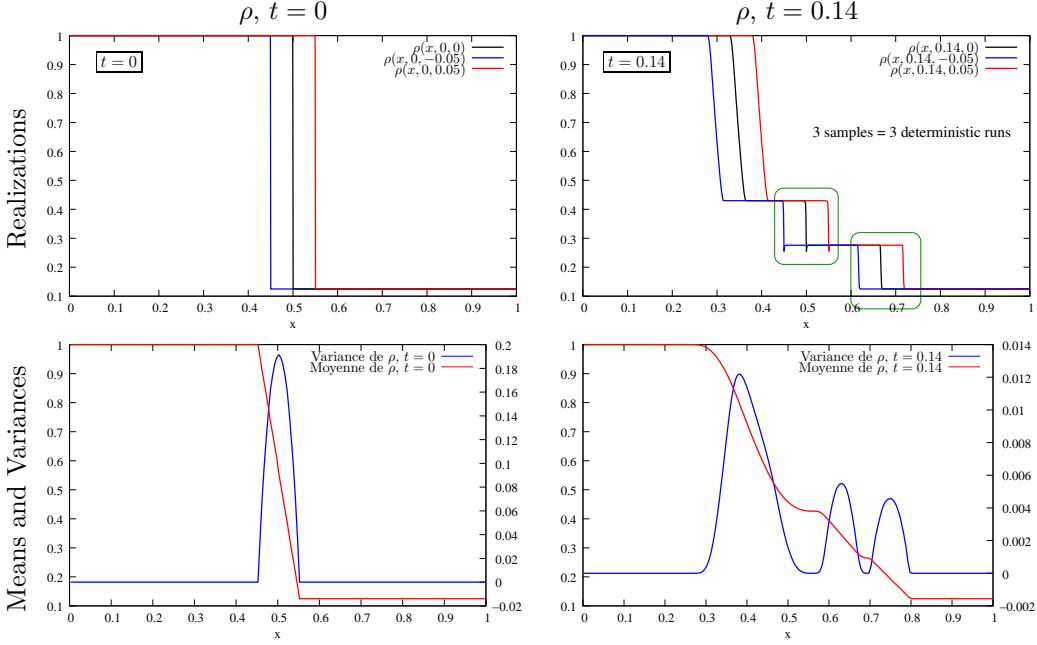


Figure 2.2: Top left: initial conditions for three realisations of the uncertain interface position for the mass density. Bottom left: initial conditions in term of mean and variance for the mass density obtained with 1000 realisations of the random variable X . Top right: three realisations of the mass density at time $t = 0.14$. Bottom right: mean and variance of the mass density at $t = 0.14$ obtained with 1000 realisations of X .

Let us comment on the bottom pictures of figure 2.2: for the variance of the mass density, it is easy identifying the spatial area of the domain \mathcal{D} affected by the uncertainty. They correspond to the non-zero areas for the variance. Note that the scale for the variance is on the right hand side. At $t = 0.14$, the bottom right picture of figure 2.2 presents the postprocessed mean and variance of the mass density obtained with the MC method with $N_{MC} = 1000$ realisations of the interface position. First, one can notice that the mean of the mass density is smoother than for one realisation but still display four constant states separated by three identifiable waves. The variance of the mass density is non-zero only in the vicinities of the three waves corresponding to the uncertainties in the rarefaction fan, the interface and the shock. For this particular configuration, the initial uncertainty is shared between the three physical waves at later times.

As tackled before, our aim is to solve a SPDE. Mean and variance are not, in general, sufficient or relevant enough in order to fully characterise a random variable or a stochastic process (see appendix A for an illustration). Until now, we focused on spatial profiles at $t = 0.14$. Let us now focus on $t = 0.14$, spatial locations $x = 0.38$ (rarefaction fan), $x = 0.61$ (interface) and $x = 0.73$ (shock) and the random variables $\rho(x = 0.38, t = 0.14, X)$, $\rho(x = 0.61, t = 0.14, X)$ and $\rho(x = 0.73, t = 0.14, X)$. The MC method allows approaching the probability density functions (pdf) of the three latter random variables thanks to histograms (see appendix A). Figure 2.3 presents the histograms of the pdfs of the latter random variables obtained by postprocessing the $N_{MC} = 1000$ realisations of the MC method. The top picture corresponds to the mass density random variable in the vicinity of the rarefaction fan. It exhibits a *continuous/smooth* behaviour, the support of the histogram being convex (as the one of the initial random variable X was). With such probabilistic observable, one can for example conclude that the state $\rho = 0.85$ has about the same probability of occurrence as the state $\rho = 0.87$. On the contrary, the histograms of the mass density in the vicinities of the interface and of the shock (bottom pictures of figure 2.3) exhibit *discontinuous/discrete* behaviours. The supports of the states are non-convex: in fact, here, it even corresponds to two Dirac masses for the states $\rho \approx 0.27$ and $\rho \approx 0.42$ for the interface and $\rho \approx 0.13$ and $\rho \approx 0.26$ for the shock. In the vicinities of the interface and the shock, the mass densities probabilistically behaves as rigged coins (with one side having twice more probability of occurrence than the other). The configuration of interest, apparently simple, has been chosen to emphasize this important

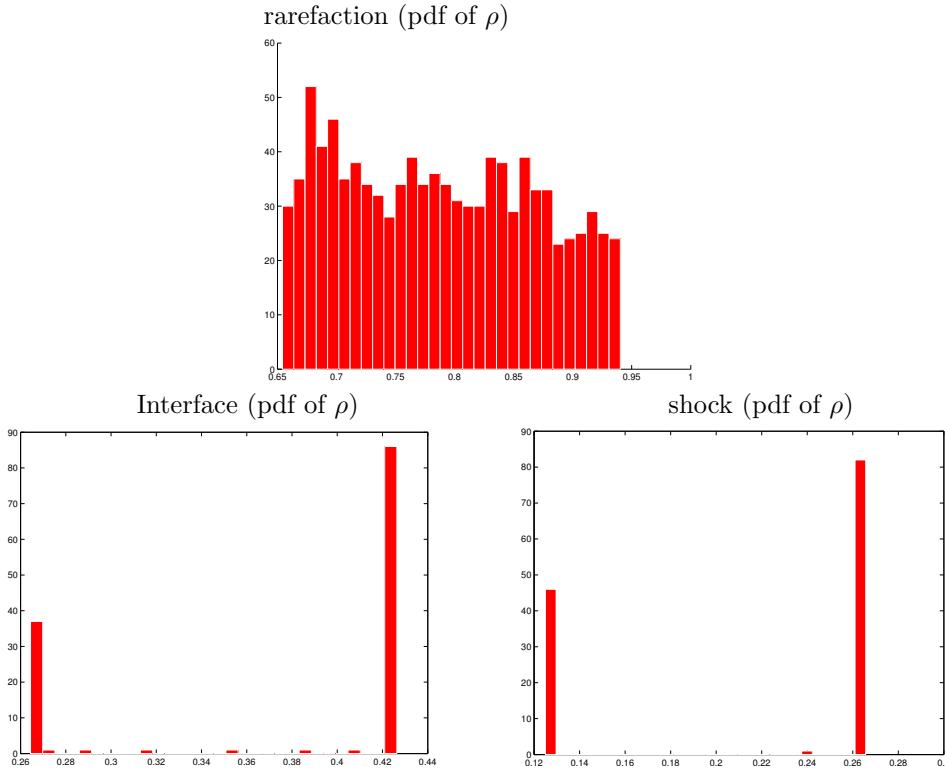


Figure 2.3: Histograms of the pdfs of the random variables $\rho(x = 0.38, t = 0.14, X)$, $\rho(x = 0.61, t = 0.14, X)$ and $\rho(x = 0.73, t = 0.14, X)$ obtained by postprocessing the $N_{MC} = 1000$ MC realisations.

aspect: the initially continuous random variable X is transformed into a discrete one, presenting what is commonly called a threshold effect. A small perturbation of the initial interface position can cause a complete change on some observables such as the mass density in the vicinity of the shock. With such threshold effect, it is obvious that mean and variance are not relevant probabilistic quantities: the mean of the random variable $\mathbb{E}[\rho(x = 0.73, t = 0.14, X)] = 0.24$ has no physical value. The mean value of the random variable 'throw a 6 faced dice' is 3.5 and is never computed. The state 3.5 has a zero probability of occurrence in practice, the mean is not informative, not relevant.

Note that once the uncertainty propagation performed, one can have access to every aspects of an uncertainty analysis, see [271, 272]. It is often only a first step toward sensitivity analysis, robust optimization, probability of failure, resolutions of an inverse problem etc. With this simple example, we wanted to depict the general conditions of the next studies: we aim at taking into account uncertainties in systems of conservation laws. The example is in 1D spatial dimension x and 1D stochastic dimension X but we have in mind applying such analysis to more complex configurations in 3D spatial dimensions for example, for more complex systems of conservation laws (multi-physics) with several sources of uncertainties (stochastic dimension Q).

Regarding the sources of uncertainties, we consider the number of uncertain parameters is not very important $Q \sim 5 - 8$. We suppose a previous study has been performed in order to identify the most sensitive ones for the considered outputs. For an interesting and pedagogical overview of the methods to do so, we refer to [145, 166, 145] but we will not tackle the subject anymore in this document. Our aim is now to perform an accurate uncertainty analysis on a relatively low number of random variables. Concerning the physics of interest, we want to tackle multidimensional multiphysics systems of conservation laws (i.e. developing discontinuous solutions) for which a deterministic resolution can be very expensive in term of computational time. Consequently, an MC method can not be applied, needing an unaffordable number of system resolutions (runs of a simulation code as a black-box). For this reason we will study alternative numerical methods for solving SPDEs. For these, all along our studies, we will have to keep in mind the discontinuous behaviours of the solutions and the threshold effects they

will trigger from a probabilistic point of view. The numerical methods will have to be computationally *efficient, accurate and robust* with respect to such unavoidable difficulties, imposed by the PDEs of interest. As an alternative numerical method to the MC one, in relatively low stochastic dimension, we got interested in a recent one referred as *Polynomial Chaos* in the literature. The next chapters aim at presenting an illustrated state-of-the-art of Polynomial Chaos with a systematic application of the resolution algorithms to the above hydrodynamical 'fil rouge' problem.

Chapter 3

Polynomial Chaos as an alternative to Monte-Carlo methods for UQ

An illustrated state-of-the-art on Polynomial Chaos methods

Contents

3.1 Wiener's Homogeneous Chaos [295] and Cameron-Martin's theorem [55]	30
3.1.1 On Stone-Weierstrass' approximation theorem	31
3.1.2 On Wiener's Homogeneous Chaos [295]	31
3.1.3 On Cameron and Martin's theorem [55]	33
3.2 Polynomial Chaos for uncertainty quantification (UQ)	35
3.2.1 Transformation of a gaussian random variable into a uniform one	36
3.2.2 Mapping of a uniform random variable into an Arcsinus and a Binomial one .	37
3.3 Introduction of generalized Polynomial Chaos (gPC) for UQ	39
3.4 The construction of the gPC basis	41
3.4.1 Inner product defined by an arbitrary probability measure	42
3.4.2 Moments of a probability measure and Hankel determinants	42
3.4.3 Christoffel's formulae, Jacobi's matrix and construction procedures	44
3.4.4 Taking into account discrete/categorical input variables with gPC	46
3.5 Curse of dimensionality and Gibbs phenomenon	46
3.5.1 Curse of dimensionality	46
3.5.2 Sensitivity to the Gibbs phenomenon	47
3.6 Summary for generalized Polynomial Chaos	48

In this chapter we suggest an illustrated state-of-the-art of Polynomial Chaos for uncertainty quantification. We aim at describing its pros and cons having in mind the issues triggered by the hydrodynamical problem of chapter 2. The material of this chapter was published (in a much less detailed version) in some vulgarization journals [234, 233].

It is commonly accepted that at the basis of Polynomial Chaos for uncertainty quantification, two main references are unavoidable, [295] and [55]. The term *Polynomial Chaos* was first introduced in the seminal work of Wiener, in "Homogeneous Chaos Theory" [295], in 1939. The paper is rich in many ways: first for its mathematical content and theoretical results. But also through the indirectly tackled fields of application such as probability theory, approximation theory, ergodic theory, homogeneous turbulence and even numerical analysis of hyperbolic systems at its very end. It will be discussed in the next section together with the second one, written in 1957, by Cameron and Martin [55]. In both papers, the authors demonstrate a convergence theorem on an unbounded function space. The statements are complex and

general and it may appear hard understanding all their implications in term of approximation theory and resolution algorithm. In section 3.1, we present a (too) brief analysis of both papers in order to replace them in nowadays' practical context. Our main contributions (except for chapter 4), see chapters 6, 7 and 8, often only consist in new algorithmic interpretations of the work presented in both of these papers.

Regarding uncertainty quantification, the term *Polynomial Chaos* together with its first application were first introduced by Ghanem and Spanos [124]. The approximation method has since then been successfully applied to solve many uncertain problems (stochastic elastic materials [124], finite deformations [6], heat conduction [290], incompressible flows [307, 215, 186], reacting flows and detonation [179]...). However, most of these approaches failed in the case of "complex" flows involving discontinuities with respect to the random variables. Many authors analysed the approximation method, highlighted some of its weaknesses and even suggested practical solutions, see for example [67, 66, 185, 294, 121]¹. There has been active ongoing research on this topic over the last ten years and this part II of the document presents our contributions to the discipline.

In the next sections, we aim at performing an illustrated state-of-the-art of Polynomial Chaos for uncertainty quantification. By illustrated, we mean that care will be taken to analyse, isolate the points/notions of interest and imagine simple test-cases emphasizing them. We want to make the state-of-the-art the more progressive possible and for this, we often introduce difficulties one by one, illustrate them on simple problems and discuss the material. The reader familiar with Polynomial Chaos may find some examples and discussions *naive* and skip this chapter.

The chapter is organized as follow: we first (briefly) discuss Wiener's Homogeneous Chaos [295] and Cameron-Martin's theorem [55]. Section 3.1 is born from notes aiming at investigating the subtleties of both papers. The papers are very rich and kind of hard to penetrate due to their quite disruptive notations (at least for me). We here claim no originality and the section, in any way, does not represent a substitute to the referred works. We hope it may help the beginner with their decryption. The discussion is general and will probably first seem far from the concrete problem of uncertainty quantification. Attempts will be made to shorten the gap between the paper and this field of application. As we are aiming at presenting Polynomial Chaos as a good alternative (under some hypothesis) to Monte-Carlo methods, we want to identify the spectral counterpart of the Central Limit Theorem [256] invoked when applying Monte-Carlo approximation strategies. This will lead us to recall first Stone-Weierstrass' approximation theorem. In the following sections, we focus on uncertainty quantification applications as introduced first by [124]. We insist on Polynomial Chaos (PC) first and on the need for the introduction of *generalized* Polynomial Chaos (gPC) afterward. The step from Polynomial Chaos to generalized Polynomial Chaos is the key of the methodology. It allows a considerable gain in accuracy, hence its efficient application to even more complex physics. In section 3.4, we recall some details about the already intensively studied orthogonal polynomials [273, 5, 117]. They remain the basis of the PC/gPC approximation methods. Our main contributions to the discipline are presented in the next chapters 4–6–7–8, motivated by systems of conservation laws and the 'fil rouge' problem described in chapter 2.

3.1 Wiener's Homogeneous Chaos [295] and Cameron-Martin's theorem [55]

Before tackling Wiener's and Cameron-Martin's publications, we suggest reminding the well-known Stone-Weierstrass theorem. The work described in [295, 55] may be understood as an attempt to generalize the latter to unbounded intervals under prescribed hypothesis. Note that the theorem is stated with *uncertainty quantification-friendly* notations which will hold all along the document. We aim at briefly emphasizing what Wiener, Cameron and Martin added in term of results with respect to Stone and Weierstrass.

¹Note that the list here is non exhaustive but will be completed all along this chapter.

3.1.1 On Stone-Weierstrass' approximation theorem

Stone-Weierstrass' theorem is at the basis of spectral methods and will be considered a reference for the analysis of the theorems invoked for PC/gPC representations. Care will be taken to state the different approximation theorems and properties in similar conditions in order to make comparisons easier. We recall our aim here is to identify the spectral counterpart of the Central Limit Theorem for Monte-Carlo methods.

Theorem 3.1 *Stone-Weierstrass' approximation theorem: suppose $u : X \in \prod_{i=1}^Q [a_i, b_i] \subset \mathbb{R}^Q \rightarrow u(X) \in \mathbb{R}$, $u \in \mathcal{C}^0 \left(\prod_{i=1}^Q [a_i, b_i] \right)$, the set of continuous real-valued function together with its L^∞ -norm*

$$\|u\|_\infty = \sup_{x \in \prod_{i=1}^Q [a_i, b_i]} |u(x)|.$$

Then there exists a sequence of polynomials $(\phi_n)_{n \in \mathbb{N}}$ converging uniformly toward u on $\prod_{i=1}^Q [a_i, b_i]$.

The proof is not detailed here but can easily be found in many books, see [105] for example. It implies the use of Bernstein polynomials and the application of the Central Limit theorem. Both are of importance in this document: particular families of polynomials will be studied in this part II and every Monte-Carlo schemes described in part III will rely on the Central Limit Theorem for convergence. The strength of the theorem comes from two aspects:

- first the convergence is uniform. But designing efficient engineering/numerical algorithms from the L^∞ -norm remains complicated. Less constraining and strong norms may be more convenient.
- Second, the hypothesis of regularity of the solution does not hold for our applications. The fact that discontinuous functions are in $L^2(\prod_{i=1}^Q [a_i, b_i])$ and that polynomials are dense in this space make it look like an interesting trade-off. The L^2 space is also very convenient to build approximation algorithms based on scalar products and projections on a space of both finite and infinite dimension. For these reasons, in this document, the L^2 -norm is mainly considered.
- Note there are ongoing researches on Optimal Control and L^1 -minimization problems, see [85]. They are out of the scope of this document but we will probably study and work on these in the future in order to understand their subtleties.

In [295], a new approximation theorem (*the weak approximation theorem* see [295]) is stated under general conditions but hints toward considering the L^2 space come very naturally along the paper. In [55], the authors directly focus on the L^2 space. In the two following sections, we briefly discuss papers [295] and [55]. Both papers end with a new approximation theorem extending the Stone-Weierstrass one but in order to avoid redundancies, we focus on their particularities: in Wiener's paper, the notion of ergodicity is central. On another end, in [55], the authors state and prove their approximation theorem in a more general form. These two aspects are described in sections 3.1.2 and 3.1.3.

3.1.2 On Wiener's Homogeneous Chaos [295]

The most astonishing parts of Wiener's publication [295] are in my opinion the introduction, untitled *Physical need for theory*, and the conclusion, untitled *The physical problem*. Those parts are short but hint at a wide range of applications, from homogeneous gas and liquids, states of turbulence and finally considerations on solutions of nonlinear systems of conservation laws. The last ones are still hot topics as testifies a recent paper by Tadmor *et al.* [109].

In his paper, the author focuses on homogeneous dynamical systems for which an ergodicity property holds. He emphasizes the need for mathematical tools to represent, understand *to accurately approximate*, solutions of dynamical systems bearing this property. Ergodicity allows translating averages over an infinite range taken with respect to a given measure λ into averages over a set of finite measure. In physical applications, the variable associated to the infinite set λ is usually the time whereas the spatial set (simulation domain \mathcal{D} such that $|\mathcal{D}| < \infty$ for example) corresponds to the finite one. Basically, if this property holds, instead of observing a physical phenomenon during an infinite amount of time, averaging spatially is enough.

The main aim of paper [295] is to *generalize Birkhoff's ergodic theorem* to multi-dimensional measure λ without any identification with the time variable. This implicitly paves the path toward the possibility to consider unstationary problems: solutions at a given time can be obtained from averages over some multi-dimensional measures (λ should not be confined to the time and to any infinite set).

As explained before, the author aims at generalizing the ergodic theorem. Chaoses are the objects on which it applies. To state the ergodic theorem as in [295], we need to introduce several notions and definitions: in the following section, we try to introduce them in more UQ-friendly notations but we must admit it is still a long way to achieve a satisfying enough section. Anyway, we make this attempt. A *homogeneous chaos*, see [295], is defined as a scalar² measurable function u from

$$u : x_1, \dots, x_Q, X \in \mathbb{R}^Q \times [0, 1] \longrightarrow u(x_1, \dots, x_Q; X) \in \mathbb{R}.$$

In the above expression, $X \sim \mathcal{U}([0, 1])$ is such that if $\forall (y_1, \dots, y_Q) \in \mathbb{R}^Q, \forall S \subset \mathbb{R}$

$$\mathbb{E} [\mathbf{1}_S(u(x_1 + y_1, \dots, x_Q + y_Q; X))] < \infty,$$

then $\forall (y_1, \dots, y_Q), (y'_1, \dots, y'_Q) \in \mathbb{R}^Q \times \mathbb{R}^Q$

$$\mathbb{E} [\mathbf{1}_S(u(x_1 + y_1, \dots, x_Q + y_Q; X))] = \mathbb{E} [\mathbf{1}_S(u(x_1 + y'_1, \dots, x_Q + y'_Q; X))].$$

If u is integrable, it is easier and conciser defining *homogeneous chaoses* as a set-function of $\Sigma_1 \times \dots \times \Sigma_Q = \Sigma$ such that $(x_1, \dots, x_Q) \in \Sigma_1 \times \dots \times \Sigma_Q = \Sigma$, i.e.

$$U(\Sigma, X) = \int_{\Sigma} \dots \int u(x_1, \dots, x_Q; X) dx_1 \dots dx_Q.$$

Of course, U reduces to u , i.e. $U(\Sigma, X) = u(x_1, \dots, x_Q; X)$, when the set Σ reduces to the point (x_1, \dots, x_Q) , i.e. $(x_1, \dots, x_Q) = \Sigma$. Now define the set $\Sigma(y_1, \dots, y_Q)$ as such,

$$\text{if } (x_1, \dots, x_Q) \in \Sigma \iff (x_1 + y_1, \dots, x_Q + y_Q) \in \Sigma(y_1, \dots, y_Q).$$

Then in the new notations, the definition of a *homogeneous chaos* becomes: if $\forall (y_1, \dots, y_Q) \in \mathbb{R}^Q, \forall S \subset \mathbb{R}$

$$\mathbb{E} [\mathbf{1}_S(U(\Sigma(y_1, \dots, y_Q), X))] < \infty,$$

then $\forall (y_1, \dots, y_Q), (y'_1, \dots, y'_Q) \in \mathbb{R}^Q \times \mathbb{R}^Q$

$$\mathbb{E} [\mathbf{1}_S(U(\Sigma(y_1, \dots, y_Q), X))] = \mathbb{E} [\mathbf{1}_S(U(\Sigma(y'_1, \dots, y'_Q), X))].$$

Thanks to the previous notions, Wiener's generalization of the ergodic theorem can be stated as below.

Theorem 3.2 *Wiener's generalization of Birkhoff's ergodic theorem: let $U(\Sigma, X)$ be a homogeneous chaos. Let the functional*

$$\Phi(U(\Sigma, X)) = g(X),$$

be a measurable function of X such that

$$\int g(X) \log(|g(X)|) d\mathcal{P}_X < \infty.$$

Then for almost all values of X ,

$$\lim_{r \rightarrow \infty} \frac{1}{V(r)} \int_R \dots \int \Phi(U(\Sigma(y_1, \dots, y_Q), X)) dy_1 \dots dy_Q < \infty.$$

In the above expression, $r^2 = \sum_{i=1}^Q y_i^2$, $R = \{(y_1, \dots, y_Q) \in \mathbb{R}^Q | \sum_{i=1}^Q y_i^2 \leq r^2\}$ and $V(r)$ the volume of

²or vector valued, the generalization is straightforward.

R. If $U(\Sigma, X)$ is metrically transitive (see [295]),

$$\lim_{r \rightarrow \infty} \frac{1}{V(r)} \int_R \dots \int \Phi(U(\Sigma(y_1, \dots, y_Q), X)) dy_1 \dots dy_Q = \int \Phi(U(\Sigma, X)) d\mathcal{P}_X, \quad (3.1)$$

for almost all values of X .

Now, polynomial chaos, which will be defined in the next section, can be understood as a mathematical object such that any function of a polynomial chaos of arbitrary order satisfies the ergodic property by construction.

The latter property of polynomial chaos is scarcely tackled in the literature. I must admit it is still hard for me acknowledging its relative importance in everyday applications. Still, in [243], more developed in chapter 7, an attempt is made to take advantage of it.

3.1.3 On Cameron and Martin's theorem [55]

In this section, we state Cameron-Martin's theorem [55] with notations specific to the manuscript and, we hope, friendlier in an uncertainty quantification context. This may appear of poor interest but the reader unfamiliar with the two results may find easier identifying the common points and the differences, the stakes and the subtleties of the different invoked theorems. Its statement is complex and is more general than the way it is applied for uncertainty quantification. Care will be taken to highlight this point. Understanding every aspect of the theorem may lead to new interpretations and new resolution schemes (an attempt is made in that direction in chapter 8).

Statement of Cameron-Martin's theorem

Let $\mathcal{C}^0([a, b])$ be the space of continuous functions $f : x \in [a, b] \rightarrow f(x) \in \mathcal{D}_u \subset \mathbb{R}$. Cameron-Martin's theorem uses both the Wiener measure on $\mathcal{C}^0([a, b])$ and the completeness of the Hermite polynomials on \mathbb{R} in order to introduce a complete orthonormal set of functionals on $\mathcal{C}^0([a, b])$ such that every real-valued function $u(f(x))$ in $L^2(\mathcal{C}^0([a, b]))$ has a converging development, in L^2 , on this complete set. We denote by $(\phi_k^H)_{k \in \mathbb{N}}$ the Hermite polynomials, orthonormal with respect to the inner product defined by the gaussian measure, i.e. such that³

$$\int \phi_k^H(x) \phi_l^H(x) d\mathcal{P}_G(x) = \int \phi_k^H(x) \phi_l^H(x) \mathbf{1}_{]-\infty, \infty[}(x) \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \delta_{k,l}, \forall (k, l) \in \mathbb{N}^2.$$

Let us introduce $(\phi_k^\alpha)_{k \in \mathbb{N}}$, a set of real orthonormal functions of $L^2(\mathcal{C}^0([a, b]))$, such that the quantity

$$\int_a^b \phi_k^\alpha(x) df(x) = \int_a^b \phi_k^\alpha(x) f'(x) dx \text{ exists } \forall f \in \mathcal{C}^0([a, b]).$$

Besides, $\forall h(x_1, \dots, x_p)$ such that

$$\int_{-\infty}^{\infty} h(x_1, \dots, x_p) \prod_{i=1}^p d\mathcal{P}_{G_i}(x_i) < \infty,$$

exists, we write (notations)

$$\int_{\mathcal{C}^0([a, b])}^w h \left(\int_a^b \phi_1^\alpha(x) df(x), \dots, \int_a^b \phi_p^\alpha(x) df(x) \right) d_w f = \int_{-\infty}^{\infty} h(x_1, \dots, x_p) \prod_{i=1}^p d\mathcal{P}_{G_i}(x_i).$$

³Note that we here introduce the notation $d\mathcal{P}_G(x) = \mathbf{1}_{]-\infty, \infty[}(x) \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$.

We furthermore define a complete orthonormal set of real functions $(\Phi_{m,n})_{(m,n) \in \mathbb{N}^2}$ belonging to $L^2(\mathcal{C}^0([a,b]))$ by

$$\Phi_{m,n}(f) = \phi_m^H \left(\int_a^b \phi_n^\alpha(x) df(x) \right), \forall (m,n) \in \mathbb{N}^2. \quad (3.2)$$

Besides, $\forall p \in \mathbb{N}$ we have

$$\Psi_{m_1, \dots, m_p}(f) = \Phi_{m_1,1}(f) \times \dots \times \Phi_{m_p,p}(f). \quad (3.3)$$

Expression (3.3) is called the Fourier-Hermite set in [55]. Then the Cameron-Martin's theorem is stated as follows.

Theorem 3.3 *Cameron-Martin's theorem: the Fourier-Hermite series of any real functional $u(f) \in L^2(\mathcal{C}^0([a,b]))$ converges in the $L^2(\mathcal{C}^0([a,b]))$ sense. This means that if $u(f)$ is such that*

$$\int_{\mathcal{C}^0([a,b])}^w |u(f)|^2 d_w f < \infty,$$

then

$$\int_{\mathcal{C}^0([a,b])}^w \left| u(f) - \sum_{m_1, \dots, m_P}^P u_{m_1, \dots, m_P} \Psi_{m_1, \dots, m_P}(f) \right|^2 d_w f < \infty \xrightarrow{P \rightarrow \infty} 0. \quad (3.4)$$

In the above expression, u_{m_1, \dots, m_P} are the Fourier-Hermite coefficients defined by

$$u_{m_1, \dots, m_P} = \int_{\mathcal{C}^0([a,b])}^w u(f) \Psi_{m_1, \dots, m_P}(f) d_w f.$$

Of course, regarding uncertainty quantification, the set of *orthogonal* functions $(\phi_k^\alpha)_{k \in \mathbb{N}}$ of $L^2(\mathcal{C}^0([a,b]))$ is replaced by *uncorrelated continuous* random variables, i.e. orthogonal continuous applications $(X_k)_{k \in \mathbb{N}}$ with $\forall k \in \mathbb{N} X_k : \omega \in \Omega \rightarrow \mathcal{D}_X \subset \mathbb{R}$ where \mathcal{D}_X is not necessarily bounded. At this stage of the discussion, the continuity for the considered random variables may appear strong (as random variables are generally only considered measurable) but more general cases (discrete/discontinuous input random variables) will be discussed later in the document.

A special case of Cameron-Martin's theorem [55]

One detail may trigger the curiosity of the uncertainty analyst familiar with the application of Polynomial Chaos: the summation over P in (3.4) implies a tensorized basis (see expression (3.2)) with components growing with respect to the number of components of the set $(\phi_k^\alpha)_{k \in \mathbb{N}}$, i.e. in an uncertainty quantification context, with respect to the set of random variables $(X_k)_{k \in \mathbb{N}}$. When commonly applied in such context, the expansion depends only on a finite number of random variables $(X_k)_{k \in \{1, \dots, Q\}}$ modeling the uncertain *input* parameters of interest. The theorem usually applied in uncertainty quantification problems corresponds in fact to a special case of Cameron-Martin's theorem dealt with (and called 'special case') in the same publication [55]. Wiener in [295] in fact only stated this special case, but under less constraining hypothesis with respect to the regularity of the solutions.

Theorem 3.4 *A special case of Cameron-Martin's theorem: let $f(u_1, \dots, u_Q)$ be any Q -dimensional function such that*

$$u(u_1, \dots, u_Q) e^{-(u_1^2 + \dots + u_Q^2)} \in L^2(-\infty, \infty), \quad (3.5)$$

and

$$u(f) = u \left(\int_a^b \phi_1^\alpha(x) df(x), \dots, \int_a^b \phi_Q^\alpha(x) df(x) \right).$$

then

$$\int_{C^0([a,b])}^w \left| u(f) - \sum_{m_1, \dots, m_Q=0}^P u_{m_1, \dots, m_Q} \Psi_{m_1, \dots, m_Q}(f) \right|^2 d_w f < \infty \xrightarrow[P \rightarrow \infty]{} 0, \quad (3.6)$$

where u_{m_1, \dots, m_Q} are the Fourier-Hermite coefficients defined by

$$u_{m_1, \dots, m_Q} = \int_{C^0([a,b])}^w u(f) \Psi_{m_1, \dots, m_Q}(f) d_w f.$$

The difference between the statement of 3.3 and its special case 3.4 is identifiable comparing expressions (3.4) and equation (3.6). In the second one, the number of components Q does not grow with P (the polynomial order is the only convergence parameter). The special case of theorem 3.3 states that any Q -dimensional functional from any subset of \mathbb{R} (even unbounded ones) into any subset of \mathbb{R} (even unbounded) verifying (3.5) can be expanded in an infinite sum of Hermite-Fourier coefficients on a Hermite basis of dimension Q .

Remark 3.1 At this stage of the discussion, it is interesting noticing that Grad's 13 moment model briefly presented in chapter 1 (see also [207]) is nothing more than a $P = 1$ -truncated Polynomial Chaos development on the Hermite basis with $Q = 3$ (with respect to $X = (v_1, v_2, v_3)^t$).

In the next chapters (until chapter 8 in fact), we mainly focus on/invoke the special case of Cameron-Martin's theorem. We insist it is implicitly invoked in the literature for uncertainty quantification problems. In chapter 8, we present our attempt to build a new approximation algorithm, based on gPC, from the general Cameron-Martin's theorem (i.e. not its special case).

3.2 Polynomial Chaos for uncertainty quantification (UQ)

Theorem 3.4 is the convergence result at the basis of Polynomial Chaos for uncertainty quantification. It corresponds to the spectral counterpart of the Central Limit Theorem [165, 256] for Monte-Carlo methods. The aim of this section is to illustrate what can be asymptotically expected from a PC approximation. For this, we consider a *very* special case of theorem 3.4: we focus on a 1D application and state the theorem in a probability space.

Corollary 3.1 A very special case of Cameron-Martin's Theorem: let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let $\mathcal{G} \sim \mathcal{G}(0, 1)$ be a centered normalized gaussian variable with probability measure

$$d\mathcal{P}_{\mathcal{G}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Let $(\phi_k^H)_{k \in \mathbb{N}}$ be the basis of Hermite Polynomials, and let $u(\mathcal{G})$ be an unknown random variable written as a transformation of the gaussian one \mathcal{G} .

Suppose that $\int u^2(\mathcal{G}) d\mathcal{P}_{\mathcal{G}} = \mathbb{E}[u^2(\mathcal{G})] < \infty$, then we have

$$u_P(\mathcal{G}) = \sum_{k=0}^P u_k \phi_k^H(\mathcal{G}) \xrightarrow[P \rightarrow \infty]{L^2(\Omega, \mathcal{P})} u(\mathcal{G}).$$

In other words, the polynomial development converges in the $L^2(\Omega, \mathcal{P})$ -norm toward the unknown random variable $u(\mathcal{G})$. The polynomial coefficients $(u_k)_{k \in \mathbb{N}}$ are defined as $u_k = \int u(\mathcal{G}) \phi_k^H(\mathcal{G}) d\mathcal{P}_{\mathcal{G}}$, the projection of the solution on the Hermite basis with respect to the gaussian scalar product.

Without further details, we illustrate the practical application of the above corollary 3.1 in an uncertainty quantification analysis. To do so, we suggest considering several simple uncertainty propagation problems.

3.2.1 Transformation of a gaussian random variable into a uniform one

The first uncertainty quantification problem we consider here transforms a gaussian random variable into a uniform one. This can be done by the application of u defined by $X \rightarrow u(X) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\sqrt{2}}{2}X)$, to $X \sim \mathcal{G}(0, 1)$. Then, $u(X) \sim \mathcal{U}[0, 1]$ is a uniform random variable on the interval $[0, 1]$. Let us apply the material of corollary 3.1 in order to approximate $X \rightarrow u(X)$. Of course, the key step of the method

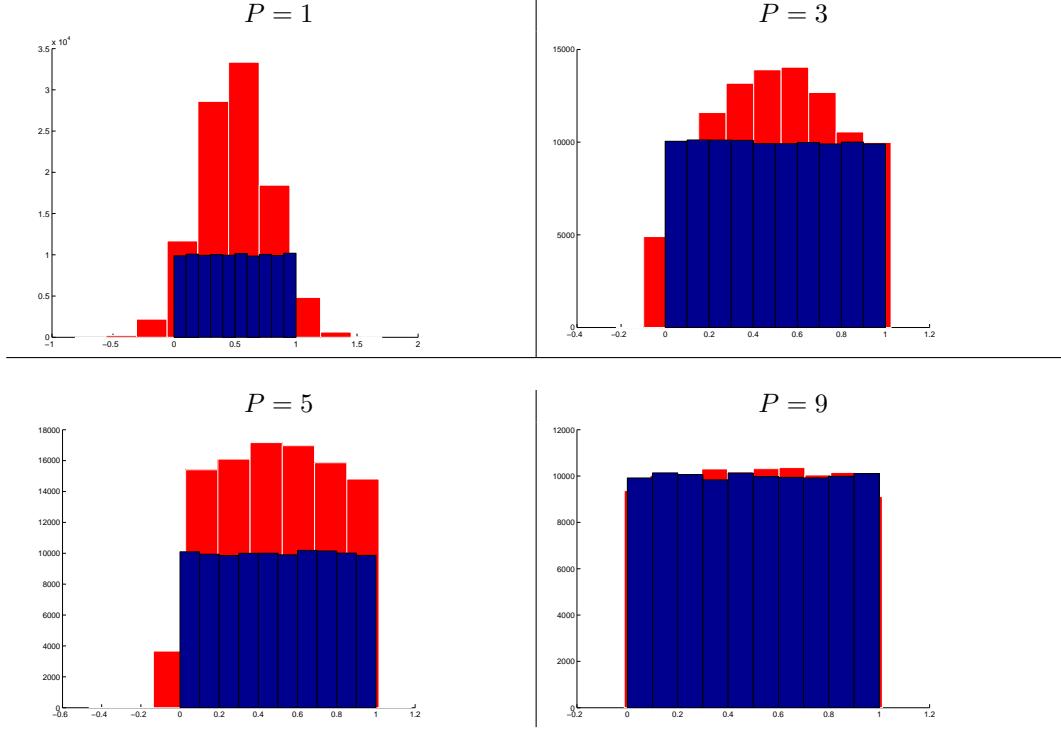


Figure 3.1: Histograms of the sampled PC developments of the random variable $X \sim \mathcal{G}(0, 1) \rightarrow u(X) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\sqrt{2}}{2}X) \sim \mathcal{U}[0, 1]$ for $P = 1, 3, 5, 9$ with analytically computed coefficients $(u_k)_{k \in \{0, P\}}$. The blue histograms are the analytical ones obtained from sampling a uniform random variable $\mathcal{U}[0, 1]$.

consists in the computation of the polynomial coefficients $(u_k)_{k \in \{0, \dots, P\}}$. We will explain how they are computed later in the document: chapters 4 and 5 are dedicated to numerical methods allowing so. In this chapter, consider they are computed very accurately (even analytically whenever it is possible). We aim at presenting what can be asymptotically expected from Polynomial Chaos.

Figure 3.1 compares the histograms obtained thanks to an MC method (reference in blue) and the one obtained by sampling a Gaussian random variable $X \sim \mathcal{G}(0, 1)$ and applying the PC_P developments for different truncation orders P (red histograms). Qualitatively, we identify a convergence behaviour: increasing P allows the approximated histograms to get closer to the target one (in blue). For $P = 9$, the PC_9 and the MC histograms are almost indistinguishable.

Once the polynomial coefficients $(u_k)_{k \in \{0, \dots, P=9\}}$ computed, the PC_P histograms are obtained by sampling from a polynomial $X \rightarrow \sum_{k=0}^P u_k \phi_k(X)$ rather than from a complex functional $X \rightarrow u(X)$ potentially costly. In the literature, the PC_P expansions are often called *metamodels or surrogate models*. In this manuscript, a metamodel is also denoted by *reduced model* (mainly in chapter 4) or more often a *gPC approximation* (in chapter 5).

Once the histogram available, one can work on many statistical features such as mean, variance, high order moments, failure probabilities etc. depending on the purpose of the uncertainty analysis (see the main chart in [271]). Consequently, once the histogram is accurately enough computed, we consider the resolution of the uncertainty propagation step done, the rest being less costful postprocessings.

Now, let us consider the same uncertainty quantification problem as before but from a more quantitative point of view. Figure 3.2 presents a convergence study with respect to P in L^2 -norm for the

PC_P developments of $X \sim \mathcal{G}(0, 1) \rightarrow u(X) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\sqrt{2}}{2}X) \sim \mathcal{U}[0, 1]$. As can be seen on figure 3.2,

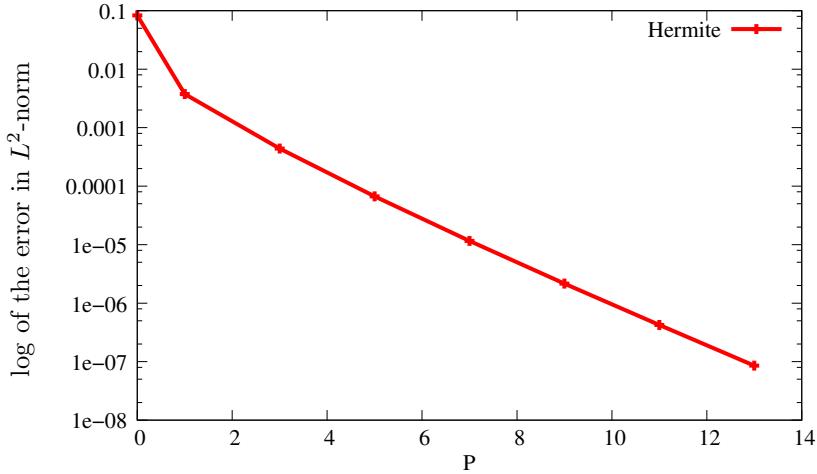


Figure 3.2: Convergence study of the PC_P approximation for the transformation of a gaussian random variable $X \sim \mathcal{G}(0, 1) \rightarrow u(X) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\sqrt{2}}{2}X) \sim \mathcal{U}[0, 1]$ into a uniform one.

for $P > 1$, the curve exhibits an exponential convergence rate: linear dependence with respect to P of $\log(e_{L^2})$ where e_{L^2} is the L^2 -norm of the error. This exponential convergence rate is characteristic of spectral methods. From a practical point of view, the gain is considerable: with $P = 9$, the error in L^2 -norm is $\approx 10^{-6}$. Such accuracy, with a MC method can only be reached with $N_{MC} \approx 10^{12}$ samples which remains unaffordable for many applications. As a consequence, if we are able to compute the coefficients $(u_k)_{k \in \{0, \dots, P\}}$ efficiently enough (in comparison to a certain amount of MC runs), PC_P stands for an interesting alternative.

3.2.2 Mapping of a uniform random variable into an Arcsinus and a Binomial one

We now consider two other simple transformations. They will help understand the subtlety of the introduction of *generalized Polynomial Chaos* [305, 291] few years after Polynomial Chaos [124] in an uncertainty quantification context. For this, we first perform the same kind of convergence study as before but on the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X)$ into an Arcsin law. We then apply the PC development to the transformation of a uniform random variable into a Binomial one *via* the transformation $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \mathbf{1}_{]-\infty, -\frac{1}{2}]}(X)$. The latter transformation is studied to anticipate the behaviour of the approximation when discontinuous solutions are appearing (cf. the 'fil rouge' problem presented in chapter 2).

Mapping of a uniform random variable into an Arcsinus law

We first apply PC_P developments for several truncation orders P for transformation $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X)$ and display the same convergence study as in figure 3.2. Note that in order to apply PC here, we first had to map the input random variable X into a gaussian one in order to compute the PC coefficients on the Hermite basis. The results are presented in figure 3.3 and once again, the test problem suggests PC_P yields an exponential convergence rate. But the curve is much flatter than in the previous example. Indeed, a 10^{-2} accuracy is not even reached with $P = 9$ on this second example. To attain an equivalent accuracy, the Monte-Carlo method needs about 10^4 samples: for such transformation, we can consider both methods compete. The fact that the convergence rate of spectral methods strongly depends on the regularity of the solution is well-known [42]. This represents a drawback with respect to MC methods for which it is relatively insensitive. For MC methods, the regularity of u mainly affects the constant multiplying the convergence rate, not the convergence rate, see section 5.2 of chapter 5.

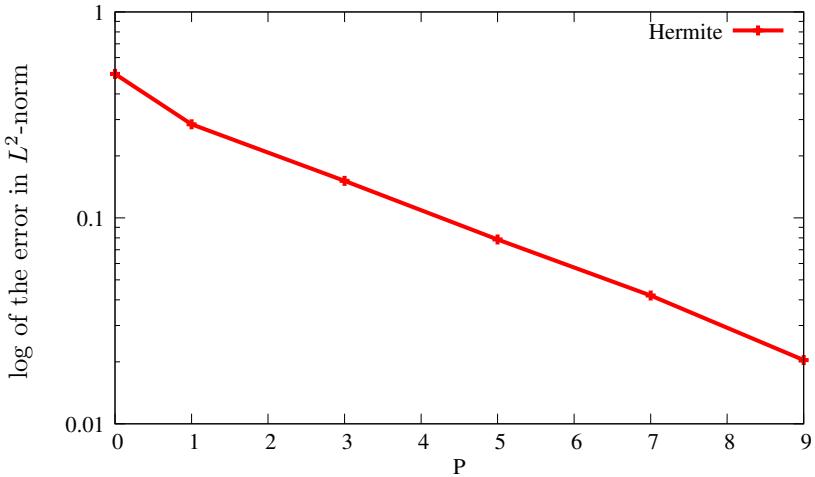


Figure 3.3: Convergence study of the PC_P approximation for the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X) \sim \mathcal{A}$ into an Arcsinus one.

Next, in order to anticipate with the kind of solution which can be encountered when dealing with hyperbolic systems of conservation laws, let us consider another simple transformation of a uniform random variable into a binomial one.

Mapping of a uniform random variable into a Binomial law

We here apply PC_P developments for several truncation orders P on the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1]$ via the application $u(X) = \mathbf{1}_{]-\infty, -\frac{1}{2}]}(X)$. In this case, the output $u(X) \sim \mathcal{B}_{0,1}(\frac{3}{4}, \frac{1}{4})$ is a binomial one with state 0 having probability $\frac{3}{4}$ and state 1 having probability $\frac{1}{4}$. Once again, the input X being non-gaussian, we first need to map X into a gaussian random variable to compute the PC coefficients on the Hermite basis. We display the same kind of convergence study as before in figure 3.4. For such discrete output, the PC_P approximation does not yield an exponential convergence rate. Obtaining an accuracy below 10^{-1} in the L^2 -norm is very hard. It needs the compu-

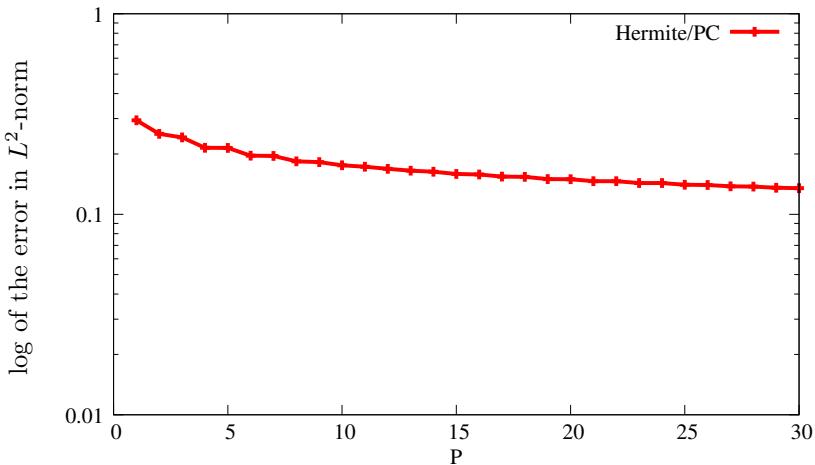


Figure 3.4: Convergence study of the PC_P approximation for the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \mathbf{1}_{]-\infty, -\frac{1}{2}]}(X) \sim \mathcal{B}$ into a Binomial one.

tation of more than $P = 30$ coefficients. The same accuracy in L^2 -norm can be attained with an MC method with about $N_{MC} = 100$ samples.

With these last two examples, it is not obvious Polynomial Chaos represents an interesting alternative to the Monte-Carlo method. In fact, Polynomial Chaos has been quite condemned in many papers between 1970 and 1980, see for example [66, 67, 103], emphasizing the need of very high order developments for real life applications (study of turbulence for example, see [66]). The introduction of *generalized Polynomial Chaos* (gPC) by Karniadakis *et al.* [305, 291] gave a new breath to the spectral approximation. It suggests a new approximation method obtained by only slightly changing the approximation basis. It is described in the next section.

3.3 Introduction of generalized Polynomial Chaos (gPC) for UQ

In order to fully understand the differences between *generalized Polynomial Chaos* and *Polynomial Chaos*, we suggest stating a second 'version/corollary', still in 1D stochastic dimension, of Cameron-Martin theorem which corresponds to the gPC application in the literature. It has been *conjectured and applied* in many numerical examples by Karniadakis *et al.* (2002-2006) [305, 291] and *fully demonstrated* in [104].

Corollary 3.2 *Convergence of generalized Polynomial Chaos: let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let X be an arbitrary random variable of given probability measure $d\mathcal{P}_X(x)$. Let $(\phi_k^X)_{k \in \mathbb{N}}$ be the basis of orthonormal polynomials with respect to $d\mathcal{P}_X(x)$*

$$\int \phi_k^X \phi_t^X d\mathcal{P}_X = \delta_{k,t}, \forall (k,t) \in \mathbb{N}^2.$$

Suppose we define a dense orthogonal set in $L^2(\Omega, \mathcal{F}, \mathcal{P})$. Let $u(X)$ be an unknown random variable. Suppose that $\int u^2(X) d\mathcal{P}_X < \infty$. Then the polynomial development $u_P^X(X) = \sum_{k=0}^P u_k^X \phi_k^X(X) \xrightarrow[P \rightarrow \infty]{L^2(\Omega, \mathcal{P})}$ $u(X)$, converges in the $L^2(\Omega, \mathcal{P})$ -norm toward the unknown random variable $u(X)$. The polynomial coefficients $(u_k)_{k \in \mathbb{N}}$ are defined by $u_k^X = \int u(X) \phi_k^X(X) d\mathcal{P}_X$. They are the projection of the solution on above polynomial basis with respect to the scalar product defined by the probability measure $d\mathcal{P}_X$.

Note that the general convergence theorem seems to demand an additional condition in comparison to Stone-Weierstrass' or Cameron-Martin's on the input distribution: it must have a uniquely solvable moment problem⁴ for the set of polynomial to be dense in L^2 . This assumption is in fact already in both seminal theorems because the Hermite polynomials are dense in L^2 . The hypothesis deserves some more details, it will be dealt with in the next section regarding the construction of the gPC basis. Paper [104] also recalls the convergence is ensured for quantiles⁵, relative moments⁶, in probability⁷: in other words, for every classical mathematical tools to perform a (converging) uncertainty quantification analysis. In the following sections, we consider random input variable X for which the gPC convergence holds [104].

To sum-up, some complementary theorems ensure the convergence for some particular measures⁸ $d\mathcal{P}_X$ but the complete convergence theorem for gPC has been demonstrated in [104]. Now, the efficiency of gPC vs. PC has been *numerically observed* in many fields of applications [291, 294, 293, 299, 167, 76, 304, 306, 305, 208]. In the next paragraphs, we suggest revisiting the previous numerical examples with gPC and compare its performances with PC.

⁴It is not restrictive for the distribution of the Askey scheme [12, 305, 303, 291] nor discrete and mixed distributions but may be problematic for the lognormal one [104].

⁵Relevant when one is interested in approximating a probability of failure.

⁶Relevant when one is interested in central quantities.

⁷Relevant when one is interested in approximating the probability density function of the output variable by a histogram.

⁸For example if the support of $d\mathcal{P}_X$ is bounded (uniform, arcsinus, beta laws etc.) the Stone-Weierstrass theorem ensures the convergence of the gPC expansion. For the Poisson distribution (unbounded discrete distribution), the convergence is ensured by the very same theorem ensuring the convergence of the Gram-Charlier expansion etc.

Mapping of a uniform random variable into an Arcsinus law: comparison of PC and gPC

Without any further comment, we apply the gPC material to the transformation of a uniform random variable into an Arcsinus laws, just as in one of the previous paragraph and compare the results obtained by both approximations, PC_P and gPC_P . Figure 3.5 (left) compares the convergence studies for $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X) \sim \mathcal{A}$ with PC_P and gPC_P up to order $P = 9$. The red curve (PC_P) is the

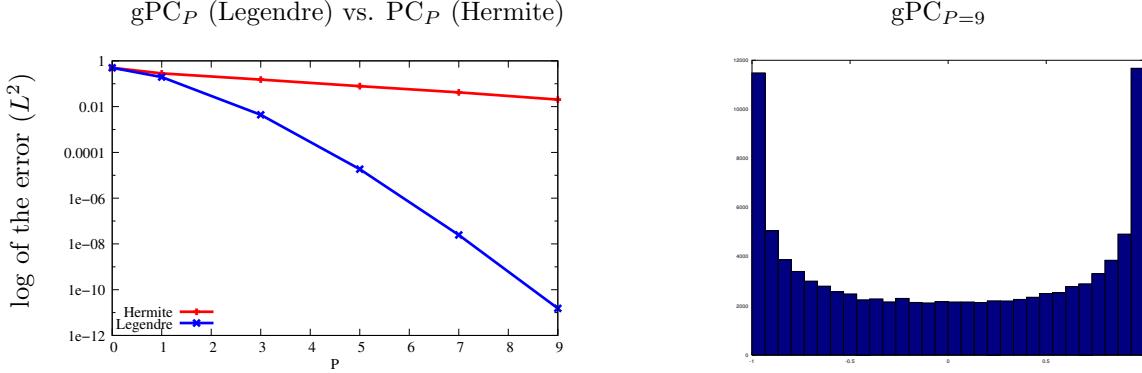


Figure 3.5: Left: Comparison of the convergence rates of PC_P and gPC_P for the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X) \sim \mathcal{A}$ into an Arcsinus one. Right: comparison of the target histogram and the $\text{gPC}_{P=9}$ one (they match exactly).

same as in section 3.2. The blue one (gPC_P) is obtained applying the new version of Cameron-Martin theorem. The convergence rate with gPC is way better on this test-problem. For $P = 9$, the level of accuracy of the approximation is $\approx 10^{-10}$. Such level of accuracy was far from being reached with PC. Figure 3.5 (right) presents the histogram of the pdf obtained with $\text{gPC}_{P=9}$: it is not distinguishable from the target histogram of the Arcsin law.

Mapping of a uniform random variable into a Binomial law: comparison of PC and gPC

Let us revisit the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \mathbf{1}_{[-\infty, -\frac{1}{2}]}(X) \sim \mathcal{B}$ into a Binomial one with gPC instead of PC. The red curve in figure 3.6 (left) is the same as in figure

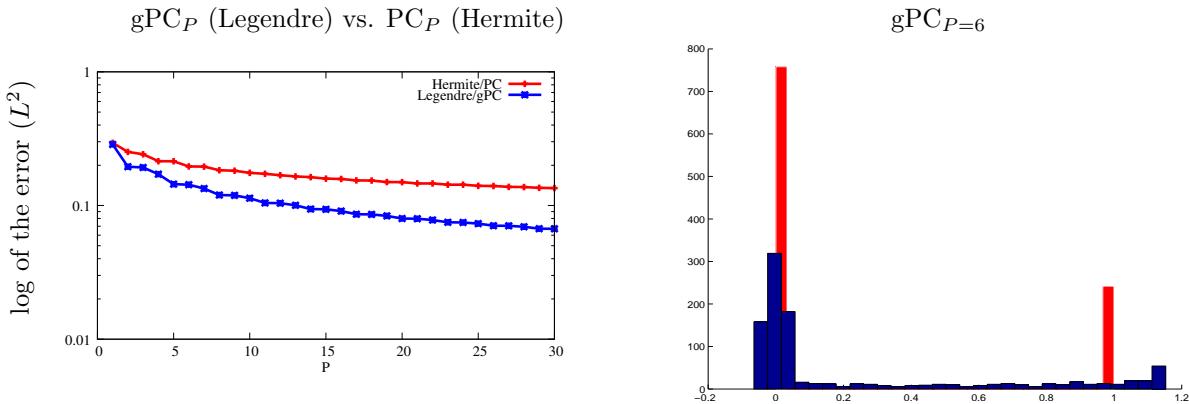


Figure 3.6: Left: Comparison of the convergence rates of PC_P and gPC_P for the transformation of a uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \mathbf{1}_{[-\infty, -\frac{1}{2}]}(X) \sim \mathcal{B}$ into a Binomial one. Right: comparison of the target histogram obtained with MC and the gPC_6 one.

3.4 obtained with PC_P whereas the blue one corresponds to the gPC_P convergence study. The gPC approximation is always (i.e. $\forall P$) better than the PC one but the convergence remains non-spectral (this is predicted by theory, see [42]). Figure 3.6 (right) shows the target histogram obtained with an

MC method (red) and the one obtained with the gPC₆ approximation: the gPC approximation poorly captures the discrete behaviour of the solution. Such oscillating approximation corresponds to what is commonly called the Gibbs phenomenon in the literature.

Summary

With the previous examples, we aimed at highlighting there exists more or less adapted basis for approximating a given random variable $u(X)$. Above all, this was already implicitly contained in the original Cameron-Martin's theorem [55]. In fact, in [291, 305], the authors re-interpreted Cameron-Martin's theorem and put forward there even exists an optimal basis for any given L^2 random variable $u(X)$. In table 3.1, few correspondances between optimal basis and random variables are given. Some of those correspondances were also given in the *Askey scheme* see [259]. In fact, table 3.1 must be understood

	Random variables	Askey Polynomials
Continuous laws	gaussian gamma beta uniform Arcsinus	Hermite Laguerre Jacobi Legendre Chebyshev
Discrete laws	Poisson Binomial negative-Binomiale Hypergeometric	Charlier Krawtchouk Meixner Hahn

Table 3.1: Askey scheme: correspondances between random variables and optimal gPC basis.

this way: let X be a uniform law, let $u(X)$ be a transformation of X distributed according to an Arcsin law. Then the development of the target random variable $u(X)$ on a *Chebyshev basis* yields the fastest convergence rate: it is analytical with only 2 polynomial coefficients, i.e. order $P = 1$, u_0 representing the mean and u_1 the variance. More than that, the same correspondance exists also for discrete laws, hence for transformation $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \mathbf{1}_{]-\infty, -\frac{1}{2}]}(X) \sim \mathcal{B}$ for which, according to table 3.1, the Krawtchouk polynomials yields the best convergence rate (perfect accuracy for $P = 1$). Of course, knowing the optimal basis implies knowing the output distribution which is precisely the unknown. Still, having such existence information for the optimal basis allows looking for it and designing algorithm aiming at it: this is the purpose of chapter 6 describing our contribution to non-intrusive gPC (see iterative gPC or i-gPC [238, 31, 242]).

Remark 3.2 Corollary 3.2 does not mean gPC is superior to PC: assume u transforms a uniform random variable X into a gaussian one. Then PC will hold the optimal convergence rate.

The two last sections presented two approximation methods, PC and gPC. They are closely related in the sense the approximation algorithms for the two approaches come from two slightly different interpretations of Cameron-Martin's theorem [55]. Its complexity and completeness makes it a probable source for even newer interpretations, hence new approximation tools.

Before tackling the different practical ways for computing the gPC coefficients, we present in the next section the construction of the gPC basis for arbitrary random variables (i.e arbitrary probability measures). Indeed, the first condition for applying corollary 3.2 is to be able to build the orthonormal basis with respect to the inner product defined by the probability measure of the input random variable X .

3.4 The construction of the gPC basis

A gPC basis is nothing more than an orthonormal polynomial basis. The gPC procedure only helps the uncertainty analyst *a priori* choosing an efficient one. Orthonormal polynomials have been intensively studied in the literature, see amongst others [273, 5, 117, 129]. In this document, we do not aim at

being exhaustive on the subject. We only recall their main properties, those useful for the construction of an arbitrary gPC basis associated to the inner product defined by an arbitrary probability measure $d\mathcal{P}_X$. We also prepare some notions and notations for the question of numerical integration with Gauss quadrature rules which is central for the non-intrusive application in chapter 5.

3.4.1 Inner product defined by an arbitrary probability measure

Let $d\mathcal{P}_X$ be an arbitrary probability measure related to an arbitrary random variable X . Care will be taken in this section to be able to consider *any* random variable, continuous (gaussian, uniform,...) and even discrete/categorial (binomial, multinomial,...). We introduce the inner product defined by the probability measure $d\mathcal{P}_X$ as

$$\langle f, g \rangle_X = \int f(x)g(x)d\mathcal{P}_X(x). \quad (3.7)$$

The above inner product, for a probability measure of a discrete random variable having $D + 1 < \infty$ states $(x_j)_{j \in \{0, \dots, D\}}$ with probabilities $(p_j)_{j \in \{0, \dots, D\}} > 0$, resumes to

$$\langle f, g \rangle_X = \int f(x)g(x)d\mathcal{P}_X(x) = \int f(x)g(x) \sum_{j=0}^D p_j \delta_{x_j}(x) = \sum_{j=0}^D p_j f(x_j)g(x_j). \quad (3.8)$$

In this particular case, the output random variable is in a finite vector space (of size $D + 1 < \infty$). With expression (3.8), we insist on the fact that the notation (3.7) is compatible with continuous and discrete input random variables. The notations are inspired from the ones of [117].

A sequence of P orthogonal polynomials $(\phi_k^X)_{k \in \{0, \dots, P\}}$ associated to the probability measure $d\mathcal{P}_X$ has the following properties:

- ϕ_k^X is of degree k ,
- $\langle \phi_k^X, \phi_l^X \rangle_X = 0, \forall (k, l) \in \{0, \dots, P\}^2$ such that $k \neq l$.

The sequence is said

- *orthonormal* if $\langle \phi_k^X, \phi_k^X \rangle_X = 1 \forall k \in \{0, \dots, P\}$,
- *monic* if the coefficient of the highest degree (i.e. of x^k for ϕ_k^X) for every polynomials of the sequence $(\phi_k^X)_{k \in \{0, \dots, P\}}$ is 1. In this document, we use the additional superscript m , i.e. $(\phi_k^{X,m})_{k \in \mathbb{N}}$, to denote monic (orthogonal) polynomials.

Obviously, the families $(\phi_k^{X,m})_{k \in \mathbb{N}}$ of monic orthogonal polynomials and $(\phi_k^X)_{k \in \mathbb{N}}$ of orthonormal polynomials are related. Their relative expression will be given in the next section once an additional notion introduced. For conciseness in the following document, the sequence of polynomials orthonormal with respect to the inner product defined by a given probability measure $d\mathcal{P}_X$ is referred as *the polynomials associated to X or $d\mathcal{P}_X$* . Such sequence is closely related to the *statistical moments* of its corresponding random variable X or probability measure $d\mathcal{P}_X$. In the next section, we recall this important link.

3.4.2 Moments of a probability measure and Hankel determinants

Consider an arbitrary random variable X having probability measure $d\mathcal{P}_X$ and suppose $\forall k \in \mathbb{N}$

$$s_k^X = \int x^k d\mathcal{P}_X(x) < \infty. \quad (3.9)$$

Then the sequence of numbers $(s_k^X)_{k \in \mathbb{N}}$ is called the sequence of moments of the random variable X or of the probability measure $d\mathcal{P}_X$. Note that if the support of $X/d\mathcal{P}_X$ is bounded, the existence

of the sequence (3.9) of moments such that (3.9) $\forall k \in \mathbb{N}$ is straightforward. We define the Hankel matrices/determinants of the random variable X or of the probability measure $d\mathcal{P}_X$ by $\forall k \in \mathbb{N}$

$$\underline{H}_{2k}^X = \begin{vmatrix} s_0^X & s_1^X & \dots & s_k^X \\ \dots & \dots & \dots & \dots \\ s_n^X & s_{n+1}^X & \dots & s_{n+k}^X \\ \dots & \dots & \dots & \dots \\ s_k^X & \dots & \dots & s_{2k}^X \end{vmatrix}, \quad \underline{H}_{2k+1}^X = \begin{vmatrix} s_1^X & s_2^X & \dots & s_{k+1}^X \\ \dots & \dots & \dots & \dots \\ s_n^X & s_{n+1}^X & \dots & s_{n+k}^X \\ \dots & \dots & \dots & \dots \\ s_{k+1}^X & \dots & \dots & s_{2k+1}^X \end{vmatrix}. \quad (3.10)$$

Determining whether a sequence of finite real numbers $(s_k)_{k \in \mathbb{N}}$ are moments of a unique or not random variable/probability measure is what is commonly called the *classical moment problem*. If such random variable/probability measure *exists* and is *unique*, the problem is said *determinate*. If it *exists* and is *not unique*, the problem is said *indeterminate*. Depending on the support of the random variable/probability measure, the moment problem is called [204, 7]:

- the *Hausdorff moment problem* when $X/d\mathcal{P}_X$ has a bounded support, i.e. $X \in \prod_{i=1}^Q [a_i, b_i]$. In this particular case, if a sequence $(s_k^X)_{k \in \mathbb{N}}$ is a sequence of moments of a random variable/probability measure, then the problem is determinate [204, 156, 7].
- The *Stieltjes moment problem* when $X/d\mathcal{P}_X$ has a half-line support, i.e. $X \in \prod_{i=1}^Q [a_i, \infty[$. In such conditions, if a sequence $(s_k^X)_{k \in \mathbb{N}}$ is a sequence of moments of a random variable/probability measure, the problem may be indeterminate. A sufficient condition for uniqueness can be express as

$$\sum_{k=0}^{\infty} (s_k^X)^{-\frac{1}{2k}} = \infty. \quad (3.11)$$

It is called *Carleman's condition* see [7].

- The *Hamburger moment problem* when $X/d\mathcal{P}_X$ has an unbounded support, i.e. $X \in \mathbb{R}^Q$. In such conditions, a sequence of moments may also be indeterminate and the *Carleman's condition* see [7] in this case is given by

$$\sum_{k=0}^{\infty} (s_{2k}^X)^{-\frac{1}{2k}} = \infty. \quad (3.12)$$

Consequently, depending on the support of the random variable/probability measure of interest, if existence holds, uniqueness is not always straightforward. In the indeterminate cases, the solutions of the moment problem form a convex set. Most of all for our applications, in the *determinate* moment problem case, the set of polynomials are *dense* in the associated Hilbert space.

So far, we mainly dealt with uniqueness and assumed existence. Let $(s_k^X)_{k \in \mathbb{N}}$ be a sequence of numbers satisfying $\forall k \in \mathbb{N} s_k^X < \infty$. Suppose the Hankel determinants defined by (3.10) satisfies:

- either $\forall k \in \mathbb{N} \underline{H}_{2k}^X > 0$ and $\underline{H}_{2k+1}^X > 0$,
- or $\forall (2k, 2k+1) \in \{0, \dots, D\}^2 \underline{H}_{2k}^X > 0$ and $\underline{H}_{2k+1}^X > 0$ and $\underline{H}_{2k}^X = \underline{H}_{2k+1}^X = 0$ for larger k ,

then the sequence is a sequence of moments of a random variable/probability measure for the classical moment problem (i.e. independently of being a Hausdorff, Stieltjes or Hamburger moment problem). Note that in the second case, the Hilbert space associated to the existing measure is finite-dimensional and of size $D + 1$.

As briefly tackled before, the existence of a set of dense polynomials associated to a given random variable/probability measure $X/d\mathcal{P}_X$ is closely related to the existence and the determinacy of its moments. In the following section, we present Christoffel's formulae which explicits the relation between the moments of a random variable and the orthonormal polynomials associated to this same one.

3.4.3 Christoffel's formulae, Jacobi's matrix and construction procedures

Christoffel's formulae [7, 156] explicits the relation between the sequence of moments $(s_k^X)_{k \in \mathbb{N}}$ of a random variable X and the set of orthonormal polynomials $(\phi_k^X)_{k \in \mathbb{N}}$ associated to X . It is given by

$$\forall n \in \{0, \dots, P\}, \phi_n^X(x) = \frac{1}{\sqrt{\underline{H}_{2(n-1)}^X \underline{H}_{2n}^X}} \begin{vmatrix} s_0^X & s_1^X & \dots & s_n^X \\ \dots & \dots & \dots & \dots \\ s_k^X & s_{k+1}^X & \dots & s_{n+k}^X \\ \dots & \dots & \dots & \dots \\ 1 & x^1 & \dots & x^n \end{vmatrix}. \quad (3.13)$$

In (3.13) appears the previous Hankel determinants (3.10). From (3.13), it is easy recovering the polynomial sequence may only exist up to a certain order D (whether the Hankel determinants are all strictly positive or if there exists an order after which they are all zero). In the following document, we keep considering, for convenience, polynomials orders $k \in \mathbb{N}$ even if for some probability measure they exist only up to a certain order $P \in \mathbb{N}$. With (3.13), it is easy verifying, see [7], the monic orthogonal polynomials $(\phi_k^{X,m})_{k \in \mathbb{N}}$ can be expressed with respect to both the orthonormal ones $(\phi_k^X)_{k \in \mathbb{N}}$ and the Hankel determinants as we have $\forall k \in \mathbb{N}$:

$$\phi_k^X(x) = \Gamma_k^X \phi_k^{X,m}(x) = \sqrt{\frac{\underline{H}_{2(k-1)}^X}{\underline{H}_{2k}^X}} \phi_k^{X,m}(x). \quad (3.14)$$

Christoffel's formulae (3.13) obviously represents a way to build the gPC basis associated to the probability measure $d\mathcal{P}_X$. It definitely has a theoretical interest but it is scarcely used in practice due to the difficulty to accurately numerically compute the Hankel determinants. The problem is more and more ill-conditioned as the polynomial order increases. To illustrate this, suppose the sequence of moments $(s_k^X)_{k \in \{0, \dots, 2P\}}$ of an *existing* random variable X are not accurately known⁹ and assume a perturbation of these moments such that

$$(s_k^\varepsilon)_{k \in \{0, \dots, 2P\}} = (s_k^X + \varepsilon_k)_{k \in \{0, \dots, 2P\}} \approx (s_k^X)_{k \in \{0, \dots, 2P\}}.$$

For the sake of simplicity of the following developments, we suppose a particular form for the perturbation $\varepsilon = (\varepsilon_0, \dots, \varepsilon_{2P}) = (0, \dots, 0, \delta)$ so that we suppose every moments of order $n \in \{0, \dots, 2P-1\}$ are accurately computed whereas the last one s_{2P}^X is perturbed by δ . The polynomials orthonormal with respect to the perturbed moments coincide with the one of X up to order $P-1$ and we have

$$\phi_P^\varepsilon(x) = \frac{1}{\sqrt{\underline{H}_{2(P-1)}^X \underline{H}_{2P}^\varepsilon}} \begin{vmatrix} s_0^X & s_1^X & \dots & s_n^X \\ \dots & \dots & \dots & \dots \\ s_k^X & s_{k+1}^X & \dots & s_{n+k}^X \\ \dots & \dots & \dots & \dots \\ 1 & x^1 & \dots & x^P \end{vmatrix}. \quad (3.15)$$

In other words, ε only affects the last Hankel determinant $\underline{H}_{2P}^\varepsilon$ hence the last polynomial ϕ_P . Now, from the definition of $\underline{H}_{2P}^\varepsilon$ and by a development of the last line of the determinant, we have

$$\underline{H}_{2P}^\varepsilon = \underline{H}_{2P}^X + \delta \underline{H}_{2(P-1)}^X \quad \text{so that} \quad \phi_P^\varepsilon(x) = \frac{1}{\sqrt{1 + \delta \frac{\underline{H}_{2(P-1)}^X}{\underline{H}_{2P}^X}}} \phi_P^X(x). \quad (3.16)$$

Consequently, the sequence of polynomials $(\phi_k^X)_{k \in \mathbb{N}}$ associated to X and the sequence of polynomials $(\phi_k^\varepsilon)_{k \in \mathbb{N}}$ associated to the perturbed moments are such that

$$- \forall n \in \{0, \dots, P-1\}, \phi_n^\varepsilon = \phi_n^X,$$

⁹they may be only computed, estimated etc.

– and for the last component, we have

$$\phi_P^\varepsilon(x) = \phi_P^X(x) - \frac{1}{2} \frac{H_{2(P-1)}^X}{H_{2P}^X} \phi_P^X(x) \delta + \mathcal{O}(\delta^2). \quad (3.17)$$

It is known in the litterature [14, 118, 7, 156] that in the previous conditions $\forall n \in \mathbb{N}$

$$\frac{H_{2n}^X}{H_{2(n-1)}^X} \leq 2^{-(4n+2)} \quad \text{leading to} \quad \left| \frac{d\phi_P^X}{d\delta} \right| \geq 2^{4P+1}.$$

It testifies for a higher and higher sensivity of the orthonormal basis components with respect to a small inaccuracy in the statistical moments/Hankel determinants (δ) as the polynomial order P increases. The inaccuracy δ may come from an approximation of the moments but also from roundoff errors due to the determinant computation algorithm. In order to avoid such aliasing errors, more stable algorithm are available such as the Chebyshev one or the modified Chebyshev one (see [117]). Those algorithms intensively use another important property of orthonormal polynomials. For any set of orthonormal polynomials up to order P , there exists two sequences of coefficients $(\alpha_k)_{k \in \{0, \dots, P\}}$ and $(\beta_k)_{k \in \{0, \dots, P\}}$ such that $\forall k \in \{0, \dots, P\}$

$$\sqrt{\beta_{k+1}} \phi_{k+1}^X(x) = (x - \alpha_{k+1}) \phi_k^X(x) - \sqrt{\beta_k} \phi_{k-1}^X(x). \quad (3.18)$$

In the above expression, $\forall k \in \{0, \dots, P\}$, we have

$$\alpha_k = \frac{\int x \phi_k^X(x) \phi_k^X(x) d\mathcal{P}_X(x)}{\langle \phi_k, \phi_k \rangle_X},$$

and $(\beta_k)_{k \in \{0, \dots, P\}}$ are such that $\forall k \in \{0, \dots, P\}$

$$\beta_k = \frac{\langle \phi_k^X, \phi_k^X \rangle_X}{\langle \phi_{k-1}^X, \phi_{k-1}^X \rangle_X}.$$

Equation (3.18) is refered to as the *three term recurrence formulae* in the literature. The Chebyshev algorithms, even if known to be more stable, may also suffer inaccuracy¹⁰ in the estimations of $(\alpha_k, \beta_k)_{k \in \mathbb{N}}$. The relation between the moments and the coefficients of the three-term recurrence formulae can be explicited, we refer to [273, 117] for the interested reader. It must be kept in mind that the orthonormal basis associated to any random variable with a too important order may bear unworkable inaccuracies.

The three-term recurrence formulae can be written in a matrix form by introducing the *Jacobi matrix* of order P defined by

$$J_P^X = \begin{pmatrix} \alpha_1 & \sqrt{\beta_1} & 0 & 0 & \dots & 0 \\ \sqrt{\beta_1} & \alpha_2 & \sqrt{\beta_2} & 0 & \dots & 0 \\ 0 & \sqrt{\beta_2} & \alpha_3 & \sqrt{\beta_3} & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \sqrt{\beta_P} \end{pmatrix}. \quad (3.19)$$

Denote by $\Phi_P^X = (\phi_0^X, \dots, \phi_P^X)^t$ the vector of $P+1$ components of the sequence of orthonormal polynomials associated to X , the three-term recurrence formulae ensures that

$$x \Phi_P^X(x) = J_P^X \Phi_P^X(x) + \sqrt{\beta_P} \phi_{P+1}^X(x) e_P, \quad (3.20)$$

where $e_P = (0, \dots, 0, 1)^t$ is of size $P+1$. It is then interesting noticing that the $P+1$ roots $(X_i)_{i \in \{0, \dots, P\}}$ of polynomial ϕ_{P+1}^X are the eigenvalues of the Jacobi matrix of order P . We have $\forall k \in \{0, \dots, P\}$

$$X_k \Phi_P^X(X_k) = J_P^X \Phi_P^X(X_k).$$

¹⁰As the sequence $(\beta_k)_{k \in \mathbb{N}}$ is obviously related to the sequence of moments, see (3.13).

Vectors $(\Phi_P^X(X_k))_{k \in \{0, \dots, P\}}$ are consequently the corresponding eigenvectors. This property is intensively used in order to build Gauss quadrature rules, see chapter 5.

3.4.4 Taking into account discrete/categorical input variables with gPC

In the previous sections, care has been taken to detail the case of discrete random variables with $D + 1$ states even if the notations and the results did not particularly need it. We here want to insist on the fact that discrete input random variables are included in the gPC framework (already in [295] or in the Askey scheme of table 3.1). This property is not obvious in the literature. It has triggered questionings from some of my students. The construction procedure for the gPC basis of any discrete law is exactly the same as for continuous random variables or at least the same as for random variables having an *infinite* sequence of orthonormal polynomials. The only difference comes from the fact that one should not try to build the gPC basis after a certain order depending on the number of states of the categorial random variable. By convention in the following document and without loss of generalities, we denote by $(\phi_k^X(X))_{k \in \mathbb{N}}$ the gPC basis associated to any random variable X : if X is discrete then $\exists k_0 \in \mathbb{N}$ such that $\forall k > k_0, \phi_k^X(X) = 0$ and the notation still holds.

We do not spend too much time on this particular case but let us consider a practical example derived from the 'fil rouge' problem of chapter 2. Suppose the initial interface position can be modeled by a discrete random variable with three equiprobable states $X_0 = 0.45, X_1 = 0.5, X_2 = 0.55$ with probabilities $p_0 = p_1 = p_2 = \frac{1}{3}$. In this particular conditions, the initial and final realisations of the discrete interface position are also given by the three initial realisations displayed on the top pictures (but the bottom ones are not representative of this choice here) of figure 2.2. The gPC basis exists up to order $P = 2$, it corresponds to the Meixner polynomials (see table 3.1). The output random variables have at most three discrete states and we know in the vicinity of the shock or the interface, they have *only two discrete ones*. We assure they can efficiently be captured by the three term Meixner basis.

This section is almost allusive but we wanted to insist on this point. We now spend few sections illustrating the main drawbacks of gPC, especially having in mind the 'fil rouge' problem of chapter 2.

3.5 Curse of dimensionality and Gibbs phenomenon

In this section, we focus on the two main drawbacks of gPC, the curse of dimensionality and the sensitivity to Gibbs phenomenon. As explained before, the dimensionality problem is not our main issue, we are dealing with a fairly reasonable number of input uncertainties (see chapter 2). The Gibbs phenomenon on the contrary is unavoidable in our applications, even in 1D stochastic dimension. In the next section, we briefly describe and illustrate both difficulties.

3.5.1 Curse of dimensionality

The first drawback is quite simple to understand and illustrate: the previous corollaries (PC with 3.1 and gPC with 3.2) were stated in 1D stochastic dimension. Their multi-dimensional counterparts corresponds to a tensorization of 1D polynomial basis, exactly as detailed in the original paper [55] and recalled in theorem 3.4. Assume $X = (X_1, \dots, X_Q)^t$ is a vector of *independent* components, a gPC approximation of the transformation $u(X)$ can be obtained by introducing Q one-dimensional gPC basis $(\phi_k^i(X_i))_{k \in \mathbb{N}, i \in \{1, \dots, Q\}}$ orthonormal with respect to the independent probability measures of each component of the random vector $\forall (k, l) \in \mathbb{N}^2, i \in \{1, \dots, Q\}$:

$$\int \phi_k^i(x_i) \phi_l^i(x_i) d\mathcal{P}_{X_i}(x_i) = \delta_{k,l}.$$

Approximating $u(X)$ on the Q -tensorized P -truncated polynomial basis implies the determination of $(P+1)^Q$ coefficients. An increase in the dimension Q leads to an exponential increase of the number of coefficients which are directly related to the computational resources needed (see chapters 4 to 8). Note that we supposed $X \in \mathbb{R}^Q$ is a Q -dimensional *independent* random vector. Independence here is convenient but does not imply any loss of generality. Correlated random variables can be treated the

very same way conditionally to a primary modeling *via* copulae, see [169, 170, 100, 34] for some very pedagogical examples.

Several authors aimed at relieving the impact of the curse of dimensionality in high dimensions. Each attempt comes with an additional hypothesis. For example in [34], the author build some sparse representations assuming many coefficients are zero. For this, an L^1 penalization is introduced. Other authors [34, 35, 262, 158, 258, 72] choose to compute the coefficients on some simplexes based on hypothesis of regularity of the output random variables with respect to the different input dimensions: the truncation orders $(P_i)_{i \in \{1, \dots, Q\}}$ is different depending on the stochastic directions $(X_i)_{i \in \{1, \dots, Q\}}$ for which some smoothness criterions are satisfied *a posteriori* by the output. In the following sections, due to the appearance of discontinuous solutions in our applications, we do not make any of the above assumptions on the regularity of the solution.

3.5.2 Sensitivity to the Gibbs phenomenon

If the curse of dimensionality can be considered a minor difficulty in the studies we are interested in, this is quite different for the Gibbs phenomenon. Let us consider the previously tackled problem of transforming a uniform random variable $X \sim \mathcal{U}([-1, 1])$ into a binomial one *via* application $u(X)$ defined as

$$u(x) = \mathbf{1}_{]-\infty, -\frac{1}{2}]}(x) = \begin{cases} 1 & \text{if } x \leq -\frac{1}{2}, \\ 0 & \text{else.} \end{cases} \quad (3.21)$$

The transformation u maps a uniform random variable into a binomial one with states $\{0, 1\}$ with respective probabilities $\{\frac{3}{4}, \frac{1}{4}\}$. Discontinuous functions are in L^2 so there is no surprise the classical theorems apply. We verify it experimentally: figure 3.7 compares the gPC_{P=6} approximation to the analytical solution in term of functional representation (left) and histogram of the pdf of the output (right). The functional representation of figure 3.7 (left) testifies of the discontinuous behaviour of the

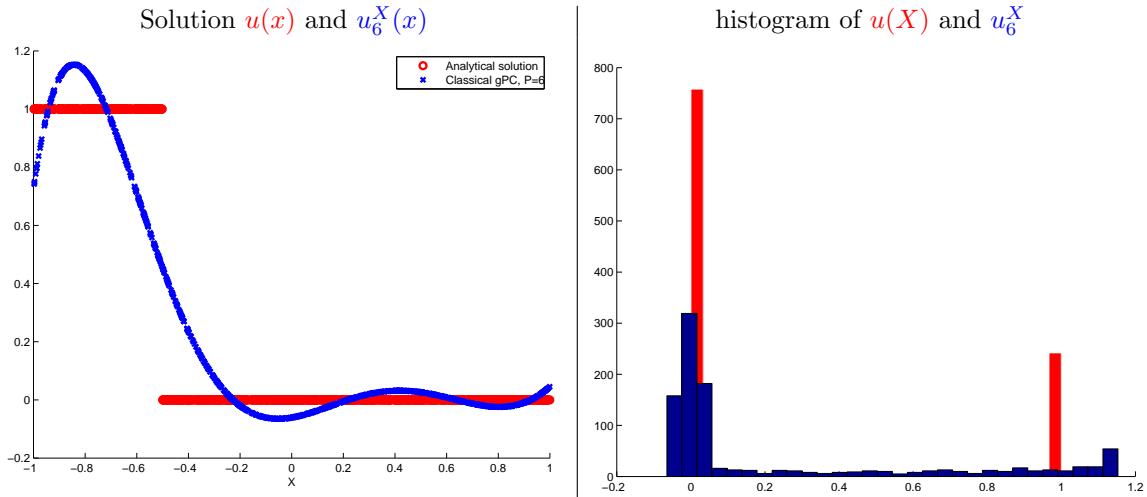


Figure 3.7: gPC and Gibbs phenomenon in term of functional representation (left) and histogram (right)

output $u(x)$ with respect to the uncertain parameter (red curve for the reference solution). The histogram of the pdf, figure 3.7 (right), testifies of the bimodal behaviour of the output random variable $u(X)$. Note that the above simple test-problem gives a cheap and efficient emulation of the hydrodynamical problem detailed in chapter 2: the histogram of figure 3.7 (right) is close enough to the ones obtained in the vicinities of the shock or of the interface in figure 2.3.

In figure 3.7, the gPC_{P=6} approximation fails to capture the discontinuous behaviour of the output random variable. The support of the output random variable is missed (the maximum principle is not respected) and the behaviour is misevaluated. In fact, for such problem, increasing the polynomial order does not allow any great improvement: the convergence rate is very slow as testifies figure 3.8. In fact

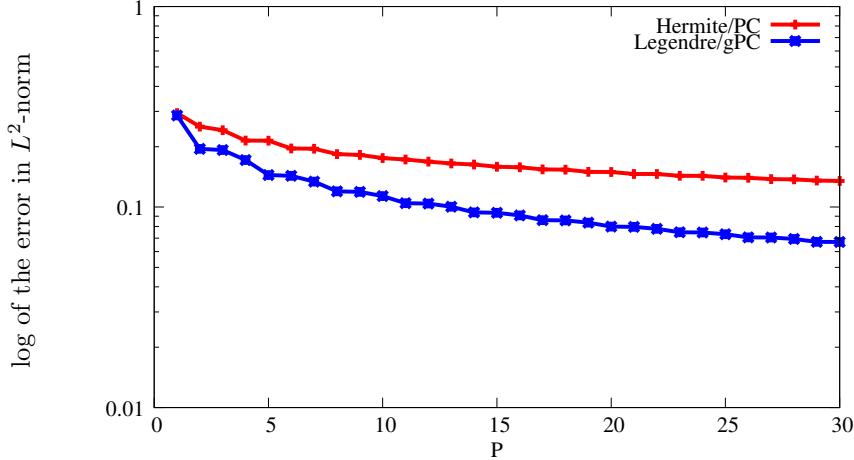


Figure 3.8: Convergence study of the PC_P and gPC_P approximations for the transformation of a uniform random variable $X \sim \mathcal{U}[-1, 1] \rightarrow u(X) = \mathbf{1}_{]-\infty, \frac{1}{2}]}(X) \sim \mathcal{B}$ into a Binomial one.

for such low regularity, the convergence rate is even sublinear: in figure 3.8, the logarithm of the error with respect to P is not a straight line.

Several authors suggested algorithmic solutions in order to reduce the effect of the Gibbs phenomenon on the gPC approximations. The main ones are gPC/Haar Wavelets representations (Lemaître *et al.* [185]), Multi-Element gPC (Karniadakis *et al.* [294]) or stochastic WENO-like gPC (Abgrall [4]) etc. To describe them briefly, the Haar wavelets are powerful in order to represent discontinuous solutions but fail for smooth ones (whereas gPC does not). The complementary use of both representations is very interesting but needs an arbitrary criterion in order to choose which one to use. The ME-gPC approximation consists in tracking and locating the discontinuity in the random space before decomposing it in N_s subspaces in which are solved N_s independent uncertainty propagation problems (instead of only one initially). It consequently raises other complex algorithmic questions. The WENO-like approximations on the contrary consists in accepting locally lower degree representations based on oscillatory behaviours/criterions.

As explained before, the Gibbs phenomenon will be intensively encountered in the application of interest and our contributions mainly consist in dealing with it, from an intrusive point of view (robustness difficulties in chapter 4), or a non-intrusive one (accuracy issues in chapter 5).

3.6 Summary for generalized Polynomial Chaos

With this chapter, we wanted to illustrate under which conditions gPC stands for an interesting alternative to MC methods: especially in relatively low stochastic dimensions and for relatively smooth solutions. The next chapters are dedicated to a presentation of our contributions to uncertainty quantification thanks to gPC based reduced model. They are of two different natures, opening to two different possibilities:

- the first possibility consists in working on the resolution method, i.e. on gPC, to increase its efficiency. We are motivated by hyperbolic systems and in this context, the main weakness concerns the sensitivity to Gibbs phenomenon¹¹. In chapters 4, 6 and 8 our different contributions will aim at alleviating the hypothesis of regularity of the solution for the gPC approximations. We will not tackle the curse of dimensionality in this part of the document. High (physical and stochastic) dimension problems will rather be considered in part III.
- The second possibility consists in accepting the flaws of gPC (with respect to dimensionality and the regularity of the solutions) and apply it in well-suited relevant situations. This will be the

¹¹In the sense it also occurs in small stochastic dimensions (also in 1D as in the example of chapter 2).

purpose of chapter 7 in which non-intrusive gPC reduced models will be applied to accelerate computations to predict the growth rate of hydrodynamical instabilities.

Until now, we presented what can asymptotically be expected of gPC and did not detail how to compute the gPC coefficients $(u_k)_{k \in \{0, \dots, P\}}$ which are the keys of the approach. Two main ways exist in the literature for this. Intrusive methods imply building a reduced model, consequently solving a new set of PDE and developing a new resolution code. They are detailed, together with our contributions for systems of conservation laws, in chapter 4. Non-intrusive methods use a resolution code for the deterministic PDE of interest as a black-box (just as MC methods do). They are detailed in chapter 5, together with our contributions in chapters 6, 7 and 8.

Chapter 4

Intrusive application of gPC for systems of conservation laws

Some P_n -like or M_n -like gPC based moment models

Contents

4.1	Intrusive application of gPC	51
4.1.1	The P -truncated gPC reduced model: a P_n -like closure	51
4.1.2	Roe solver for the P -truncated intrusive gPC reduced model	54
4.1.3	Application to the 'fil rouge' problem of chapter 2	55
4.2	A step-by-step study of intrusive gPC for systems of conservation laws	56
4.2.1	The particular case of a scalar conservation law	56
4.2.2	Possible loss of wellposedness for non-scalar systems of conservation laws	59
4.2.3	A closure ensuring wellposedness for general systems of conservation laws	61
4.2.4	Application to the 'fil rouge' problem of chapter 2	66
4.3	Summary for intrusive gPC and the entropy closure reduced models	70

In this chapter, we present a first way to compute the coefficients $(u_k^X)_{k \in \{0, \dots, P\}}$ of a gPC expansion. It is commonly called *intrusive* gPC. The methodology is general and can be applied to any systems but we focus on conservation laws. Its application results in the construction and resolution of a *reduced model of order P* , a new set of PDEs derived from the initial uncertain one. Intrusiveness refers to the fact this resolution comes with more or less¹ important modifications of a simulation code. The initial uncertain system of interest is not directly solved, it is first pretreated to obtain a new model, built from it. In the following sections, we show that the reduced model of order P of any hyperbolic system of conservation of size d is a new system of conservation law of size $d \times (P + 1)$ whose *wellposedness is not necessarily guaranteed*. In order to understand the stakes and the importance of the property, we recall few notions specific to hyperbolic systems. These are stated in simple forms² as we do not intend to be exhaustive on this topic. For the latter purpose, we rely on [81, 260, 81, 261].

One dimensional uncertain systems of conservation laws can be written in the general form

$$\begin{cases} \partial_t u(x, t, X) + \partial_x f(u(x, t, X)) = 0, \\ u(x, 0, X) = u_0(x, X). \end{cases} \quad (4.1)$$

We suppose system (4.1) is hyperbolic $\forall X \in \Omega$. Hyperbolicity ensures existence and uniqueness of its solutions, hence wellposedness. A system of conservation laws of the form of (4.1) is called *hyperbolic* if

¹From rewriting completely a simulation code, as in this section, to minor modifications of an existing one, as in section 9.11.2.

²for example in 1D space dimension whereas they are also true in nD , etc.

- either, the jacobian matrix of the flux $\nabla_u f(u(x, t, X))$ is diagonalizable in a complete basis of eigenvectors $\forall u(x, t, X) \in \mathcal{D}_u$.
- or³ there exists a *strictly convex mathematical entropy* s for system (4.1). A mathematical entropy is a real function

$$s : u \in \mathcal{D}_u \subset \mathbb{R}^d \longrightarrow s(u) \in \mathbb{R},$$

such that there exists

$$g : u \in \mathcal{D}_u \subset \mathbb{R}^d \longrightarrow g(u) \in \mathbb{R}^d,$$

called the entropy flux, such that for smooth solution u of (4.1), we have

$$\partial_t s(u) + \partial_x g(u) = 0. \quad (4.2)$$

Besides, if $s(u)$ is strictly convex on \mathcal{D}_u , it satisfies

$$\partial_t s(u) + \partial_x g(u) \leq 0, \quad (4.3)$$

for discontinuous solutions.

In [81, 260, 81, 261], *physical systems* are defined as systems of conservation laws satisfying the second point. In this part II, we only consider physical system in the sense of [81, 260, 81, 261].

Now that few notions have been briefly reminded, we suggest applying intrusive gPC as described in the literature to our 'fil rouge' system (Euler, see chapter 2) and identify different practical issues.

4.1 Intrusive application of gPC

The methodology consists in two main steps:

- first, the construction of a P -truncated gPC reduced model.
- Second, the development of a numerical solver and its implementation in a resolution code for the newly built system of equations.

Both steps are detailed in sections 4.1.1 and 4.1.2 for general conservation laws before focusing on the Euler system and our 'fil rouge' configuration in section 4.1.3.

4.1.1 The P -truncated gPC reduced model: a P_n -like closure

Intrusive gPC is also called stochastic Galerkin gPC (sG-gPC) in the literature [185, 186, 215, 168]. Its practical application resumes to few steps: the first one consists in building the gPC basis $(\phi_k^X)_{k \in \mathbb{N}}$ associated to the probability measure $d\mathcal{P}_X$ of the input random variable X . This step has been described in section 3.4. Once the gPC basis at hand, the construction of the P -truncated reduced model consists in formally assuming

$$u(x, t, X) = \sum_{k=0}^{\infty} u_k^X \phi_k^X(X),$$

³Note that the second condition implies the first one whereas the inverse is not true but once again we refer to [81, 260, 81, 261] for more details on hyperbolic systems of conservation laws.

and introducing the above expansion in (4.1). The next step consists in performing a Galerkin projection up to order P to get

$$\left\{ \begin{array}{l} \partial_t u_0^X(x, t) + \partial_x \int f \left(\sum_{k=0}^{\infty} u_k^X(x, t) \phi_k^X(X) \right) \phi_0^X(X) d\mathcal{P}_X = 0, \\ \dots, \\ \partial_t u_k^X(x, t) + \partial_x \int f \left(\sum_{k=0}^{\infty} u_k^X(x, t) \phi_k^X(X) \right) \phi_l^X(X) d\mathcal{P}_X = 0, \\ \dots, \\ \partial_t u_P^X(x, t) + \partial_x \int f \left(\sum_{k=0}^{\infty} u_k^X(x, t) \phi_k^X(X) \right) \phi_P^X(X) d\mathcal{P}_X = 0. \end{array} \right. \quad (4.4)$$

The above system is not closed. To obtain $d \times (P + 1)$ unknowns and equations, a P_n -like closure hypothesis⁴ is introduced and can be stated as follow: assume that $\forall x \in \mathcal{D}, t \in [0, T], X \in \Omega$ the solution $u(x, t, X)$ of (4.1) can be accurately described with exactly $P + 1$ terms, i.e.

$$u(x, t, X) = \sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \text{ with } u_k^X = 0 \ \forall k > P. \quad (4.5)$$

With (4.5) in mind, (4.4) simply becomes

$$\left\{ \begin{array}{l} \partial_t u_0^X(x, t) + \partial_x \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \phi_0^X(X) d\mathcal{P}_X = 0, \\ \dots, \\ \partial_t u_k^X(x, t) + \partial_x \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \phi_l^X(X) d\mathcal{P}_X = 0, \\ \dots, \\ \partial_t u_P^X(x, t) + \partial_x \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \phi_P^X(X) d\mathcal{P}_X = 0. \end{array} \right. \quad (4.6)$$

This process is intimately related to the methodology applied to build a Finite Element scheme [43]. For this reason, intrusive gPC is also denoted as stochastic Finite Element in some publications relative to elliptic equations [272, 75, 123, 157, 124, 192, 125, 142, 35, 36, 114, 300, 269, 122]. System (4.6) is a new system of conservation laws with main conservative variable $U(x, t) = (u_0^X(x, t), \dots, u_P^X(x, t))^t$ and flux

$$F(U(x, t)) = \left(\int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \phi_0^X(X) d\mathcal{P}_X, \dots, \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \phi_P^X(X) d\mathcal{P}_X \right)^t.$$

It is closed, in the sense it has $d \times (P + 1)$ equations and the same amount of unknowns. But the nonlinearity⁵ of f may induce the need for an additional hypothesis to obtain a more explicit expression of the flux with respect to U . To illustrate this, let us focus on the P -truncated Euler system⁶. Reduced

⁴The term ' P_n -like closure hypothesis' is not commonly used in the literature, it is introduced in this document to put forward some analogies with kinetic theory, see chapter 1.

⁵The case of a linear f is straightforward and is not considered in this manuscript. For more details we refer to [152, 278].

⁶With a perfect gas closure as in chapter 2.

model (4.7) is obtained applying the previously detailed process to (2.2):

$$\left\{ \begin{array}{l} \partial_t \rho_0 + \partial_x (\rho u)_0 = 0, \\ \dots \\ \partial_t \rho_P + \partial_x (\rho v)_P = 0, \\ \partial_t (\rho v)_0 + \partial_x \left(\frac{3-\gamma}{2} \sum_{i,j,l=0}^P (\rho v)_i (\rho v)_j \left(\frac{1}{\rho} \right)_l c_{i,j,l,0} + (\gamma-1)(\rho e)_0 \right) = 0, \\ \dots \\ \partial_t (\rho v)_P + \partial_x \left(\frac{3-\gamma}{2} \sum_{i,j,l=0}^P (\rho v)_i (\rho v)_j \left(\frac{1}{\rho} \right)_l c_{i,j,l,P} + (\gamma-1)(\rho e)_P \right) = 0, \\ \partial_t (\rho e)_0 + \partial_x \left(\begin{array}{l} \gamma \sum_{i,j,l=0}^P (\rho v)_i (\rho e)_j \left(\frac{1}{\rho} \right)_l c_{i,j,l,0} \\ - \frac{\gamma-1}{2} \sum_{i,j,l,m,t=0}^P (\rho v)_i (\rho v)_j (\rho v)_t \left(\frac{1}{\rho} \right)_l \left(\frac{1}{\rho} \right)_m c_{i,j,l,t,m,0} \end{array} \right) = 0, \\ \dots \\ \partial_t (\rho e)_P + \partial_x \left(\begin{array}{l} \gamma \sum_{i,j,l=0}^P (\rho v)_i (\rho e)_j \left(\frac{1}{\rho} \right)_l c_{i,j,l,P} \\ - \frac{\gamma-1}{2} \sum_{i,j,l,m,t=0}^P (\rho v)_i (\rho v)_j (\rho v)_t \left(\frac{1}{\rho} \right)_l \left(\frac{1}{\rho} \right)_m c_{i,j,l,t,m,P} \end{array} \right) = 0. \end{array} \right. \quad (4.7)$$

In (4.7), we introduced $c_{i,j,\dots,l} = \int \phi_i^X \phi_j^X \dots \phi_l^X d\mathcal{P}_X$. In the expression of the flux of (4.7) appears a new quantity $((\frac{1}{\rho})_k)_{k \in \{0, \dots, P\}}$, highlighted in red, which remains to be defined. In fact, its appearance is even the result of an implicit assumption: the same development as in hypothesis (4.5) applies also to $\frac{1}{\rho}$, i.e. $(\frac{1}{\rho}) = \sum_{k=0}^P (\frac{1}{\rho})_k \phi_k^X$, and not only to the components of the main variable $(\rho, \rho u, \rho e)^t$. Closing the reduced model (4.7) consists in defining $((\frac{1}{\rho})_k)_{k \in \{0, \dots, P\}}$ with respect to $(\rho_0, \dots, \rho_P)^t$. Many numerical strategies can be found in the literature to 'treat' such nonlinearity [76, 299] and achieve the closure purpose here. Some needs additional hypothesis on the variance of the random variable, some intensively use the structure of the nonlinearities at play, etc. The aim here is not to go through a review of all of the proposed methods of [76, 299]. We only insist that in [239, 232], we studied and analysed some of them for the uncertain p -system⁷ and explained in which way they were not satisfying enough with respect to the hyperbolicity of the reduced model. One closure, for example, transforms the hyperbolic p -system into an only *weakly hyperbolic* reduced model, see [232, 239]. The choice of the closure consequently confers more or less satisfying properties to (4.7). In practice, due to the important size of the reduced model, these properties may be complex to study and it is tempting assuming hyperbolicity for (4.7) (once a closure chosen) before going straight to its numerical resolution. In the next section 4.1.2, we typically succumb to the previous temptation and settle for a closure choice dictated by the possibility to develop a converging numerical scheme for (4.7) rather than on a hyperbolicity-based one.

In this section, we went through the general methodology to build a reduced model from intrusive gPC. We put forward an analogy with the construction of P_n models for the linear transport equation, see chapter 1 for example. The P_n models rely on the same assumption as (4.5), the gPC basis being only replaced by the Legendre one (or spherical harmonics, see for example [281, 196, 141]). The main difference between both (P_n and gPC ones) comes from the fact that the closure bearing relevant properties is not straightforward as soon as f is nonlinear. In the following section, we detail one first way to close system (4.7) based on some *practical* needs: the closure ensures the existence of a converging numerical scheme for the resolution of the reduced model.

⁷System of conservation laws of size $d = 2$ such that $p(\rho, \varepsilon) = p(\rho)$ also called *isentropic Euler* system.

4.1.2 Roe solver for the P -truncated intrusive gPC reduced model

The way the system is closed confers particular properties to the system. These properties, in general, are hard to study mainly due to the size of the built reduced model. Moreover, even once the system closed and its properties studied, it still remains to solve it. Designing a numerical solver for (4.7) for arbitrary truncation orders P is not more straightforward than the mathematical analysis of the reduced model. In the next section, we emphasize some closures make the freshly built P -truncated reduced model easier to solve: choosing them hits two birds with the same stone as they allow having the good amount of unknowns/equations together with an efficient numerical resolution. In particular, we demonstrate that some closures preserve the existence of a *homogeneous* change of variables for the reduced model if the property holds for the initial system of interest.

Definition 4.1 (Homogeneous change of variable of order K) Two vectors $u \in \mathbb{R}^d$ and $z \in \mathbb{R}^d$ are homogeneous of order $K \in \mathbb{N}$ if they verify $Ku(z) = \nabla_z u(z)z$.

Then we have the following general property:

Property 4.1 Let a hyperbolic system of conservation laws having for unknown the vector $u \in \mathbb{R}^d$. Assume there exists $z \in \mathbb{R}^d$ such that u and z are homogeneous of order K . If $U = (u_0^X, \dots, u_P^X)^t \in \mathbb{R}^{d \times (P+1)}$ is the vector of unknown of a reduced model of order P built from intrusive gPC, then there exists a vector $Z \in \mathbb{R}^{d \times (P+1)}$ such that U and Z are homogeneous of order K .

Proof Let $(u, z) \in \mathbb{R}^d \times \mathbb{R}^d$ be homogeneous of order K . Let $(\phi_k^X)_{k \in \{0, \dots, P\}}$ be the gPC basis associated to the probability measure $d\mathcal{P}_X$. Suppose $z = \sum_{k=0}^P z_k^X \phi_k^X$ and $Z = (z_0^X, \dots, z_P^X)^t$. Let us furthermore define $\forall i \in \{0, \dots, P\}$ $u_i^X = \int u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_i^X d\mathcal{P}_X$ and $U = (u_0^X, \dots, u_P^X)^t$. It now remains to verify that the vectors U and Z are homogeneous of order K . For this, let us consider

$$\nabla_Z U(Z)Z = \begin{pmatrix} \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_0^X \phi_0^X d\mathcal{P}_X & \dots & \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_0^X \phi_P^X d\mathcal{P}_X \\ \dots & \dots & \dots \\ \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_P^X \phi_P^X d\mathcal{P}_X & \dots & \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_P^X \phi_P^X d\mathcal{P}_X \end{pmatrix} \begin{pmatrix} z_0^X \\ \vdots \\ z_P^X \end{pmatrix}.$$

The t^{th} component of $\nabla_Z U(Z)Z$ is denoted $(\nabla_Z U(Z)Z)_t$ and we get

$$\begin{aligned} (\nabla_Z U(Z)Z)_t &= \sum_{i=0}^P \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_t^X z_i^X \phi_i^X d\mathcal{P}_X, \\ &= \int \nabla_z u \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \left[\sum_{i=0}^P z_i^X \phi_i^X \right] \phi_t^X d\mathcal{P}_X. \end{aligned} \tag{4.8}$$

Using the homogeneity property for u and z together with their respective definition, we obtain

$$\begin{aligned} (\nabla_Z U(Z)Z)_t &= \int Ku \left(\sum_{k=0}^P z_k^X \phi_k^X \right) \phi_t^X d\mathcal{P}_X, \\ &= KU_t, \end{aligned} \tag{4.9}$$

and the property holds. ■

A direct application of the above property leads to the construction of a Roe scheme for the reduced model if there exists a Roe scheme for the system of conservation laws of interest. For the Euler system of our 'fil rouge' configuration, there exists a homogeneous change of variable of order 2 between the conservative variable $u = (\rho, \rho v, \rho e)^t$ and $z = (\sqrt{\rho}, \sqrt{\rho}v, \sqrt{\rho}(e + \frac{p}{\rho}))^t$. Applying property 4.1, the vectors

$$U = (\rho_0, \dots, \rho_P, (\rho v)_0, \dots, (\rho v)_P, (\rho e)_0, \dots, (\rho e)_P)^t,$$

and

$$Z = \left((\sqrt{\rho})_0, \dots, (\sqrt{\rho})_P, (\sqrt{\rho}v)_0, \dots, (\sqrt{\rho}v)_P, \left(\sqrt{\rho}(e + \frac{p}{\rho}) \right)_0, \dots, \left(\sqrt{\rho}(e + \frac{p}{\rho}) \right)_P \right)^t,$$

are also homogeneous of order 2. Consequently, applying the P_n -like hypothesis (4.5) to Z rather than to U ensures there exists a Roe scheme for the reduced model defined by the above change of variable. We do not detail the equivalent of (4.7) for the previous change of variable, its expression is complex. In [232, 80], we verify the Roe conditions are satisfied and apply the Roe solver for different reduced models from different systems of conservation laws (but we did not apply it to our 'fil rouge' configuration). In [224], the same solver for the Euler system, is applied in different interesting configurations.

Remark 4.1 *The change of variable allows building a converging scheme for the reduced model whatever the truncation order P . However, just as in the previous case in which the closure needed the definition of $((\frac{1}{\rho})_k)_{k \in \{0, \dots, P\}}$ as a function of $(\rho_k)_{k \in \{0, \dots, P\}}$, we here need $(\rho_k)_{k \in \{0, \dots, P\}}$ as a function of $((\sqrt{\rho})_k)_{k \in \{0, \dots, P\}}$. In practice, during the computations, we rely on the same process as described in [76] for treating the square nonlinearity, it does not degrade the homogeneous property, see [232].*

We suggest presenting the results obtained with the built reduced model solved with the Roe scheme to the 'fil rouge' configuration of interest detailed in chapter 2.

4.1.3 Application to the 'fil rouge' problem of chapter 2

We described the methodology to build a reduced model of order P . The choice of the closure has been made so that a converging numerical resolution strategy (Roe scheme) can be designed for every truncation order P of the reduced model and we developed a simulation code. In [224], [232, 80], the resolution has been performed in several configurations but we here are interested to the configuration described in chapter 2.

We initiated our simulation code with the condition of the configuration of chapter 2 and the results are... Not presented because the code crashed at the first iteration. The obtained reduced model clearly lacks robustness, at least for low polynomial orders P . In the vicinity of the initial interface position, the mass density is discontinuous and its gPC approximation exhibits an oscillatory behaviour. For low polynomial orders P , the gPC approximation of the mass density can be negative whereas the closure presented in the previous section needs the computation of $((\sqrt{\rho})_k)_{k \in \{0, \dots, P\}}$ from $(\rho_k)_{k \in \{0, \dots, P\}}$, the calculation being ill conditioned.

For higher polynomial order P , the conditioning of the problem probably improves but increasing P may lead to unaffordable sizes of system. Besides, the same situation can occur in the middle of a simulation, dynamically, rather than initially: in [232, 237], a Richtmyer-Meshkov shock tube⁸ is studied and in such configuration, the crash occurs when a shock hits the interface between the two fluids. In [232, 80], the reduced model with Roe scheme/homogeneous closure has also been successfully applied in several configurations (in the sense the computations did not crash). But as will be emphasized in the following sections, computations with reduced models which are not wellposed are of poor interest (section 4.2.2). Furthermore, in [232, 80], some configurations did not seem to account for a loss of wellposedness: the initial conditions were such that the initial uncertainty was very small. Experimentally, we also observed the stronger the initial uncertainty is, the greater P must be to avoid a crash. The 'small variance' regime is studied more in detail in chapter 7: a parallel is made between gPC based models and perturbative ones.

As a conclusion of this first section 4.1 on intrusive gPC, the results are mitigated. First, the construction of a gPC reduced model from a given system of conservation laws is complex, mainly due to the many possible choices for the closure. These choices characterise the reduced model and its properties, hence its theoretical quality and its simplicity with respect to numerical resolution. The reduced model is often of important size and designing a numerical solver for its resolution $\forall P$ can be difficult. We put forward the possibility to choose the closure so as to have a numerical solver at hand but the (lack of) numerical results on the 'fil rouge' configuration tends to show that choices only dictated by practical considerations can lead to a waste of time (the time for developing the simulation code with the Roe scheme resulting in an unworkable computation code!). With this example, we wanted to justify the

⁸Its one dimensional counterpart is presented in chapter 7.

need for a theoretical analysis in order to close the reduced model properly *prior* to studying resolution algorithms.

In the next sections, instead of considering directly the Euler system and our configuration of interest, we adopt a more step-by-step approach in order to identify and analyse accurately the causes of the numerical instabilities encountered in this section. For this, we first consider scalar conservation law, then a system of size 2 (Shallow water) before going back to the Euler one.

4.2 A step-by-step study of intrusive gPC for systems of conservation laws

Regarding the robustness difficulties of the previous section, in order to identify and analyse the properties of the intrusive gPC reduced model, we consider a more step-by-step approach. Instead of immediately considering the Euler system, we first study scalar equations (we focus on Burgers' equation but the material of section 4.2.1 can be easily applied to any scalar conservation law). We then consider, in section 4.2.2, a system of 2 conservation laws, the Shallow water system. Finally, we go back to general systems of conservation laws. The 'fil rouge' configuration is finally treated in section 4.2.3.

4.2.1 The particular case of a scalar conservation law

Scalar conservation laws correspond to the special case $d = 1$ in the general form (2.1):

$$\partial_t u(x, t, X) + \partial_x f(u(x, t, X)) = 0, u \in \mathcal{D}_u \subset \mathbb{R}^{d=1}. \quad (4.10)$$

In this particular case, it is quite easy studying the hyperbolicity of the intrusive gPC reduced model: for any truncation order P and any closure choice with respect to f , it is given by

$$\partial_t \begin{pmatrix} u_0^X(x, t) \\ \dots \\ u_P^X(x, t) \end{pmatrix} + \partial_x \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \begin{pmatrix} \phi_0^X(X) \\ \dots \\ \phi_P^X(X) \end{pmatrix} d\mathcal{P}_X = 0. \quad (4.11)$$

System (4.11) is a system of conservation laws of size $P + 1$ and is hyperbolic. The flux of system (4.11) is given by

$$F(u_0^X, \dots, u_P^X) = \int f \left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right) \begin{pmatrix} \phi_0^X(X) \\ \dots \\ \phi_P^X(X) \end{pmatrix} d\mathcal{P}_X.$$

The general term of the jacobian of the flux (we drop the dependences with respect to x, t, X for the sake of simplicity) is given by

$$\left[\nabla_{(u_0^X, \dots, u_P^X)} F(u_0^X, \dots, u_P^X) \right]_{k,l} = \int f' \left(\sum_{i=0}^P u_i^X \phi_i^X \right) \phi_k^X \phi_l^X d\mathcal{P}_X.$$

The system being scalar, f' is scalar and consequently $\nabla_{(u_0^X, \dots, u_P^X)} F(u_0^X, \dots, u_P^X)$ is a symmetric matrix, hence the hyperbolicity of (4.11). Such property, in the scalar case, has been intensively studied in the literature [236, 232], [278, 152, 162] whereas to our knowledge, this is much less the case for systems (i.e. $d > 1$). In practice, with numerical resolutions of reduced models derived from scalar conservation laws, we did not experienced any robustness issues nor numerical instabilities. For example, in [232], a Roe scheme (almost as in the previous section) has been applied to Burgers' equation and the computations always went well (robust) and gave satisfactory results with respect to accuracy. In fact, in the case of Burgers' equation, we are even able to demonstrate spectral convergence, see [84], for early times (i.e. before the appearance of a discontinuous solution). The theorem is stated here but we refer to [84] for

the proof. Let us consider an uncertain problem for Burgers' equation:

$$\begin{cases} \partial_t u(x, t, X) + \partial_x \left(\frac{u^2(x, t, X)}{2} \right) = 0, \\ u(x, t = 0, X) = u_0(x, X), X \sim \mathcal{U}([-1, 1]) \\ \text{on the periodic domain } x \in [0, 1]_{\text{per}}. \end{cases} \quad (4.12)$$

For simplicity, we consider periodic boundary conditions $([0, 1]_{\text{per}})$. The initial data is supposed to be a smooth function for all X and we assume the time

$$T_X = -\frac{1}{\inf_x (\partial_x u_0(x, X))},$$

at which a discontinuous solution appears is bounded from below uniformly

$$\exists T, \quad 0 < T < T_X \quad \forall X.$$

We also assume the exact solution is smooth with respect to all variables

$$u \in L^\infty((0, 1) \times (0, T^\varepsilon) \times (-1, 1)) \cap L^\infty([0, 1]_{\text{per}} \times (0, T^\varepsilon) : H^k(-1, 1)),$$

for all $k \in \mathbb{N}$ where

$$H^k(\Omega) = \left\{ u \in L^2(\Omega) \mid \int \sum_{l=0}^k (u^{(l)})^2 d\mathcal{P}_X < \infty \right\}.$$

Let us solve this problem with gPC as described in the literature. This leads to the P -truncated hyperbolic reduced model

$$\begin{cases} \partial_t u_0^X(x, t) + \partial_x \int \frac{\left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right)^2}{2} \phi_0^X d\mathcal{P}_X = 0, \\ \dots \\ \partial_t u_P^X(x, t) + \partial_x \int \frac{\left(\sum_{k=0}^P u_k^X(x, t) \phi_k^X(X) \right)^2}{2} \phi_P^X d\mathcal{P}_X = 0. \end{cases}$$

Then the following spectral theorem holds:

Theorem 4.1 (Convergence of Burgers' approximation) *spectral accuracy holds in the following sense: if we denote by*

$$\|u(t)\|_{L^2(\mathcal{I} \times \Omega)}^2 = \int_{\mathcal{I}} \int_{\Omega} u^2(x, t, X) d\mathcal{P}_X(X) dx,$$

for all k there exists a constant D_k^ε such that

$$\|u(t) - u_P^X(t)\|_{L^2(\mathcal{I} \times \Omega)}^2 \leq D_k^\varepsilon \left(\|u(0) - u_P^X(0)\|_{L^2(\mathcal{I} \times \Omega)}^2 + \frac{1}{P^k} \right), \quad t \leq T^\varepsilon.$$

Let us illustrate the above theorem 4.1. We consider Burgers' equation (4.12) together with zero fluxes boundary conditions and a *smooth* uncertain initial condition. This choice is motivated by the fact that despite this smoothness, the dynamics of the system stiffen the problem in both the random and the physical space. At $t = 0$, we consider u is given by

$$u^0(x, X) = K_0 \mathbf{1}_{[0, x_0]}(x - \sigma X) + K_1 \mathbf{1}_{[x_1, L]}(x - \sigma X) + Q(x - \sigma X) \mathbf{1}_{[x_0, x_1]}(x - \sigma X),$$

with coefficients K_0, K_1 to be defined and $Q(x) = ax^3 + bx^2 + cx + d$. The coefficients (a, b, c, d) are given by

$$a = -2 \frac{K_0 - K_1}{x_0^3 + 3x_0x_1^2 - x_1^3 - 3x_1x_0^2}, \quad b = \frac{3(K_0 - K_1)(x_0 + x_1)}{x_0^3 + 3x_0x_1^2 - x_1^3 - 3x_1x_0^2},$$

$$c = -6 \frac{(K_0 - K_1)x_1x_0}{x_0^3 + 3x_0x_1^2 - x_1^3 - 3x_1x_0^2}, \quad d = \frac{-x_1^3K_0 + 3x_1^2K_0x_0 + K_1x_0^3 - 3K_1x_1x_0^2}{x_0^3 + 3x_0x_1^2 - x_1^3 - 3x_1x_0^2}.$$

They ensure the initial condition and its first derivatives are continuous with respect to the space and stochastic variable: in other words, $u^0(x, X)$ verifies the conditions of theorem 4.1 for $k = 3$. Several realisations of $X \rightarrow u_0(x, X)$, with $X \sim \mathcal{U}([-1, 1])$, are presented in figure 4.1 (left). The stochastic initial conditions consist in uniformly distributed translations along the x -axis of one deterministic curve. In practice, we take $L = 3$, $K_0 = 12$, $K_1 = 1$, $x_0 = 0.5$, $x_1 = 1.5$ and $\sigma = 0.2$ so that $\sigma X \in [-0.2, 0.2]$.

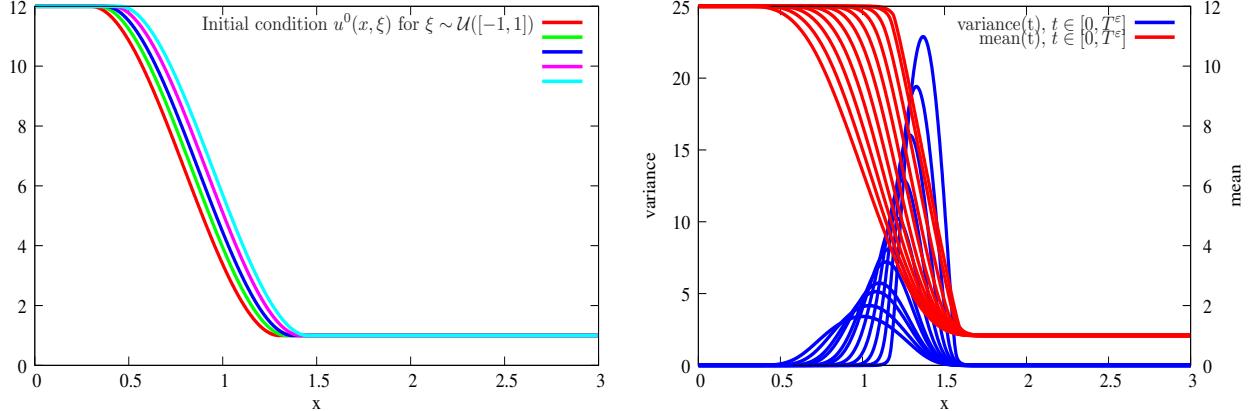


Figure 4.1: Left: initial condition $u^0(x, X)$ for several realisations of $X \sim \mathcal{U}([-1, 1])$. Right: time evolution of the mean and variance of the solution for $t \in [0, T^e]$.

For this problem, an analytical solution is available up to the critical time $T_X = \frac{3a}{b^2 - 3ac}$, which is here independent of X . The analytical solution is not reminded here but is given in [84]. For the numerical tests, we take $T^e = T_X - \varepsilon$ with $\varepsilon = 10^{-10}$.

The results are displayed in figures 4.1 and 4.2. Figure 4.1 (right) shows the time evolution of the mean and the variance with respect to the spatial variable. As time passes, the mean gets steeper and the variance increases. The computation is stopped at T^e , just before the appearance of a shock wave in both the stochastic and physical space. In figure 4.2 (left), we display the numerical solution with respect to X at point $x = 1.5$ and at different times: it represents the time evolution in the random domain at a certain point in space. We observe that the solution also gets steeper with respect to the random parameter as time increases. Figure 4.2 presents the numerical results, the relative errors in $L^2(\Omega, \mathcal{I})$ at time T^e obtained by the discretisation of the P -truncated Burgers' system with a Roe solver with, respectively, 500, 1000 and 2000 cells: spectral convergence and the result of theorem 4.1 are recovered.

Remark 4.2 *The stagnation in the final portion of the convergence curves in figure 4.2 (right) corresponds to spatial discretisation limits: it is interesting considering there is, for a given spatial discretisation Δx , an optimal choice of P . Indeed, for the 500 cells discretisation, going beyond $P = 10$ does not improve the accuracy (the error is driven by Δx) whereas it increases considerably the computational cost. With finer discretisations, the slope $\frac{1}{P^k}$ is valid for higher orders.*

In [84], numerical computations together with the proof of the above theorem are also presented. According to the material of this section, intrusive gPC gives very satisfying results in the case of scalar conservation law. To obtain an accuracy of about 10^{-5} with an MC method, we would need 10^{10} runs of a simulation code. The same accuracy is reached with the intrusive gPC method with a polynomial order $P = 7$, see figure 4.2 (bottom right), and a very fast resolution (less than a second computational time).

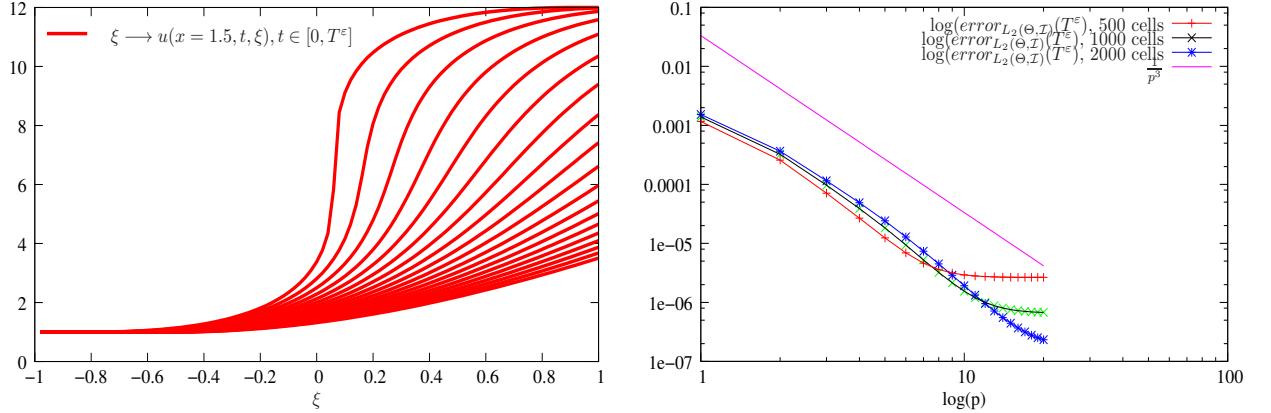


Figure 4.2: Illustration of theorem 4.1: Burgers' solution with respect to the random parameter at $x = 1.5$ for several times (left) and spectral convergence with respect to polynomial approximation order P (right).

From the previous study, we now know that for scalar conservation laws, there is no problem of wellposedness nor of accuracy. In order to understand the numerical instabilities of section 4.1, we need to consider *systems* of conservation laws ($d > 1$).

4.2.2 Possible loss of wellposedness for non-scalar systems of conservation laws

The next step consists in considering non-scalar systems of conservation laws. The first we studied in [239] was the p -system, of size $d = 2$: we showed that some closures were not satisfying enough, transforming an initially *hyperbolic* system into an only *weakly hyperbolic* reduced model. In [84], we studied the Shallow water one, also called Saint-Venant system. It describes a flow below a pressure surface. The system may be expressed as

$$\begin{cases} \partial_t h(x, t, X) + \partial_x (hv(x, t, X)) = 0, \\ \partial_t (hv(x, t, X)) + \partial_x \left(hv^2(x, t, X) + g \frac{h^2(x, t, X)}{2} \right) = 0, \end{cases} \quad (4.13)$$

where h is the water height, v the velocity of the water and $g > 0$ is the local gravity constant. In this case $d = 2$, $u = (h, hv)^t$ and $f(u) = (hv, hv^2 + g \frac{h^2}{2})$. We assume the system is uncertain, uncertainty being modeled by a scalar random variable $X \sim \mathcal{U}[-1, 1]$. Let us consider the first two Legendre polynomials

$$\phi_0^X(X) = \frac{1}{\sqrt{2}}, \quad \phi_1^X(X) = \sqrt{\frac{3}{2}}X.$$

The $P = 1$ -truncated reduced model obtained from system (4.13) can be recast as

$$\partial_t \begin{pmatrix} u_0^X \\ u_1^X \end{pmatrix} + \partial_x \begin{pmatrix} \int_{-1}^1 f(u_0^X \phi_0^X(X) + u_1^X \phi_1^X(X)) \phi_0^X(X) d\mathcal{P}_X \\ \int_{-1}^1 f(u_0^X \phi_0^X(X) + u_1^X \phi_1^X(X)) \phi_1^X(X) d\mathcal{P}_X \end{pmatrix} = 0. \quad (4.14)$$

The Jacobian matrix A of the total flux with respect to the unknown u_0^X, u_1^X is

$$A = \begin{pmatrix} \int_{-1}^1 \nabla f \phi_0^X(X) \phi_0^X(X) & \int_{-1}^1 \nabla f \phi_1^X(X) \phi_0^X(X) \\ \int_{-1}^1 \nabla f \phi_1^X(X) \phi_0^X(X) & \int_{-1}^1 \nabla f \phi_1^X(X) \phi_1^X(X) \end{pmatrix} \in \mathbb{R}^{4 \times 4}. \quad (4.15)$$

The Jacobian matrix of the Saint-Venant flux with respect to $u = (h, hv)$ is

$$\nabla f = \begin{pmatrix} 0 & 1 \\ -v^2 + gh & 2v \end{pmatrix} \in \mathbb{R}^{2 \times 2}. \quad (4.16)$$

Property 4.2 Assume that $u_0^X(x, t) = (\sqrt{2}, 0)$ and $u_1^X(x, t) = \left(0, \sqrt{\frac{2}{3}}\right)$ for arbitrary $x \in \mathcal{D}$ and $t \in [0, T]$. Then for all $0 < g < \frac{3}{25}$ the matrix A has complex eigenvalues, so the system (4.14) is not hyperbolic.

Property 4.2 proves the hyperbolicity assumption does not always hold for a reduced model of a hyperbolic system ($d > 1$) of conservation laws obtained by sG-gPC. The problematic state for the height is deterministic, $h(x, t, X) = \sqrt{2}$, $\phi_0^X(X) = 1$, and we have $hv(x, t, X) = \sqrt{\frac{2}{3}} \phi_1^X(X) = X$, implying the

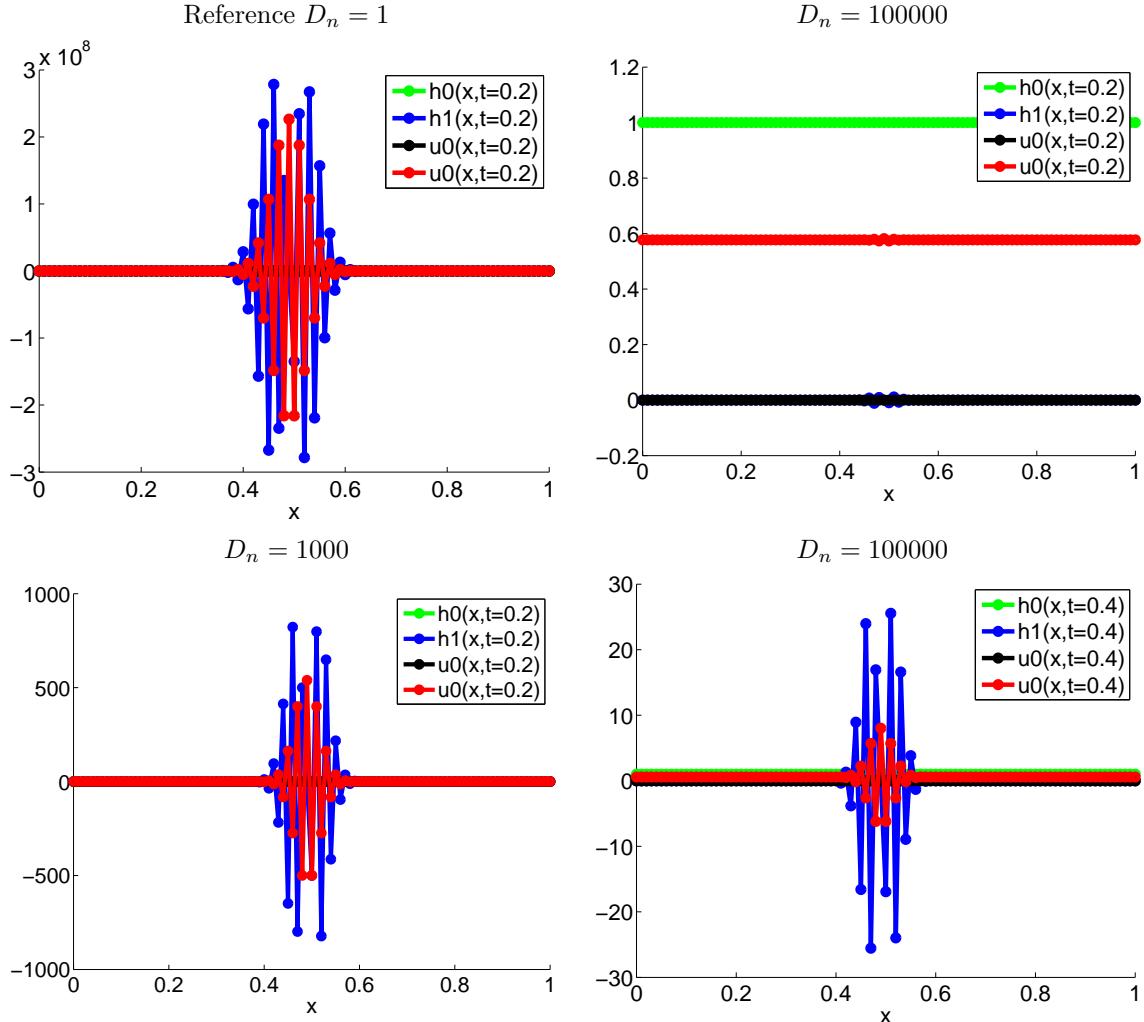


Figure 4.3: Illustration of the non-hyperbolicity of the shallow water truncated system (data are specified in Proposition 4.2). A very small germ of oscillations increases exponentially fast. The numerical diffusion coefficient is $D_n \in \{1, 10^3, 10^5\}$. Even artificially large numerical diffusion $D_n = 10^5$ is not able to control it for sufficiently large time.

problematic velocity state is uncertain, equals to $v = X$. Property 4.2 shows the construction of a well-posed intrusive gPC reduced model from a wellposed system of conservation laws is not straightforward. It may not hold for any truncation order P (the loss of hyperbolicity has only been proven for $P = 1$) or any configuration (only proven for a particular *but admissible* state). The proof is given in [84] together

with a proof of loss of hyperbolicity for the Euler system.

To end this section, we would like to emphasize what can numerically occur when one choose to deliberately solve a non-hyperbolic system (the material is the same as in [84]). Let us solve the latter truncated shallow water system ($P = 1$) with a numerical method and set the problematic state of property 4.2 $(h_0, h_1, (hv)_0, (hv)_1) = (1, 0, 0, \frac{1}{\sqrt{3}})$ with $0 < g = 0.1 < \frac{3}{25}$ as the uncertain initial condition. The analytical solution is stationary and homogeneous (i.e. constant with respect to $\forall(x, t) \in [0, 1] \times \mathbb{R}^+$). Suppose we are interested in the solution at time $T = 0.2$. The truncation order $P = 1$ should allow recovering the analytical stochastic solution $\forall(x, t, X) \in [0, 1] \times \mathbb{R}^+ \times [-1, 1]$. The numerical results are displayed in figure 4.3 in several configurations for 100 cells: on figure 4.3 (top-left), numerical instabilities appear in the center of the domain and make the solution non physical, they are even growing exponentially fast with time. This is typical of non-hyperbolic models: small perturbations (round-off errors here) are exponentially amplified with time. On figure 4.3 (bottom-left) and figure 4.3 (top-right), we revisit the same problem but we apply more and more diffusive numerical schemes. We denote by D_n the numerical diffusion coefficient (manually tuned) of our scheme. The increase in numerical diffusion artificially smoothes the solution. It even makes it look physical at the same time $T = 0.2$ on figure 4.3 (top-right), whereas it only consists in a numerical trick: if we consider the same resolution scheme as before but are interested in the solution at a later time $T = 0.4$, the small oscillations occurring at time $T = 0.2$ keep on growing (exponentially fast with time) leading to figure 4.3 (bottom-right).

In this section, we emphasized the importance of the hyperbolicity property of the P -truncated reduced model, prior to having a numerical resolution strategy. In the next section, we suggest a systematic way to close the reduced model obtained from an uncertain system of conservation laws in order to ensure hyperbolicity by construction.

4.2.3 A closure ensuring wellposedness for general systems of conservation laws

From what we learnt in the previous numerical and theoretical studies, care will be taken to build, prior to any other considerations, a hyperbolic reduced model. Only then we will authorize ourselves to solve it. In order to ease the analysis, we would like to introduce a general canvas in which both reduced models of the previous sections could be recovered. For this, we rewrite the P -truncated gPC reduced model built from a Galerkin projection of the components of a gPC basis in a more general form

$$\partial_t U(\Lambda) + \partial_x F(\Lambda) = 0. \quad (4.17)$$

In the above expression, we explicited the dependences of the main variable U with respect to an additional one, Λ , which remains to be defined at this stage of the discussion. In (4.17), the moments $U(\Lambda) = (u_0^X(\lambda_0^X, \dots, \lambda_P^X), \dots, u_P^X(\lambda_0^X, \dots, \lambda_P^X))^t$ and the moments of the flux $F(\Lambda) = (f_0^X(\lambda_0^X, \dots, \lambda_P^X), \dots, f_P^X(\lambda_0^X, \dots, \lambda_P^X))^t$ are defined by $\forall k \in \{0, \dots, P\}$

$$u_k^X = \int u^P(\lambda_0^X, \dots, \lambda_P^X) \phi_k^X d\mathcal{P}_X, \text{ and } f_k^X = \int f(u^P(\lambda_0^X, \dots, \lambda_P^X)) \phi_k^X d\mathcal{P}_X. \quad (4.18)$$

Closing reduced model (4.17) resumes to the definition of the transformation $u^P(\lambda_0^X, \dots, \lambda_P^X)$ in (4.18). At this stage, the introduction of $\Lambda = (\lambda_0^X, \dots, \lambda_P^X)^t$ may appear artificial but it allows recovering both the closure of section 4.1.1 and of section 4.1.2 for Euler system by choosing $\Lambda = (u_0^X, \dots, u_P^X)^t = U$ for the first one and $\Lambda = (z_0^P, \dots, z_P^X)^t = Z$ for the second. The choice of Λ implicitly determines the shape of u^P in (4.18) (i.e. $u^P = u$ for section 4.1.1, $u^P = z$ for section 4.1.2). Now, the first term of (4.18) defines what is commonly called a *moment problem* which can be stated as such: find u^P such that given

a vector $(u_0^X, \dots, u_P^X)^t$, u^P satisfies

$$\left\{ \begin{array}{l} \int u^P \phi_0^X d\mathcal{P}_X = u_0^X, \\ \dots \quad \dots, \\ \int u^P \phi_k^X d\mathcal{P}_X = u_k^X, \\ \dots \quad \dots, \\ \int u^P \phi_P^X d\mathcal{P}_X = u_P^X. \end{array} \right. \quad (4.19)$$

The above problem is indeterminate in the sense there is not unicity⁹ of u^P satisfying (4.19). To ensure unicity, we suggest looking for u^P as the only minimum of a strictly convex functional $u \in \mathcal{D}_u \rightarrow \theta(u)$ on \mathcal{D}_u . Problem (4.19) becomes: find $u^P \in \mathcal{D}_u$ minimizing

$$\Theta(u^P) = \int \theta(u^P) d\mathcal{P}_X, \quad (4.20a)$$

$$\left\{ \begin{array}{l} \text{under constraints} \\ \left\{ \begin{array}{l} \int u^P \phi_0^X d\mathcal{P}_X = u_0^X, \\ \dots \quad \dots, \\ \int u^P \phi_k^X d\mathcal{P}_X = u_k^X, \\ \dots \quad \dots, \\ \int u^P \phi_P^X d\mathcal{P}_X = u_P^X. \end{array} \right. \end{array} \right. \quad (4.20b)$$

Solution u^P of minimization problem (4.20), if it exists, is unique and is in \mathcal{D}_u , see [155, 198]. In [84], considerations on the existence of the solution of the moment problem are studied. It is closely related to the convexity of the state space \mathcal{D}_u (and \mathcal{D}_λ later on). We rely on [84] for these technical details and focus on the properties of the closure entropy in the following section. First, if u^P exists, the unicity comes from the strict convexity of $\Theta(u)$. The Lagrange multipliers $(\lambda_k)_{k \in \{0, \dots, P\}}$ for problem (4.20) are such that $u^P(\lambda_0, \dots, \lambda_P)$ minimizes

$$T(u(\lambda_0, \dots, \lambda_P)) = -\Theta(u(\lambda_0, \dots, \lambda_P)) + \sum_{k=0}^P \int u(\lambda_0, \dots, \lambda_P) \lambda_k \phi_k d\mathcal{P}_X - \sum_{k=0}^P u_k \lambda_k. \quad (4.21)$$

Performing a variational study of T with respect to u in (4.21) leads to

$$T(u + \delta u) - T(u) = \int \delta u \left(\sum_{k=0}^P \lambda_k \phi_k - \nabla_u \theta(u) \right) d\mathcal{P}_X + \mathcal{O}((\delta u)^2). \quad (4.22)$$

Now, T is minimal for u^P satisfying $\int \delta u \left(\sum_{k=0}^P \lambda_k \phi_k - \nabla_u \theta(u^P) \right) d\mathcal{P}_X = 0, \forall \delta u$ and consequently

$$\nabla_u \theta(u^P(\lambda_0, \dots, \lambda_P)) = \sum_{k=0}^P \lambda_k \phi_k. \quad (4.23)$$

From the strict convexity of θ , function $u \in \mathcal{D}_u \mapsto \nabla_u \theta(u) = \lambda \in \mathcal{D}_\lambda$ is invertible on \mathcal{D}_λ and we get

$$u^P(\lambda_0, \dots, \lambda_P) = (\nabla_u \theta)^{-1} \left(\sum_{k=0}^P \lambda_k \phi_k \right) \in \mathcal{D}_u. \quad (4.24)$$

⁹for example $u_1^P(X) = \sum_{k=0}^P u_k^X \phi_k^X(X)$ satisfies (4.19) and $u_2^P(X) = u_1^P(X) + \phi_{P+1}^X(X)$ too.

Once again, this is in agreement with the change of variable of section 4.1.2 in which z is developed on the gPC basis rather than u . More generally, any closure introduced *via* a closure entropy θ implies developing variable $\nabla_u \theta(u)$ on the gPC basis rather than u , cf. (4.23). In the gPC formalism presented in the previous chapter 3, the Lagrange multipliers $(\lambda_k)_{k \in \{0, \dots, P\}}$ are nothing more than the coefficients in the gPC development of $\nabla_u \theta(u)$, i.e. we have

$$\forall k \in \{0, \dots, P\}, \lambda_k = \int \nabla_u \theta(u) \phi_k d\mathcal{P}_X. \quad (4.25)$$

The closure entropy θ generalizes the changes of variables of the previous sections:

- to recover the closure of section 4.1.1, it is enough choosing $\theta(u) = \frac{u^2}{2}$. In this case, $\nabla_u \theta(u) = u$ and performing the gPC expansion of $\nabla_u \theta(u)$ or u is equivalent. Conversely, choosing $\theta(u) = \frac{u^2}{2}$ is equivalent to applying sG-gPC.
- The change of variable of section 4.1.2 corresponds to the particular choice $\nabla_u \theta(u) = (\sqrt{\rho}, \sqrt{\rho}v, \sqrt{\rho}(e - \frac{p}{\rho}))^t$. This choice of θ would be based on the *a priori* choice of being able to build a Roe scheme for system (4.17).
- The introduction of θ to close system (4.17) also opens to new possibilities. For example, it is possible to build closures based on some *a priori* knowledge of the space in which u the solution of (4.1) lives. Scalar conservation laws verifies the maximum principle. In other words, it is possible constraining the solution of the moment problem to *a priori* known bounds. The closure entropies $(\theta_i)_{i \in \{1, 2, 3, 4\}}$ defined by¹⁰

$$\begin{aligned} \theta_1(u) &= (u - u_-) \ln(u - u_-) - u + u_-, & \text{is strictly convex in } \mathcal{D}_u = I_1 =]u_-, +\infty[, \\ \theta_2(u) &= (u_+ - u) \ln(u_+ - u) - u_+ + u, & \text{is strictly convex in } \mathcal{D}_u = I_2 =]-\infty, u_+[, \\ \theta_3(u) &= \begin{pmatrix} (u - u_-) \ln(u - u_-) - u + u_- \\ +(u_+ - u) \ln(u_+ - u) - u_+ + u \end{pmatrix} & \text{is strictly convex in } \mathcal{D}_u = I_3 =]u_-, u_+[, \\ \theta_4(u) &= \frac{u^2}{2} & \text{is strictly convex in } \mathcal{D}_u = I_4 = \mathbb{R}, \end{aligned} \quad (4.26)$$

are examples of closure entropies having a unique minimum in $(\lambda_0, \dots, \lambda_P)^t$ for all $P \in \mathbb{N}$ if and only if the vector of constraints $(u_0, \dots, u_P)^t$ is realizable, see [155, 198]. If we study their respective transformation $(\nabla_u \theta_i)_{i \in \{1, 2, 3, 4\}}$, it is easy verifying they are given by

$$\begin{aligned} u_1(\lambda_0, \dots, \lambda_P) &= u_- + e^{\lambda^P} & \in]u_-, +\infty[, \\ u_2(\lambda_0, \dots, \lambda_P) &= \frac{-1 + u_+ + e^{\lambda^P}}{e^{\lambda^P}} & \in]-\infty, u_+[, \\ u_3(\lambda_0, \dots, \lambda_P) &= \frac{u_- + u_+ e^{\lambda^P}}{1 + e^{\lambda^P}} & \in]u_-, u_+[, \\ u_4(\lambda_0, \dots, \lambda_P) &= \lambda^P & \in \mathbb{R}, \end{aligned} \quad (4.27)$$

and are constrained to their respective definition domains $(I_i)_{i \in \{1, 2, 3, 4\}}$. We considered the scalar case with (4.26) but it is easy building entropies having similar properties in the non-scalar case: let $\mathcal{D}_u = I_{i_1} \times \dots \times I_{i_n}$ where $(I_{i_k})_{k \in \{1, \dots, n\}}$ be convex subsets of \mathbb{R} such that $i_k = 1, 2, 3$ or 4. Then $\theta_f = \sum_{k=1}^n \theta_{i_k}$ is strictly convex in \mathcal{D}_u function of $\theta_{i_k}, k \in \{1, \dots, n\}$ strictly convex on the $I_{i_k}, \forall k \in \{1, \dots, n\}$ where $i_k = 1, 2, 3$ or 4.

The gains one can obtain with the above examples of entropies are highlighted in [236, 232]. Improvements have even been made in [162]. Let us now focus on the nonlinear functional we choose to minimize to close system (4.17). We define the *adjoint closure variable* together with the *adjoint closure entropy* as follows.

¹⁰the results are valid $\forall u_-, u_+ \in \mathbb{R}$.

Definition 4.2 (Adjoint closure variable) *The adjoint closure variable λ is defined by*

$$\lambda : u \in \mathcal{D}_u \longmapsto \nabla_u \theta(u) \in \mathcal{D}_\lambda \subset \mathbb{R}^n,$$

where θ is the chosen closure entropy, strictly convex in \mathcal{D}_u .

From the strict convexity of θ , transformation $\lambda \in \mathcal{D}_\lambda \rightarrow u(\lambda) \in \mathcal{D}_u$ defines a bijection from \mathcal{D}_λ to \mathcal{D}_u . With the previous definition, it is convenient introducing the *adjoint closure entropy*.

Definition 4.3 (Adjoint closure entropy) *The adjoint closure entropy for system (4.38) associated to the closure entropy θ is given by θ^* such that*

$$\theta^* : \lambda \in \mathcal{D}_\lambda \subset \mathbb{R}^n \longmapsto -\theta(u(\lambda)) + \langle u(\lambda), \lambda \rangle \in \mathbb{R}. \quad (4.28)$$

Function θ^* is also strictly convex in \mathcal{D}_λ . Indeed, it is the Legendre transformation of θ which is, by hypothesis, strictly convex in \mathcal{D}_u . The new entropy θ^* will considerably ease the next calculations: it allows a concise expression for the inverse of $u \in \mathcal{D}_u \longmapsto \nabla_u \theta(u) = \lambda \in \mathcal{D}_\lambda$.

Property 4.3 (Inverse of $u \in \mathcal{D}_u \longmapsto \nabla_u \theta(u) = \lambda \in \mathcal{D}_\lambda$) *The inverse of $u \in \mathcal{D}_u \longmapsto \nabla_u \theta(u) = \lambda \in \mathcal{D}_\lambda$ is given by*

$$\lambda \in \mathcal{D}_\lambda \longmapsto \nabla_\lambda \theta^*(\lambda) = u(\lambda) \in \mathcal{D}_u. \quad (4.29)$$

Proof It is easy verifying

$$\begin{aligned} \nabla_\lambda \theta^*(\lambda) &= - \left\langle \nabla_\lambda u(\lambda), \underbrace{\nabla_u \theta(u(\lambda))}_{=\lambda} \right\rangle + \langle \nabla_\lambda u(\lambda), \lambda \rangle + u(\lambda), \\ \nabla_\lambda \theta^*(\lambda) &= u(\lambda). \end{aligned} \quad (4.30)$$

■

Now, let us go back to functional T and refine its expression with respect to $(\lambda_k)_{k \in \{0, \dots, P\}}$: from the previous analysis, it can be rewritten

$$T(\lambda_0, \dots, \lambda_P) = - \int \theta \left(u \left(\sum_{k=0}^P \lambda_k \phi_k \right) \right) d\mathcal{P}_X - \sum_{k=0}^P \langle u_k, \lambda_k \rangle_n + \sum_{l=0}^P \int \left\langle u \left(\sum_{k=0}^P \lambda_k \phi_k \right), \lambda_l \phi_l \right\rangle_n d\mathcal{P}_X. \quad (4.31)$$

In (4.31), $\lambda^P = \sum_{k=0}^P \lambda_k \phi_k$ with $\forall k \in \{0, \dots, P\}$ $\lambda_k = \int (\nabla_u \theta) \phi_k d\mathcal{P}_X$. With the above notation, (4.31) can be equivalently rewritten

$$T(\Lambda) = -\Theta(U(\Lambda)) - \langle U, \Lambda \rangle_{n \times (P+1)} + \langle U(\Lambda), \Lambda \rangle_{n \times (P+1)}. \quad (4.32)$$

It is equivalent to

$$T(\Lambda) = \Theta^*(\Lambda) - \langle U, \Lambda \rangle_{n \times (P+1)}, \quad (4.33)$$

where Θ^* is the adjoint entropy of Θ . From the above expression, it is easy verifying the following property.

Property 4.4 *Functional T is strictly convex.*

Proof Recall θ^* denotes the adjoint closure entropy relative to θ , then a simple calculation of the Jacobian matrix of T is given by

$$\nabla_\Lambda T(\Lambda) = \nabla_\Lambda \Theta^*(\Lambda) - U = \int \nabla_\lambda \theta^*(\lambda^P) \begin{pmatrix} \phi_0 \\ \dots \\ \phi_P \end{pmatrix} d\mathcal{P}_X - \begin{pmatrix} u_0 \\ \dots \\ u_P \end{pmatrix}. \quad (4.34)$$

Its Hessian is given by

$$\nabla_{\Lambda,\Lambda}^2 T(\Lambda) = \nabla_{\Lambda,\Lambda}^2 \Theta^*(\Lambda) = \int \nabla_{\lambda,\lambda}^2 \theta^*(\lambda^P) \cdot \begin{pmatrix} \phi_0 \phi_0 & \dots & \phi_0 \phi_P \\ \dots & \phi_i \phi_j & \dots \\ \phi_P \phi_0 & \dots & \phi_P \phi_P \end{pmatrix} d\mathcal{P}_X. \quad (4.35)$$

Then, for every vector $X = (x_0, \dots, x_P)^t \in \mathbb{R}^{n \times (P+1)}$

$$\begin{aligned} \langle X, \nabla_{\Lambda,\Lambda}^2 T(\Lambda) X \rangle_{n \times (P+1)} &= \langle X, \nabla_{\Lambda,\Lambda}^2 \Theta^*(\Lambda) X \rangle_{n \times (P+1)} \\ &= \int \langle \Pi^P x, \nabla_{\lambda,\lambda}^2 \theta^*(\lambda^P) \Pi^P x \rangle_n d\mathcal{P}_X > 0. \end{aligned} \quad (4.36)$$

The positiveness comes from the strict convexity of θ^* . Consequently, Θ^* is strictly convex and so T is. ■

From the above properties, if a minimum of T exists (realizability of the constraints), it is unique.

The new closure method for reduced model (4.17) consists in choosing a closure entropy θ , strictly convex, and solving a minimization problem under constraints on the moments of the main variable. It implicitly introduces a new variable, $\nabla_u \theta(u)$ which is developed on the gPC basis (rather than the main variable u). Until this point, we only generalized the change of variable of section 4.1.1 and section 4.1.2 which were not satisfactory enough with respect to hyperbolicity. At a pinch, we put forward the possibility to ensure by construction the respect of certain principles¹¹ for the solution of (4.1). The question now is: does it help regarding our hyperbolicity considerations?

Recall we are dealing, by hypothesis, with *physical* systems of conservation laws (introductory part of this chapter). Such systems have an entropy-entropy flux pair (s, g) satisfying (4.2)–(4.3). Furthermore, in kinetic theory [207, 63], variable $\nabla_u s(u)$, i.e. choosing $\theta = s$, is called the *entropic variable* and plays an important role in the closure of M_n models for example (see chapter 1). We study its properties in the following section. In [207] (p. 29-32), the entropic variable v is introduced and corresponds to the particular choice $v = \lambda$ (directly related to the choice $\theta = s$). It verifies the following property.

Property 4.5 (Symmetrizability) *The entropic variable v symetrizes system (4.1).*

Proof For smooth solutions, (4.1) becomes

$$\partial_t u + \partial_x f(u) = \nabla_{v,v} s^*(v) \partial_t v + \nabla_{v,v} g^*(v) \partial_x v = 0, \quad (4.37)$$

where $\nabla_{v,v} s^*(v)$ and $\nabla_{v,v} g^*(v)$ are symmetric (hessian matrices). The strict convexity of s^* derives from its definition with respect to s (Legendre transform). ■

The idea now, as described previously, is to develop the entropic variable on the gPC basis $(\phi_k^X)_{k \in \{0, \dots, P\}}$ together with defining the polynomial moments of v as $v_k = \int v \phi_k d\mathcal{P}_X, \forall k \in \{0, \dots, P\}$. We will have $v \approx v^P = \sum_{k=0}^P v_k \phi_k$. The truncated reduced model can now be defined thanks to the moments of v

$$\partial_t \begin{pmatrix} u_0(v_0, \dots, v_P) \\ \dots \\ u_P(v_0, \dots, v_P) \end{pmatrix} + \partial_x \begin{pmatrix} f_0(v_0, \dots, v_P) \\ \dots \\ f_P(v_0, \dots, v_P) \end{pmatrix} = 0. \quad (4.38)$$

The above system is closed minimizing (4.21) with $\theta = s$ (i.e. $(v_k)_{k \in \{0, \dots, P\}}$ minimizes s under constraints $(u_0, \dots, u_P)^t$). In other words, we almost directly applied the *extended thermodynamic of moments* closure (see [207]) at order P to the uncertain hyperbolic system of conservation laws (4.1). It is closely related to the construction of a M_n model (see chapter 1) with entropy s being different from the Shannon entropy. We can characterise the dependences of the moments $(u_k)_{k \in \{0, \dots, P\}}$ and the moments of the

¹¹maximum principle with the examples of (4.26).

flux $(f_k)_{k \in \{0, \dots, P\}}$ with respect to $(v_k)_{k \in \{0, \dots, P\}}$:

$$\forall k \in \{0, \dots, P\}, \quad u_k = \int \nabla_v s^*(v^P) \phi_k d\mathcal{P}_X = \int u(v^P) \phi_k d\mathcal{P}_X. \quad (4.39)$$

For the flux of the truncated reduced model, we have

$$\forall k \in \{0, \dots, P\}, \quad f_k = \int \nabla_v g^*(v^P) \phi_k d\mathcal{P}_X = \int f(u(v^P)) \phi_k d\mathcal{P}_X. \quad (4.40)$$

From the above expressions of the moments (4.39) and of the moments of the flux (4.40), we can verify the following theorem.

Theorem 4.2 (Hyperbolicity of the truncated reduced model (4.38)) *The vector of moments $V = (v_0, \dots, v_P)^t$ of the entropic variable v , symmetrises the truncated reduced model (4.38) $\forall P \in \mathbb{N}$ under condition $v^P \in \mathcal{D}_v$ (realizability, see [84]).*

Proof Suppose the solution of (4.38) are smooth and denote by $V = (v_0, \dots, v_P)^t$ the vector of moments of the entropic variable v . Then (4.38) is equivalent to

$$\nabla_{V,V}^2 S^*(V) \partial_t V + \nabla_{V,V}^2 G^*(V) \partial_x V = 0. \quad (4.41)$$

In (4.41), the general terms in the symmetric (hessian) matrices $\nabla_{V,V}^2 S^*(V)$ and $\nabla_{V,V}^2 G^*(V)$ are given by

$$\nabla_{V,V}^2 S^*(V)_{i,j} = \int \nabla_{v,v}^2 s^*(v^P) \phi_i \phi_j d\mathcal{P}_X, \text{ and } \nabla_{V,V}^2 G^*(V)_{i,j} = \int \nabla_{v,v}^2 g^*(v^P) \phi_i \phi_j d\mathcal{P}_X. \quad (4.42)$$

It remains to show $\nabla_{V,V}^2 S^*(V)$ is definite positive. Let $X = (x_0, \dots, x_P)^t$ be a vector of $\mathbb{R}^{n \times (P+1)}$, then

$$\langle X, \nabla_{V,V}^2 S^*(V) X \rangle_{n \times (P+1)} = \int \langle \Pi^P x, \nabla_{v,v}^2 s^*(\Pi^P v) \Pi^P x \rangle_n d\mathcal{P}_X > 0.$$

The positiveness comes from the strict convexity of s^* . Of course, this result is valid if $v^P \in \mathcal{D}_v$. Under such conditions, the system is symmetrisable, hence hyperbolic. ■

The previous proof introduces the couple (S^*, G^*) which is very useful in practice to implement the resolution of the the reduced model together with its closure (see [236]).

Remark 4.3 *We insist the application of extended thermodynamic of moments to close a gPC based reduced model ensures the construction of a wellposed reduced model even if the basis of multiplicators $(\phi_k^X)_{k \in \mathbb{N}}$ is not polynomial or not orthogonal. In fact, the less restrictive hypothesis on the family of multiplicators to apply the previous methodology is that it should be pseudo-Haar (cf. [155] and the references therein). It implies the linear dependence in L^2 . In practice, in this document, the multiplicators are an orthonormal polynomial basis of L^2 .*

The material of this section was inspired from [232, 84]. Several other points are tackled in the same publications. For example, the technical discussion on the numerical schemes and minimization algorithm necessary to solve (4.17) together with closure (4.21). We also study the mathematical structure of the built truncated system in [84], the behaviour of the characteristic waves of the uncertain entropy closure (4.17) with respect to the ones of (4.1). This behaviour is hinted at in the numerical computations of next paragraph in which the 'fil rouge' configuration is considered with the newly built hyperbolic reduced model.

4.2.4 Application to the 'fil rouge' problem of chapter 2

To finish this chapter, we suggest applying the uncertain entropy closure to the Euler system and to our 'fil rouge' configuration. We do not spend time on the technical details such as the numerical scheme, the optimization algorithm in order to obtain U from V , they are given in [236, 84].

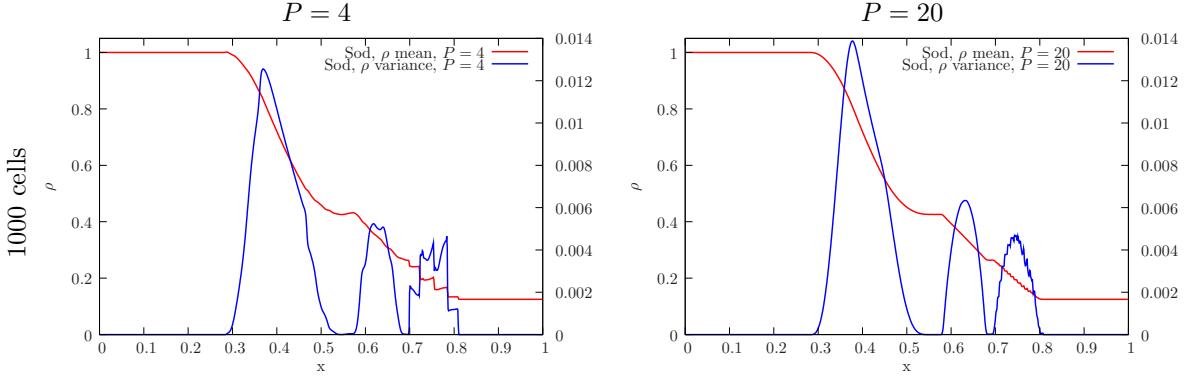


Figure 4.4: Illustration of the existence of multiple waves, see [84]. The phenomenon is especially emphasized for the mean in the vicinity of the probable shock region ($x \in [0.7, 0.8]$) where $P + 1 = 5$ discontinuities are visible on the left picture and $P + 1 = 21$ discontinuities are visible on the right one.

Figure 4.4 presents the results obtained in term of mean and variance profiles of the mass density at time $t = 0.14$. Figure 4.4 (left) presents the results for $P = 4$ and figure 4.4 (right) for $P = 20$. Those figures must be compared to figure 2.2 (bottom right) of chapter 2. The main differences with the reference solutions of chapter 2 are in the vicinities of the shock ($x \in [0.7, 0.8]$) and of the contact discontinuities ($x \in [0.55, 0.69]$). If we focus on the vicinity of the shock ($x \in [0.7, 0.8]$), $P + 1 = 5$ discontinuities can be identified on figure 4.4 (left) and $P + 1 = 21$ discontinuities on figure 4.4 (right) for the mean. They are also identifiable on the variance. The built reduced models (for $P = 4$ and $P = 20$) do not allow recovering the smooth behaviours¹² of the mean and variance profiles of the mass density as displayed in figure 2.2. The appearance of $d \times (P + 1)$ waves for a system of conservation laws of such size is classical, see [81, 260, 81, 261]. In fact, the observed smoothness in the vicinity of the contact discontinuity (for $P = 20$ but not for $P = 4$) may even be relative to the choice of the spatial discretisation: with less cells, numerical diffusion artificially smoothes the profiles. Note that figure 4.4 also allows emphasizing that the waves in the vicinity of the interface or the shock behave differently (sharp discontinuities for the shock, smoother behaviour for the interface) with respect to the numerical resolution. The study of the nature (linearly degenerate, genuinely nonlinear) of the waves of the P -truncated reduced model is complicated in general as the size of the system makes the analytical expressions of the eigenvectors of the Jacobian of the flux hard to obtain. In [84], we propose a short study of those waves.

As explained in chapter 2, the mean and variance are not always relevant probabilistic quantities, especially for systems of conservation laws developing discontinuous solutions. We suggest considering the more local observables presented in chapter 2: the pdfs of the mass density at time $t = 0.14$ and three spatial locations in the vicinities of the rarefaction fan, the interface and the shock. The Monte-Carlo references are the same as in figure 2.3 of chapter 2. The results obtained with intrusive gPC (entropy closure) are displayed in figure 4.5 for $P = 5$ and figure 4.6 for $P = 20$ in term of histograms and functional representations. Figure 4.5 for $P = 5$ first shows that in the vicinity of the rarefaction fan, the reduced model is very efficient in the sense it allows recovering exactly the pdf of the mass density at this location and time. On another hand, in the vicinities of the interface and of the shock, the Gibbs phenomenon is clearly identifiable, in term of functional representation but also in term of histograms. The two Dirac masses are not captured for this order $P = 5$. Note that mainly because of these two waves for this configuration, classical intrusive gPC approaches (i.e. non hyperbolic ones) fail to give any results (see section 4.1.3). Here, we can nonetheless verify the functional representations do not exactly behave as a polynomial (but rather as a nonlinear function of a polynomial, see [84]). This nonlinear representation ensures positive mass densities with respect to X even in presence of steep gradients but does not necessarily implies accuracy.

Remark 4.4 *In fact, on problems where the classical intrusive gPC approach and the entropy closure reduced model can be compared (typically scalar conservation laws and Burgers' equation) we can observe*

¹²In [264], the authors show that those quantities are smooth with respect to the spatial variable x for Riemann problems.

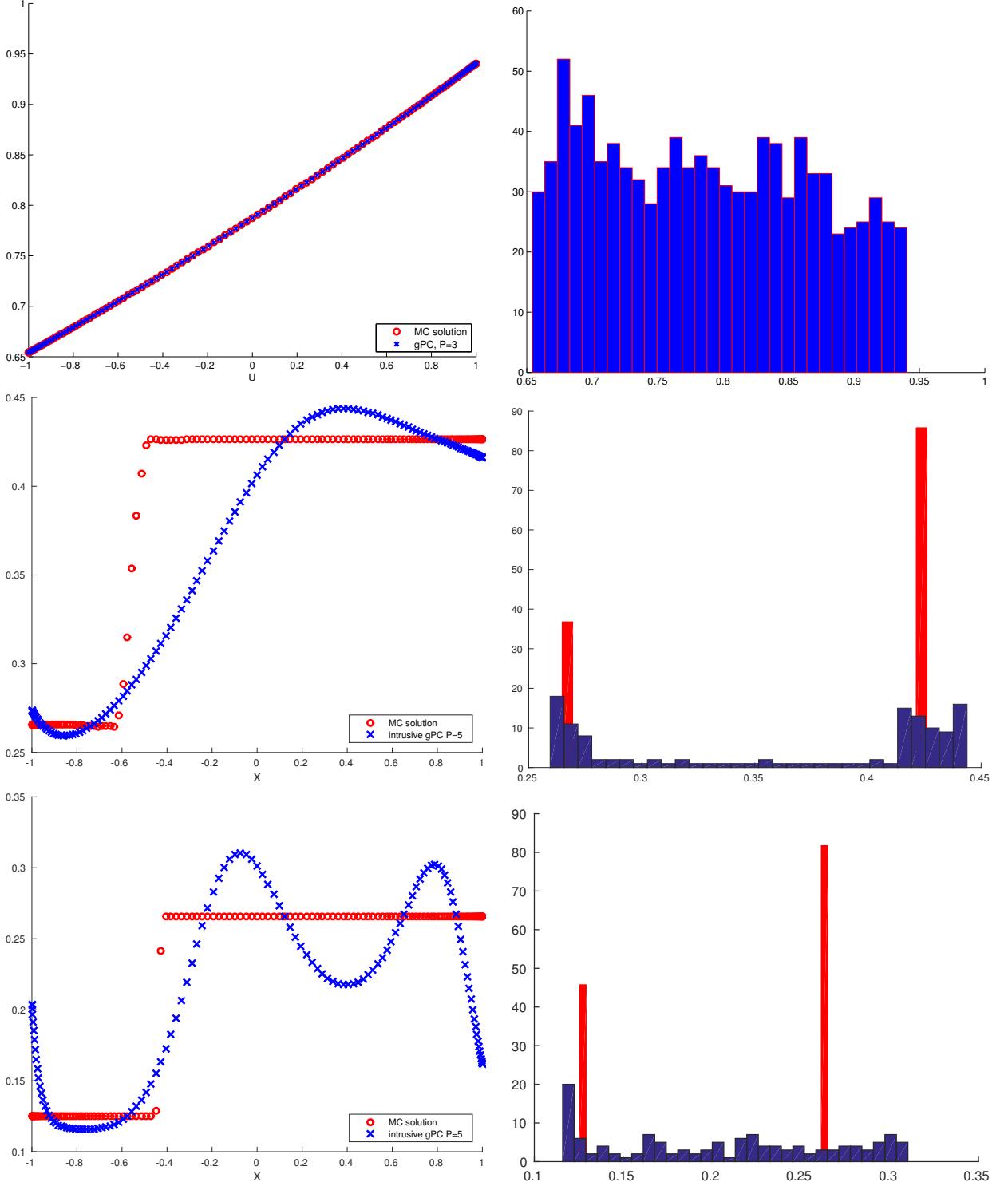


Figure 4.5: Pdfs (left) and functional representation in the random space of the mass density at $t = 0.14$ in the vicinities of the rarefaction fan ($x = 0.38$), the interface ($x = 0.61$) and the shock ($x = 0.73$) for the reduced model of order $P = 5$.

a better convergence rate for the entropy closure reduced model with respect to the classical one (see [236] and [162]). The gain for non-scalar conservation laws such as the Euler one has only been experimentally noticed via the robustness of the developed simulation code.

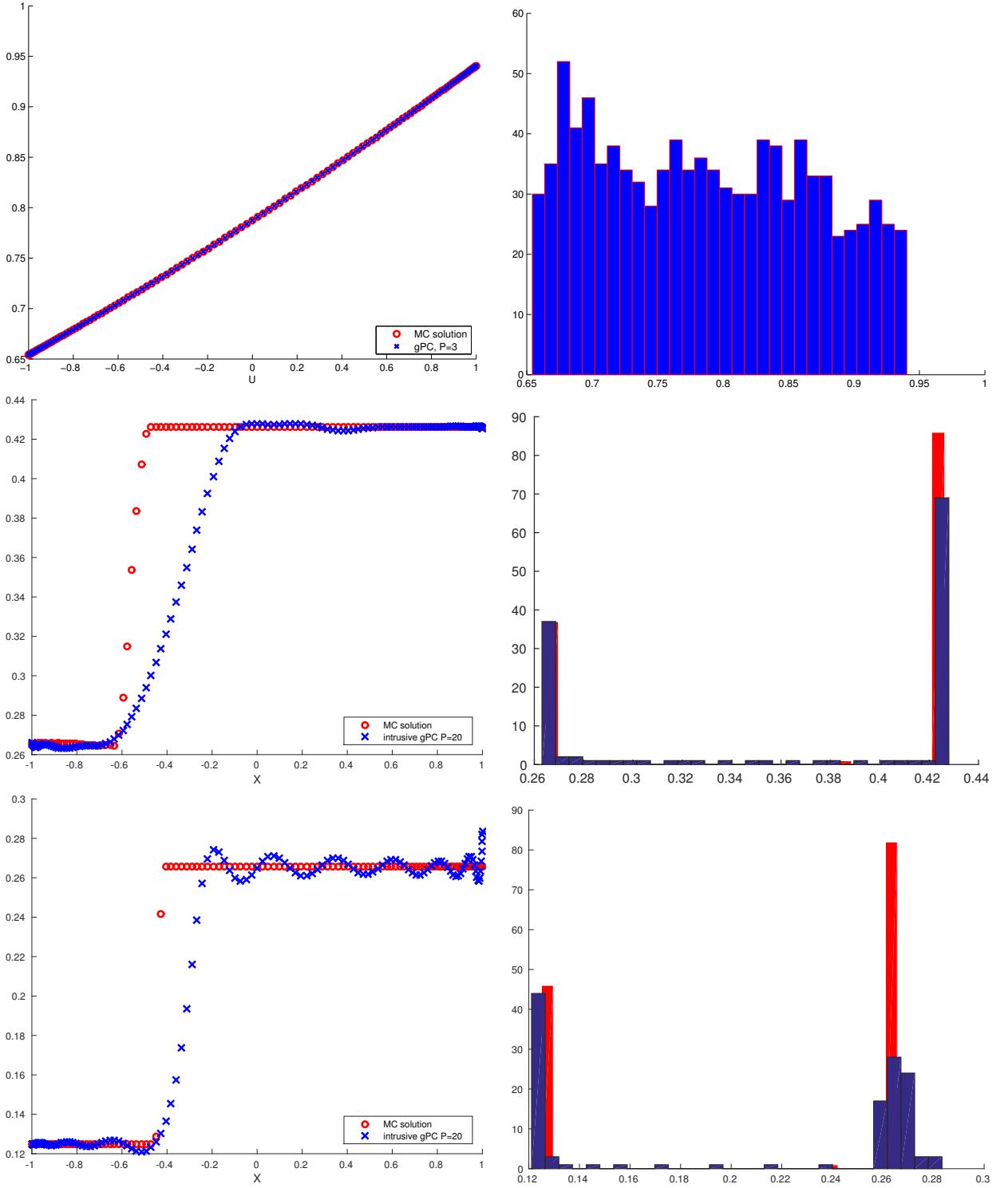


Figure 4.6: Pdfs (left) and functional representation in the random space of the mass density at $t = 0.14$ in the vicinities of the rarefaction fan ($x = 0.38$), the interface ($x = 0.61$) and the shock ($x = 0.73$) for the reduced model of order $P = 20$.

Figure 4.6 presents similar profiles but with a higher polynomial order $P = 20$. Once again, the results are very good in the vicinity of the rarefaction fan. The results are improved, with respect to figure 4.5 for $P = 5$, in the vicinities of the interface and the shock but the Gibbs phenomenon remains

clearly identifiable. It is nonetheless important noticing the behaviour of the pdf of the mass density in the vicinity of the interface: the Dirac masses are very well captured by the nonlinear approximation. The functional representation at the same location testifies the approximation does not behave like a polynomial as the states are quite well captured. In the vicinity of the shock, the discrete behaviour is captured but one of the two Dirac mass is smoothed and the approximation may need an even higher polynomial order P (with all the difficulties it implies, see section 3.4) in order to recover the second state.

4.3 Summary for intrusive gPC and the entropy closure reduced models

The material of this chapter mainly presented the work accomplished during my PhD thesis [232, 236, 80, 239, 237] and one additional publication [84]. Care has been taken to make the chapter complementary to the publications. The results are applied to a new configuration ('fil rouge' of chapter 2), common to every approximation methods detailed in this part II. To sum up the intrusive application of gPC, it implies the construction of a reduced model under truncation hypothesis and some additional ones in the nonlinear case (i.e. when the flux f is nonlinear). Building wellposed reduced models with intrusive gPC from *non-scalar conservation laws* is not straightforward. In practice, solving non-hyperbolic reduced model leads directly to a lack of robustness of the developed simulation code and non physical solutions. Nonetheless, a hyperbolic reduced model can be built applying the truncation hypothesis to the entropic variable (which symetrizes the system). The system may appear harder and more costly to close/solve (in practice) but robustness is ensured. Note that a better accuracy is also observed in some configurations in which both methodologies are comparable (this is not detailed in this manuscript but emphasized in [236]). The closure hypothesis has also been applied to more complex configurations, instable physical flows (Richtmyer-Meshkov shock tubes as in chapter 7) in 2D with strong shocks in [237]. The entropy closed reduced models are robust and efficient. At first glance, their application seems to be restricted to physical system (i.e. systems of conservation laws having a couple entropy-entropy flux (s, g)). In fact, having a physical system ensures one can choose a closure entropy (θ) ensuring the existence of a couple entropy-entropy flux for the reduced model (by taking $\theta = s$). In the case the system does not have such a couple, it is still possible to close the reduced model with an arbitrary closure entropy θ (i.e. a chosen strictly convex functional of the conservative variable). Tests have been performed and interesting results have been obtained. For Euler system, several closure entropies (different from the mathematical one) have been tested in different configurations: they yield acceptable, robust results. This study will probably be the purpose of further publications. Now those closure entropies do not ensure the theoretical wellposedness of the reduced model. This observation may be an argument in the sense that the existence of a mathematical entropy is probably strong assumption and lighter ones may be enough. Note that this latter observation and the questions arising constitute an open problem in the field of systems of conservation laws.

The closure strategy presented in the last sections of this chapter (and mainly in [236, 232, 84]) is inspired from extended thermodynamic of moments [207]. Recent work showed different analogies with kinetic theory can lead to different resolution strategies bearing interesting properties: in [83], the authors build a kinetic scheme (see the example of chapter 1 with (1.32)) for uncertain scalar conservation laws and ensure the respect of the maximum principle for the approximated uncertain solution. Oscillations are constrained and the Gibbs phenomenon is controlled by construction. Paper [162] goes beyond the material of the previous chapter (and of [84, 236]) for scalar conservation laws. It puts forward a very interesting improvement to the entropy closure method leading to considerable gains in accuracy and toward a better respect of the maximum principle. Papers [163, 164] are also very interesting, for several reasons. First, connections to kinetic theory are highlighted and new methods inspired from the latter fields (filtering) are adapted for uncertainty quantification. Second, the numerical results on Euler system confirm the observation the closure method presented in the previous chapter yields costly but accurate results. Finally, they also confirm that preserving some relevant invariants may be enough to obtain accurate and reliable results. This invariant preservation is done *via* filtering whereas in the previous chapter it is done *via* the choice of a closure entropy different from the mathematical one ($\theta \neq s$). In

[257], the authors also aim at producing hyperbolic (wellposed) reduced models. They rely on scheme limitations, also inspired from kinetic theory. Their approach is also numerically compared to the closure method of this chapter on several benchmarks (M_1 model for radiative transfer mainly). Once again, even if more costly, the closure method yields very accurate results.

Those two previous examples¹³ certainly show uncertainty quantification can benefit the methodologies and progresses of kinetic theory. On another hand, the reverse surely also applies. For modeling, for example: in chapter 1 and section 3.1.3, we put forward Grad's 13 moment model can be understood as a Polynomial Chaos development. Its generalized counterpart (gPC) may lead to the construction of more relevant models in specific regimes of interest. Or take the example of P_n models for the resolution of the linear Boltzmann equation. Legendre polynomials are introduced, implicitly assuming the regime of interest is isotropic (for the angular distribution): the first order captures the isotropic regime, the higher orders compute corrections to this regime. Suppose one is interested in another identified anisotropic angular regime. Then it is possible building the gPC basis associated to it and perform the Galerkin projection with respect to this newly built basis. These tracks will certainly be explored in further (modeling) research. More practically, in the next part III, gPC is introduced to reduce variance in Monte-Carlo computations (see section 9.12 of chapter 9).

In the next chapter, we investigate the non-intrusive counterpart of gPC. We describe its application, as detailed in the literature, and apply it to our 'fil rouge' configuration.

¹³extended thermodynamic of moments of this chapter and the kinetic methodology of [83, 257, 162, 163, 164].

Chapter 5

Non-Intrusive application of gPC for systems of conservation laws

An S_n -like (without collision term) gPC based model

Contents

5.1	Non-intrusive application of gPC	73
5.2	Choice of the experimental design (the most common ones for UQ)	74
5.2.1	The Monte-Carlo (MC) integration method	75
5.2.2	Low discrepancy sequences/Quasi Monte-Carlo	76
5.2.3	Gauss quadrature rules	77
5.2.4	Clenshaw-Curtis (CC) quadrature rule	81
5.2.5	MC vs. Quasi MC vs. Gauss vs. etc.	82
5.3	Integration vs. Regression vs. Collocation vs. Kriging	83
5.3.1	Regression-gPC approximations	83
5.3.2	Collocation-gPC approximation	89
5.3.3	Kriging-gPC approximations	91
5.4	Few other applications of gPC	97
5.4.1	Application to the 'fil rouge' problem of chapter 2	97
5.4.2	Integration vs. Regression vs. Collocation vs. Kriging vs. discontinuity	98
5.5	Summary for non-intrusive gPC for systems of conservation laws	101

In this chapter, we present the second way to compute the coefficients $(u_k)_{k \in \{0, \dots, P\}}$. It is referred to as *non-intrusive* in the literature. Once again, the methodology is general and can be applied to any system. But we here focus on systems of conservation laws. We recall general uncertain conservation laws can be written in the form

$$\begin{cases} \partial_t u(x, t, X) + \partial_x f(u(x, t, X)) = 0, \\ u(x, 0, X) = u_0(x, X). \end{cases} \quad (5.1)$$

We suppose system (5.1) is hyperbolic $\forall X \in \Omega$. Now, applying non-intrusive gPC implies the use of a deterministic simulation code (often called *black-box code* in the literature) solving (5.1) at several chosen (*and licit with respect to the wellposedness of the system*) points $(X_i)_{i \in \{1, \dots, N\}}$. It consequently allows avoiding the problem of building non-hyperbolic systems (encountered in chapter 4). All along the previous chapter, we referred to analogies with P_n (section 4.1) and M_n (section 4.2.3) models for the resolution of the linear Boltzmann equation. The non-intrusive counterpart is closer to the S_n model [16, 92, 61] for solving the linear Boltzmann equation except the collision term is zero (hence no time-dependent

coupling between the angles in example (1.31)) and the resolutions are fully independent. The coupling occurs mainly at the observation times and positions at which the pointwise discretised solutions are postprocessed to build a gPC approximation. This will be detailed in the following sections. The reader, all along the description of the methodology, may notice, through the questions arising, through the references, that the non-intrusive application of gPC leads naturally toward the field of *approximation theory* rather than *model reduction* as it was the case in chapter 4.

In the next sections, we detail the application and the characteristics of the non-intrusive resolution of system (5.1). We apply it to our 'fil rouge' problem and analyse its behaviour. Finally, we pave the path toward a new non-intrusive approach we developed (see [238, 31, 242]), based on non-intrusive gPC and moment theory.

5.1 Non-intrusive application of gPC

In this section, we present a brief state-of-the-art for non-intrusive gPC. The description of the methodology may, at first glance, look like a recipe but it is representative of its practical use. Let us first suppose the random variable¹ X has probability measure $d\mathcal{P}_X$. The methodology consists in several steps:

1. it begins by the construction of the gPC basis $(\phi_k^X)_{k \in \mathbb{N}}$. It is orthonormal with respect to the inner product defined by the probability measure $d\mathcal{P}_X$ of the input random variable X . In other words, it is such that

$$\int \phi_k^X \phi_t^X d\mathcal{P}_X = \delta_{k,t}, \forall (k,t) \in \mathbb{N}^2.$$

This step is common to intrusive gPC and is described in section 3.4.

2. The second step corresponds to the discretisation of the random variable and its probability measure $(X, d\mathcal{P}_X)$ by a numerical integration method with N points:

$$(X, d\mathcal{P}_X) \approx (X_i, w_i)_{i \in \{1, \dots, N\}}. \quad (5.2)$$

We detail how the points are usually chosen in section 5.2.

3. The next step consists in running N independent runs of a black-box code at the *a priori* chosen points $(X_i, w_i)_{i \in \{1, \dots, N\}}$ and gathering a new collection of output points²:

$$(u(x, t, X_i), w_i)_{i \in \{1, \dots, N\}} = (u(X_i), w_i)_{i \in \{1, \dots, N\}}. \quad (5.3)$$

This step is supposed to bear the main computational effort. Up to this point, the methodology can directly be compared to the S_n model without collision term, described in chapter 1.

4. Once the N runs obtained, the rest is only postprocessing at the observation points of interest (several times t and positions x for example in our 'fil rouge' problem). The estimation of the polynomial coefficients is mainly made by numerical integration in this document $\forall k \in \{0, \dots, P\}$. This means we have

$$u_k^X = \int u(X) \phi_k^X(X) d\mathcal{P}_X \approx u_k^{X,N} = \sum_{i=1}^N u(X_i) \phi_k^X(X_i) w_i. \quad (5.4)$$

Many authors apply other numerical methods to compute the coefficients (such as regression, collocation, kriging). They are addressed in section 5.3 together with their analysis.

5. Finally, one can reconstruct the truncated polynomial approximation (or the collocation or kriging ones) using the approximated coefficients (5.4)

$$u(X) \approx u_{P,N}^X(X) = \sum_{k=0}^P u_k^{X,N} \phi_k^X(X).$$

¹or vector.

²We here introduce an abusive notation as we drop the dependences with respect to x and t .

It then remains to perform the desired post-treatments in order to approximate the statistical quantities of interest (mean, variance, histograms, etc.) related to $u(X)$.

At the end of the process, one has access to an approximation³ $u_{P,N}^X(X)$. The error between $u(X)$ and $u_{P,N}^X(X)$ in the L^2 -norm can be decomposed in two main parts using the orthonormality of the gPC basis⁴:

$$\begin{aligned} \|u(X) - u_{P,N}^X(X)\|_{L^2}^2 &= \left\| \sum_{k=0}^{\infty} u_k^X \phi_k^X(X) - \sum_{k=0}^P u_k^{X,N} \phi_k^X(X) \right\|_{L^2}^2, \\ &= \underbrace{\sum_{k=0}^P (u_k^X - u_k^{X,N})^2}_{\text{integration error}} + \underbrace{\sum_{k=P+1}^{\infty} (u_k^X)^2}_{\text{truncation error}}. \end{aligned} \quad (5.5)$$

In (5.5), the error of the non-intrusive approximation have (explicitly) two parameters:

- N for the integration error,
- and P for the truncation error.

One needs to have in mind there is still a hidden parameter which corresponds to the choice of the discretisation/accuracy for the N runs of the black-box code (for example Δx and/or Δt for a deterministic resolution scheme⁵ or N_{MC} for a stochastic one as in part III for example). This additional discretisation error may be important in practice and has to be taken into account. This is reminded and illustrated in the application to our 'fil rouge' problem later on and also in section 9.11 of the next part III.

The above description of the methodology is quite simple but to be complete, two points remain to be tackled:

- we did not explain under which considerations the experimental design is chosen in practice for the discretisation (5.2) of the couple $(X, d\mathcal{P}_X)$. Section 5.2 presents some brief descriptions of the possibilities at hand, their advantages and drawbacks. They correspond to the most common strategies in uncertainty quantification.
- Some authors in the literature do not exactly use numerical integration methods in order to estimate the polynomial coefficients (5.4). Amongst the other possibilities one can cite *collocation methods*, *regression* or *kriging*. Their subtleties are briefly investigated, analysed and illustrated in section 5.3.

The two following sections address the two above points.

5.2 Choice of the experimental design (the most common ones for UQ)

The choice of the discretisation of the random variable X together with its probability measure $d\mathcal{P}_X$ obviously directly impacts the quality of the gPC approximation as testifies (5.5). Depending on the transformation u , the integration error may be preponderant with respect to the truncation one (see for example [191]). The set of points and weights (5.2) are often called an experimental design. Considerations to help choose it together with pedagogical examples can be found in [107, 15]. We here suggest a brief overview of the most commonly applied strategies in uncertainty quantification studies.

³The superscript X reminds of the approximation basis, P of the truncation order and N of the number of points for the numerical approximation of the coefficients $(u_k)_{k \in \{0, \dots, P\}}$.

⁴assuming there are no errors in the gPC basis, see section 3.4

⁵for example, the intrusive counterpart of chapter 4 counted two parameters, P and Δx as illustrated on figure 4.2, remark 4.2 of section 4.2.1.

Notation (5.2) for the punctual discretisation of $(X, d\mathcal{P}_X)$ is general and has been chosen in order to show that the material of the rest of the chapter concerning non-intrusive gPC (and even the material of chapters 6, 7 and 8) can be applied independently of this choice. Suppose the points $(X_i)_{i \in \{1, \dots, N\}}$ are chosen sampled from the probability law of X and $(w_i = \frac{1}{N})_{i \in \{1, \dots, N\}}$, then it corresponds to the Monte-Carlo integration method for the estimation of the coefficients. With the same writing, we can conveniently consider Gauss quadrature points, Latin Hypercube Samples, Sparse Grids etc. The latter sets of points in dimension Q differ only by their asymptotic error analysis with respect to integration. We insist we here implicitly deal with *converging* discretisation of $(X, d\mathcal{P}_X)$: without this assumption, most of the following results do not hold.

In the following sections, we detail different integration methods on a general function $g : X \in \Omega \rightarrow g(X) \in \mathbb{R}$. Of course, in a non-intrusive gPC context, $g \in \{u\phi_0^X, \dots, u\phi_P^X\}$ as we aim at applying the *same* experimental design to every coefficients $(u_k^X)_{k \in \{0, \dots, P\}}$. For some of the below descriptions, we may find convenient considering $X \sim \mathcal{U}[0, 1]^Q$. But we insist uniform variable on the square (i.e. independent) can be mapped into any random variable. We can consequently carry out the integration with respect to arbitrary inner products.

5.2.1 The Monte-Carlo (MC) integration method

We here recall the principles of the Monte-Carlo method for integration. These can be found in many books (see [268, 173, 256, 165] for example). We briefly mention them for the completeness of the talk. The reader familiar with the notion can easily jump to the next sections. The MC integration method corresponds to the particular experimental design for which the points $(X_i)_{i \in \{1, \dots, N\}}$ are randomly sampled from the probability law of X with uniform weights $(w_i = \frac{1}{N})_{i \in \{1, \dots, N\}}$. Two theorems are at the basis of MC integration.

Theorem 5.1 (Law of Large Numbers) *let $X \in L^1(\Omega)$ be a random vector of size Q . Let $(X_i)_{i \in \mathbb{N}}$ be some independent identically distributed (i.i.d.) random vectors with X . Then⁶*

$$\bar{X}_N = \frac{1}{N} (X_1 + \dots + X_N) \xrightarrow[N \rightarrow \infty]{a.s.} \mathbb{E}[X]. \quad (5.6)$$

This first theorem is a convergence result. The following one is stronger in the sense it describes its convergence rate.

Theorem 5.2 (Central Limit Theorem) *let $X \in L^2(\Omega)$ be a random vector of size Q with variance σ^2 . Let $(X_i)_{i \in \mathbb{N}}$ be some i.i.d. random vectors with X . Let \bar{X}_n be a random vector as in (5.6), then⁷*

$$\sqrt{N} (\bar{X}_N - \mathbb{E}[X]) \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{G}(0, \sigma^2). \quad (5.7)$$

In the expression above, $\mathcal{G}(0, \sigma^2)$ denotes a gaussian random variable of mean 0 and variance σ^2 .

Let us go back to our gPC coefficients and consider $g \in \{u\phi_0^X, \dots, u\phi_P^X\}$ as integrands. The Central Limit Theorem [256, 210] ensures

$$\sqrt{N} \left| \int g(X) \mathcal{P}_X - \sum_{i=1}^N g(X_i) w_i \right| \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{G}(0, \sigma^{g,Q}).$$

The convergence rate is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ and is independent of the smoothness of the transformation $g(X)$ nor the size Q of the random vector X . In fact, the regularity of g and the size Q of the random vector only affect the constant $\sigma^{g,Q}$. This constant can be *a posteriori* evaluated and is an estimation of the error *under condition its estimator⁸ is unbiased*, see [256]. For the MC method, the points $(X_i)_{i \in \{1, \dots, N\}}$ can be generated independently. Adding points always improves the quality of the experimental design. In other words, one can enrich *a posteriori* the experimental design with a new set of $N' \in \mathbb{N}$ points

⁶The convergence is almost surely, see [256].

⁷The convergence is in law (\mathcal{L}), see [256].

⁸This point is detailed and illustrated in section 9.12 and we do not recall the notion here to avoid redundancies.

$(X'_i)_{i \in \{1, \dots, N'\}}$. Experimental design

$$\left((X_i)_{i \in \{1, \dots, N\}} \bigcup (X'_j)_{j \in \{1, \dots, N'\}}, \left(w_j = \frac{1}{N + N'} \right)_{j \in \{1, \dots, N + N'\}} \right), \quad (5.8)$$

is still a Monte-Carlo integration method with asymptotic error $\mathcal{O}\left(\frac{1}{\sqrt{N+N'}}\right)$, independently of the choice of N' . This property is singular amongst the integration technics.

It is possible trading *sensitivity to dimensionality and regularity* for *accuracy* by using for example *low discrepancy sequences* [209, 210, 53]. They are described in the following section.

5.2.2 Low discrepancy sequences/Quasi Monte-Carlo

In this section, we assume $X \sim \mathcal{U}([0, 1]^Q)$. The discrepancy of a sequence is said low [53, 54] if the proportion of points in the sequence falling into an arbitrary set B is close to proportional to the measure of B . The aim is to avoid clustering. Specific definitions of discrepancy differ regarding the choice of B (hyperspheres, hypercubes, etc.) and how the discrepancy for every B is computed (usually normalized) and combined (usually by taking the worst value). Amongst the low discrepancy sequences, the most famous ones are the Halton, the Van Der Corput, the Faure, the generalized Faure, the Niederreiter and the Sobol sequence, see [134, 283, 267, 209, 210].

Their asymptotic error analysis can be described assuming g is of bounded variation on $[0, 1]^Q$, i.e.

$$V(g) = \int_{[0,1]^Q} |g(x)| d\mathcal{P}_X(x) < \infty. \quad (5.9)$$

With this property, we have [210]

$$\left| \int g(X) d\mathcal{P}_X - \sum_{i=1}^N g(X_i) w_i \right| = V(g) D_N(X_1, \dots, X_N),$$

where $D_N(X_1, \dots, X_N)$ is the discrepancy of the sequence of point $(X_i)_{i \in \{1, \dots, N\}}$. It is defined by

$$D_N(X_1, \dots, X_N) = \sup_{I \subset [0,1]^Q} \left| \frac{1}{N} \sum_{i=1}^N \mathbf{1}_I(X_i) - V_I \right|.$$

In the above expression, $V_I = \int \mathbf{1}_I(x) dx$ is the volume of I and $\sum_{i=1}^N \mathbf{1}_I(X_i)$ is in fact the number of points of $(X_i)_{i \in \{1, \dots, N\}}$ in I . Note that in general,

- $V(g)$ is not known,
- and has nothing to do with the variance σ of g .
- Finally, D_N is hard to estimate.

Nevertheless, one can show [53] that the convergence rate

- is at best $V(g)D_N \leq \frac{\log^k(N)}{N}$ with $k < Q$,
- at worst $V(g)D_N \leq \frac{\log^Q(N)}{N}$ depending on the smoothness of the integrand.
- For example for a function with second derivative Lipschitz-continuous, the worst case convergence rate becomes $V(g)D_N \leq \frac{\ln^{\frac{Q-1}{2}}(N)}{N^{\frac{3}{2}}}$, see [54] (p. 306).

The convergence rate remains faster than $\frac{1}{\sqrt{N}}$ for relatively small values of Q . For efficiency with respect to MC points, the number of dimensions needs to remain low and N should be large. In opposition

to the MC experimental design, the construction of a low discrepancy one must remain sequential for efficiency: it is difficult adding points to a first low discrepancy experimental design without degrading its properties (i.e. increasing its discrepancy).

In this document, Monte-Carlo methods are more completely studied in part III, mainly for the resolution of the linear and nonlinear Boltzmann equations rather than for uncertainty quantification problems. In this part II, we focus on deterministic integration strategies as the ones described below.

5.2.3 Gauss quadrature rules

The N Gauss points of any arbitrary measure $d\mathcal{P}_X$ are the roots of the $(N + 1)^{th}$ degree polynomial orthonormal with respect to the inner product defined by $d\mathcal{P}_X$. The Jacobi matrix J_N^X , see section 3.4, is symmetric and consequently diagonalizable. Its N eigenvalues are the N Gauss points $(X_i)_{i \in \{1, \dots, N\}}$. They verify $\forall i \in \{1, \dots, N\}$

$$\begin{aligned} X_i \Phi_N^X(X_i) &= J_N^X \Phi_N^X(X_i) + \underbrace{\sqrt{\beta_N} \phi_{N+1}^X(X_i)}_{=0} e_P, \\ X_i \Phi_N^X(X_i) &= J_N^X \Phi_N^X(X_i). \end{aligned}$$

The $(\Phi_N^X(X_i))_{i \in \{1, \dots, N\}}$ are the (unnormalized) eigenvectors. There are several ways to introduce the weights of the Gauss quadrature rule. For example, if we introduce $p_N(x)$ an arbitrary polynomial of exact degree N and express it in term of Lagrange polynomials⁹ $(L_i(X))_{i \in \{1, \dots, N\}}$, defined at the Gauss points $(X_i)_{i \in \{1, \dots, N\}}$, we have

$$p_N(x) = \sum_{i=1}^N p_N(X_i) L_i(x).$$

By definition of the quadrature rule, we have

$$\begin{aligned} \int p_N(x) d\mathcal{P}_X(x) &= \int \left(\sum_{i=1}^N p_N(X_i) L_i(x) \right) d\mathcal{P}_X(x), \\ &= \sum_{i=1}^N p_N(X_i) \int L_i(x) d\mathcal{P}_X(x), \\ &= \sum_{i=1}^N p_N(X_i) w_i. \end{aligned}$$

It allows identifying

$$\forall i \in \{1, \dots, N\}, w_i = \int L_i(x) d\mathcal{P}_X(x). \quad (5.10)$$

Introducing the weights as above implicitly presents them as the coefficients ensuring exact integration of polynomials up to order N . Such N -point quadrature rule is said to have *degree of exactness* N and is denoted as *interpolatory*, see [117]. Obviously, from the definition of the weights (5.10), given any N points, any quadrature rule can be made interpolatory. Such definition is convenient but not optimal: the optimal N -point quadrature rule has degree of exactness $2N$ and is called a Gauss quadrature rule [117]. It verifies (5.10) but also [117]

$$\forall i \in \{1, \dots, N\}, w_i = \int L_i(x) d\mathcal{P}_X(x) = \int L_i^2(x) d\mathcal{P}_X(x). \quad (5.11)$$

⁹They are polynomials of degree N verifying $L_i(X_j) = \delta_{i,j}, \forall (i, j) \in \{1, \dots, N\}^2$.

The weights can also be defined by normalizing every eigenvectors $(\Phi_N^X(X_i))_{i \in \{1, \dots, N\}}$ and taking their first squared component: as $\phi_0^X(x) = 1$ in our case, dealing with probability measures, we have

$$\forall i \in \{1, \dots, N\}, w_i = \frac{1}{\sum_{k=0}^P (\phi_k^X(X_i))^2}.$$

The two last definition of the Gauss weights emphasizes their positiveness, important in practice for robustness. There exists other ways to define them but we do not aim at being exhaustive on the subject: depending on the theoretical problem of interest, one formulation may be more interesting than another. The main drawback of Gauss quadratures remains linked to the curse of dimensionality. Building a multidimensional Gauss quadrature rule implies tensorizing the points in each directions (just as was done for the gPC basis in section 3.5.1). The number of Gauss points increases exponentially fast with the dimension Q . Regarding asymptotic error analysis, for a Gauss quadrature rule, we have the following general property (see [117]):

$$\left| \int g(x) d\mathcal{P}_X - \sum_{i=1}^N g(X_i) w_i \right| = \int H_{2N}(X, g) d\mathcal{P}_X. \quad (5.12)$$

The term $H_{2N}(x, g)$ denotes the Hermite interpolation polynomial of order $2N$ relative to function g . We recall the Hermite interpolation polynomials relative to the points $(X_i)_{i \in \{1, \dots, N\}}$ satisfy

$$H_{2N}(X_i, g) = g(X_i), \text{ and } H'_{2N}(X_i, g) = g'(X_i), \forall i \in \{1, \dots, N\}.$$

If furthermore, g is $2N$ times differentiable then the same asymptotic error can be expressed in term of ξ , existing in the support of the probability measure $d\mathcal{P}_X$, such that

$$\left| \int g(x) d\mathcal{P}_X - \sum_{i=1}^N g(X_i) w_i \right| = \frac{\langle \phi_N^{X,m}, \phi_N^{X,m} \rangle}{2N!} g^{(2N)}(\xi). \quad (5.13)$$

In (5.13), $g^{(n)}$ denotes the n^{th} derivative of g . The constant in the error analysis strongly depends on the smoothness (of order $2N$) of the integrand. The coefficients $(\langle \phi_k^{X,m}, \phi_k^{X,m} \rangle)_{k \in \mathbb{N}}$ corresponds to the norm of the monic orthogonal polynomial $(\phi_k^{X,m})_{k \in \{0, \dots, P\}}$ associated to the probability measure $d\mathcal{P}_X$. This normalization coefficient may also depend strongly on N : for Legendre polynomials¹⁰ for example, $\langle \phi_N^{L,m}, \phi_N^{L,m} \rangle = \frac{1}{2N+1}$. Obviously, for smooth solutions, the convergence rate is fast. The Gauss quadrature rules for an arbitrary measure $d\mathcal{P}_X$ are designed/defined [7] to ensure a good accuracy up to order $2N$ for the statistical moments of the input X i.e. up to order $2N$ for polynomials (hence the non optimality of the definition (5.10)). With the above remark in mind, assume g is polynomial of order $2N-1$, then $g^{(n)} = 0, \forall n \geq 2N$ and the quadrature rule is exact, see (5.12). In dimension Q with $X = (X_1, \dots, X_Q)^t$, due to the tensorisation of the points, the above asymptotical error becomes

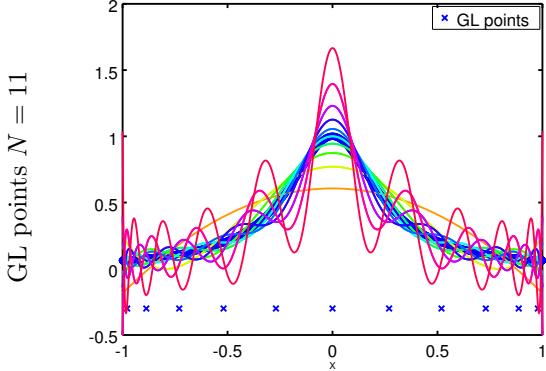
$$\left| \int g(x) d\mathcal{P}_X - \sum_{i=1}^N g(X_i) w_i \right| = \max_{i \in \{1, \dots, Q\}} \left[\frac{\langle \phi_{N_i}^{X_i,m}, \phi_{N_i}^{X_i,m} \rangle}{2N_i!} g_i^{(2N_i)}(\xi_i) \right]. \quad (5.14)$$

In (5.14), $(N_i)_{i \in \{1, \dots, Q\}}$, $(\phi_k^{X_i})_{k \in \mathbb{N}, i \in \{1, \dots, Q\}}$ and $(g_i^{(2N_i)})_{i \in \{1, \dots, Q\}}$ are respectively the number of points in each direction, the gPC basis in each direction and the $2N_i^{th}$ derivative of g in each direction.

Let us come back to our final goal. We want to apply the quadrature rule to function g having the particular form $g = u\phi_l^X$ with $l \in \{0, \dots, P\}$ to build a gPC approximation of order P . Formally, we can expand $u \in L^2$ on the gPC basis, i.e. $u = \sum_{k=0}^{\infty} u_k^X \phi_k^X$, and introduce $g_l = \sum_{k=0}^{\infty} u_k^X \phi_k^X \phi_l^X$,

¹⁰with $d\mathcal{P}_X(x) = \frac{1}{2} \mathbf{1}_{[-1,1]}(x) dx$.

gPC approximations for $3 \leq P \leq 20$



Convergence w.r.t. P for fixed N

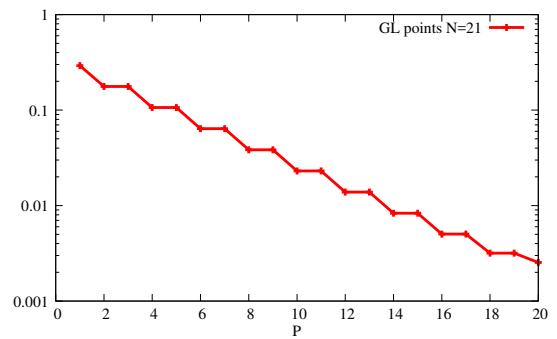
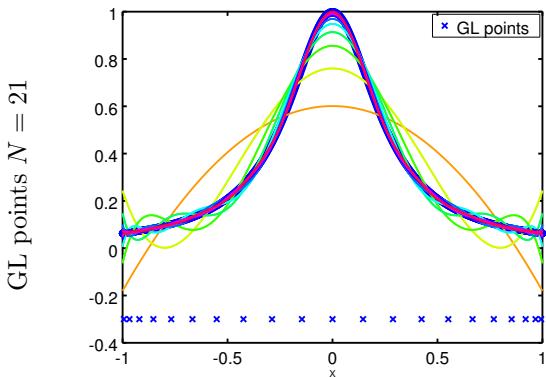
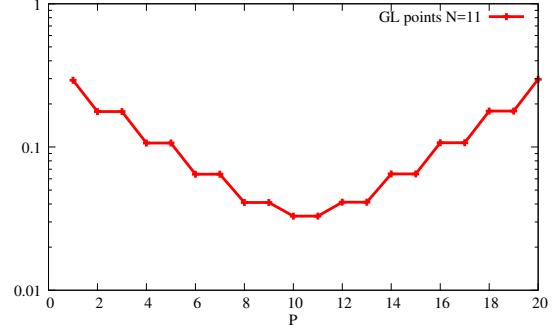


Figure 5.1: Application of Gauss-Legendre quadrature rule for the *integration* of the gPC coefficients of the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N .

$\forall l \in \{0, \dots, P\}$. With such notations, error analysis (5.13) becomes

$$\begin{aligned} \left| \int g_l(x) d\mathcal{P}_X - \sum_{i=1}^N g_l(X_i) w_i \right| &= \frac{\langle \phi_N^{X,m}, \phi_N^{X,m} \rangle}{2N!} \sum_{k=0}^{\infty} u_k^X (\phi_k^X \phi_l^X)^{(2N)}(\xi_l), \\ \left| u_l^X - u_l^{X,N} \right| &= \frac{\langle \phi_N^{X,m}, \phi_N^{X,m} \rangle}{2N!} \sum_{k=0}^{\infty} u_k^X \mathbf{1}_{k+l \geq 2N} (\phi_k^X \phi_l^X)^{(2N)}(\xi_l). \end{aligned} \quad (5.15)$$

Expression (5.15) comes from the fact (5.13) holds for every gPC coefficients $\forall l \in \{0, \dots, P\}$. Each $(\xi_l)_{l \in \{0, \dots, P\}}$ echoes ξ as defined earlier for each $(g_l)_{l \in \{0, \dots, P\}}$. To study (5.15) more in detail, recall $\forall k \in \mathbb{N}$ we can rewrite $\phi_k^X(X) = \Gamma_k^X \phi_k^{X,m}(X) = \Gamma_k^X \prod_{i=1}^k (X - X_i^k)$ with $(\phi_k^{X,m})_{k \in \mathbb{N}}$ the monic orthogonal polynomial relative to $(\phi_k^X)_{k \in \mathbb{N}}$ and with $(X_i^k)_{i \in \{1, \dots, k\}}$ its roots¹¹. Due to the fact we decomposed $(\phi_k^X)_{k \in \mathbb{N}}$ as a product of monomials, the previous notation allows rewriting

$$(\phi_k^X \phi_l^X)^{(2N)}(\xi_l) = 2N! \Gamma_k^X \Gamma_l^X P_{k+l-2N}(\xi_l).$$

In the above expression, P_{k+l-2N} is a monic polynomial of order $k + l - 2N$. With the above equality,

¹¹The roots of the orthogonal polynomials are within the support of the probability measure $d\mathcal{P}_X$, distinct and real, see [117]. Nevertheless, the decomposition could be done in the complex plane.

(5.15) becomes

$$\begin{aligned}
|u_l^X - u_l^{X,N}| &= \sum_{k=0}^{\infty} u_k^X \mathbf{1}_{k+l \geq 2N} \frac{\Gamma_k^X \Gamma_l^X}{(\Gamma_N^X)^2} P_{k+l-2N}(\xi_l). \\
|u_l^X - u_l^{X,N}|^2 &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} u_k^X u_j^X \mathbf{1}_{j+l \geq 2N} \mathbf{1}_{k+l \geq 2N} \frac{\Gamma_j^X \Gamma_l^X (\Gamma_l^X)^2}{(\Gamma_N^X)^4} P_{k+l-2N}(\xi_l) P_{j+l-2N}(\xi_l). \\
\sum_{l=0}^P |u_l^X - u_l^{X,N}|^2 &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} u_k^X u_j^X \frac{\Gamma_k^X \Gamma_j^X}{(\Gamma_N^X)^2} \underbrace{\sum_{l=0}^P \mathbf{1}_{j+l \geq 2N} \mathbf{1}_{k+l \geq 2N} \left(\frac{\Gamma_l^X}{\Gamma_N^X} \right)^2 Q_{k,j,l,N}}_{\text{term } (*) \text{ we can control by choosing } P}.
\end{aligned} \tag{5.16}$$

The last term in the left hand side of (5.16) is exactly the L^2 -norm of the integration error in (5.5). The term $(*)$ in (5.16) corresponds to the term we can control by choosing P , independently of any smoothness assumptions for u . Now, for a given k , we have (cf. section 3.4)

$$\Gamma_k^X = \sqrt{\frac{H_{2(k-1)}^X}{H_{2k}^X}} \geq 2^{2k-1}.$$

Now, assume

- N grow and P is fixed: then $(*)$ is bounded and the integration error is $\mathcal{O}(\frac{1}{(\Gamma_N^X)^2}) = \mathcal{O}(\frac{1}{2^{2N-1}})$ ensuring a fast convergence rate for Gauss points. Besides, term $(*)$ in (5.16) can be controloed by making sure $l \leq N, \forall l \in \mathbb{N}$, i.e. choosing $P \leq N$, ensuring the minimization of the residue in (5.15) independently of the smoothness of the solution u .
- In the opposite case, a gPC reconstruction with increasing P for fixed N (exponentially) accumulates errors as $(*)$ becomes $\mathcal{O}((\Gamma_P^X)^2) = \mathcal{O}(2^{2P-1})$.

The explosion of the L^2 error with P as soon as $P \geq N$ (second point above) has been observed on many numerical experiments, see [191, 72] for example. The above numerical analysis shows it is *independent of $d\mathcal{P}_X$ and of u* for the associated Gauss quadrature rule.

We suggest illustrating this behaviour on a simple uncertainty propagation problem: consider the transformation of a uniform random variable X via the Runge function

$$X \sim \mathcal{U}_{[-1,1]} \longrightarrow \frac{1}{1 + 15X^2}. \tag{5.17}$$

We build a non-intrusive gPC approximation with coefficients integrated thanks to a N points Gauss-Legendre (GL) quadrature rule. Figure 5.1 presents the results of the study with $N = 11$ and $N = 21$ points. The top right picture presents a convergence study with respect to P for $N = 11$ GL points. The L^2 -norm of the error first decreases exponentially fast (logarithmic scale for the ordinate) before *increasing almost as quickly* after $P = 11 = N$. The explosion of the error after $P = N$ is in agreement with the previous numerical analysis (cf. (5.16) and the discussion below). The quality of the obtained gPC approximations can be observed on figure top left in the same conditions. For the bottom pictures of figure 5.1, care has been taken to keep $P \leq N = 21$ GL points. The exponential convergence of the gPC approximation is ensured up to $P = 20$ as testifies the bottom right picture: the increasing quality with P is observable on figure 5.1 (bottom-left) with a less and less oscillating approximation. Note that the readability of the left column of figure 5.1 can be discussed: this column is mainly qualitative and its sense will be revealed mainly when tackling section 5.3 and comparisons with the approximations obtained with regression, collocation, kriging in the same conditions.

Note that in agreement with the latter analysis and example, in every of our applications in the next chapters, we always keep $P \leq N$ in each stochastic directions.

The construction of a Gauss quadrature rule is sequential in the sense one can not in general *a posteriori* add a set of points to a given Gauss quadrature rule with N points in order to enrich it. The

roots of the orthonormal polynomials of degree N associated the measure $d\mathcal{P}_X$, does not in general have the same roots as the one of degree $N + 1$ for example, see [273, 5, 117]. This is particular for the Chebyshev polynomials and of practical interest: their use for numerical integration is presented in the next paragraph through the introduction of the Clenshaw-Curtis quadrature rule.

5.2.4 Clenshaw-Curtis (CC) quadrature rule

The Clenshaw-Curtis (CC) quadrature rule is well-known for interpolation approximations on a bounded interval $[a, b]$. The points of the experimental design are related to the roots of the Chebyshev polynomials denoted by $(\phi_k^C)_{k \in \mathbb{N}}$. Chebyshev polynomials are orthonormal with respect to the probability measure

$$d\mathcal{P}_C(x) = \frac{2}{\pi\sqrt{1-x^2}} dx,$$

of the Arcsinus random variable. The roots of the Chebyshev polynomials have two convenient properties: first, analytical formulae are available. Second, the roots of the Chebyshev polynomial of degree P are also roots of the Chebyshev polynomial of degree $2P$. The points of the CC quadrature rule of level k have $n_k = 2^{k-1} + 1$ points and are given by

$$x_j^k = -\cos\left(\frac{\pi(j-1)}{n_k-1}\right). \quad (5.18)$$

The corresponding weights are given by

$$\begin{aligned} w_1^k &= \frac{1}{n_k(n_k-2)}, && \text{for } j = 1, \\ w_j^k &= \frac{2}{n_k-1} \left(1 + 2 \sum_{l=1}^{\lfloor \frac{n_k-1}{2} \rfloor} \frac{1}{1-4l^2} \cos\left(\frac{2\pi(j-1)l}{n_k-1}\right) \right), && \text{for } 2 \leq j \leq n_k - 1. \end{aligned} \quad (5.19)$$

The CC quadrature rule of level k (i.e. having n_k points) presented above allows integration with respect to the uniform probability measure on $[-1, 1]$. Such weights are designed to integrate exactly polynomials of degree at most $n_k + 1$ with respect to the uniform measure. In practice, comparisons between Gauss points and CC ones for a given number/level n_k/k with respect to the uniform measure are of equivalent accuracy on arbitrary g , see [277]. The CC points being Gauss points (for the Chebyshev polynomials), their asymptotical error analysis is the same as above in the case of an integration with respect to the Arcsinus measure.

We here want to focus on the second property of the roots of the CC quadrature rule: the roots of ϕ_k^C are also the roots of ϕ_{2k}^C , for $n \in \mathbb{N}$. This implies a CC experimental design of level k can be *a posteriori* enriched with level $k + 1$ reusing the previous runs from the k^{th} level. The experimental design can not be enriched adding one point after another: the previous description implies a size of experimental design multiplied by 2 at each level.

Such property makes the quadrature rule very attractive for

- integration with respect to arbitrary measures. Any random variable can be mapped into a new one provided their respective cumulative density functions (and their inverse). For integration on \mathbb{R} , the Clenshaw-Curtis points can be modified into the Fejer [288] quadrature rule which do not have nodes at 1 and -1 (values which shall be mapped to ∞ and $-\infty$ otherwise).
- adaptive algorithms, see for example [72, 34]. One may decide to run simulations quadrature level after quadrature level and add points only in the stochastic directions needing it (by detecting steep gradients using the differential between two quadrature levels for example as in [72, 34]). We do not aim at being exhaustive on adaptive and sparse technics, we rely on [72, 191, 34] and the reference therein.

In this part II of this document, Gauss quadrature rules are intensively applied. This choice is based mainly on the observation that the construction of an arbitrary gPC basis needs a good accuracy on the moments of the measure of interest (see section 3.4). Practical considerations and experiments showed

that the Gauss points outpower every other integration rule based on this criterion. Note that quadrature rules with negative weights such as the ones built from a Smolyak [263] procedure showed poor efficiency in the same context. MC methods are also very robust but remain the purpose of part III.

5.2.5 MC vs. Quasi MC vs. Gauss vs. etc.

In the previous sections, several experimental designs have been presented. Their asymptotical numerical analysis have been recalled. But we are still far away from being able to choose the relevant experimental design in the relevant situation¹². In this section, we suggest a general way to choose an experimental design to compute an arbitrary integral under some basic assumptions.

Let us begin by the introduction of a common notation. For each experimental design of the previous sections, we rewrite their asymptotical error ϵ with respect to N and Q in the more concise notation

$$\epsilon = \epsilon(\sigma, N, Q). \quad (5.20)$$

The notation is general. For example, for an MC experimental design, see section 5.2.1, we have

$$\epsilon = \epsilon(\sigma_{\text{MC}}, N, Q) = \frac{\sigma_{\text{MC}}}{\sqrt{N}}.$$

For an uniform integration we have

$$\epsilon = \epsilon(\sigma_{\text{unif}}, N, Q) = \sigma_{\text{unif}} \max_{i \in \{1, \dots, Q\}} [N_i^{-1}].$$

For a general low discrepancy sequence, see section 5.2.2, we have

$$\begin{aligned} \epsilon = \epsilon(\sigma_{\text{LHS}}, N, Q) &= \sigma_{(\text{best}) \text{ LHS}} \frac{\log^k(N)}{N}, \text{ with } k = Q - 1 \text{ which may be optimistic in general,} \\ &= \sigma_{(\text{worst}) \text{ LHS}} \frac{\log^Q(N)}{N}, \\ &= \sigma_{(\text{Lipschitz}) \text{ LHS}} \frac{\ln^{\frac{Q-1}{2}}(N)}{N^{\frac{3}{2}}}. \end{aligned}$$

For a Gauss quadrature in Q dimensions, see section 5.2.3, we have

$$\epsilon = \epsilon(\sigma_{\text{Gauss}}, N, Q) = \sigma_{\text{Gauss}} \max_{i \in \{1, \dots, Q\}} \left[\frac{1}{2N_i!} \right].$$

With the few above lines, we summed-up the previous asymptotical analysis of the different experimental designs into four lines exhibiting:

- the constant σ_{name} of the method multiplying,
- the convergence rate depending on N and Q .

Let us comment on the first above point: depending on the method, the estimation of the constant σ_{name} of the method is not straightforward. For an MC experimental design, it is enough computing the variance, see (5.7). For a low discrepancy sequence, one needs to estimate (5.9). For a Gauss quadrature, one would need to estimate a derivative of order $2N$ of the integrand, see (5.13). The estimation of these constants can be complex and problem dependent. But assume (this may be considered a strong assumption) they are comparable, of the same order $\sigma_{\text{MC}} \approx \sigma_{\text{unif}} \approx \sigma_{\text{LHS}} \approx \sigma_{\text{Gauss}}$. Under such hypothesis, we can compare the number of point needed to obtain a given accuracy ϵ with a problem of dimension Q .

Table 5.2.5 presents the number of points required to reach some given accuracies $\epsilon = 10^{-1}$, $\epsilon = 10^{-3}$ and $\epsilon = 10^{-5}$ with respect to $Q \in \{1, 3, 5, 10, 20\}$. In blue we display the smallest number of points needed to ensure ϵ for a given Q . In red, we display the largest one. The colors for the names of the experimental designs correspond to the one used in figure 5.2. In table 5.2.5, for low dimensions, the most efficient

¹²In [289], the author presents an interesting introductory example in the case of the computation of a failure probability with an MC experimental design and the risk of making bad choices.

$\epsilon = 10^{-1}$	$Q = 1$	$Q = 3$	$Q = 5$	$Q = 10$	$Q = 20$
MC	$N = 100$	$N = 100$	$N = 100$	$N = 100$	$N = 100$
Uniform	$N = 10$	$N = 1000$	$N = 10^5$	$N = 10^{10}$	$N = 10^{20}$
(best) LHS	$N = 10$	$N = 339$	$N = 2.3 \times 10^5$	$N = 6.1 \times 10^{14}$	$N = 7.4 \times 10^{37}$
(worst) LHS	$N = 35$	$N = 6909$	$N = 1.2 \times 10^7$	$N = 7.9 \times 10^{16}$	$N = 2.3 \times 10^{40}$
(Lipschitz) LHS	$N = 5$	$N = 9$	$N = 21$	$N = 2064$	$N = 10^9$
Gauss	$N = 2$	$N = 8$	$N = 32$	$N = 1024$	$N = 1048576$
$\epsilon = 10^{-3}$	$Q = 1$	$Q = 3$	$Q = 5$	$Q = 10$	$Q = 20$
MC	$N = 10^6$	$N = 10^6$	$N = 10^6$	$N = 10^6$	$N = 10^6$
Uniform	$N = 1000$	$N = 10^9$	$N = 10^{15}$	$N = 10^{30}$	$N = 10^{60}$
(best) LHS	$N = 9118$	$N = 1.4 \times 10^5$	$N = 1.2 \times 10^8$	$N = 2.7 \times 10^{17}$	$N = 2.5 \times 10^{40}$
(worst) LHS	$N = 1000$	$N = 3.4 \times 10^6$	$N = 5.7 \times 10^9$	$N = 3.4 \times 10^{19}$	$N = 7.7 \times 10^{42}$
(Lipschitz) LHS	$N = 100$	$N = 322$	$N = 1403$	$N = 176269$	$N = 9.0 \times 10^9$
Gauss	$N = 4$	$N = 64$	$N = 1024$	$N = 1048576$	$N = 1.099 \times 10^{12}$
$\epsilon = 10^{-5}$	$Q = 1$	$Q = 3$	$Q = 5$	$Q = 10$	$Q = 20$
MC	$N = 10^{10}$	$N = 10^{10}$	$N = 10^{10}$	$N = 10^{10}$	$N = 10^{10}$
Uniform	$N = 10^5$	$N = 10^{15}$	$N = 10^{25}$	$N = 10^{50}$	$N = 10^{100}$
(best) LHS	$N = 10^5$	$N = 2.9 \times 10^7$	$N = 1.2 \times 10^8$	$N = 9.1 \times 10^{19}$	$N = 7.9 \times 10^{42}$
(worst) LHS	$N = 1.4 \times 10^6$	$N = 8.7 \times 10^8$	$N = 1.8 \times 10^{12}$	$N = 1.2 \times 10^{22}$	$N = 2.4 \times 10^{45}$
(Lipschitz) LHS	$N = 2155$	$N = 9426$	$N = 51780$	$N = 8810510$	$N = 4.0 \times 10^{12}$
Gauss	$N = 5$	$N = 125$	$N = 3125$	$N = 9765625$	$N = 9.5 \times 10^{12}$

Table 5.1: Number of points of the different experimental designs for an expected accuracy of $\epsilon = 10^{-1}, \epsilon = 10^{-3}, \epsilon = 10^{-5}$.

experimental design is the Gauss quadrature whereas the MC one is the worst. As dimension increases, the MC experimental design becomes more and more competitive. On another hand, the uniform one becomes the worst solution to integrate in high dimensions. Table 5.2.5 provides the numbers of points required for few samples of Q and ϵ . Figure 5.2 presents the most efficient experimental design with respect to Q and ϵ . The number of points is not represented. Each color corresponds to a particular experimental design. The Gauss points are the most efficient for small dimensions. Their efficiency (quite erratically) increases with the demanded accuracy. The seesaw efficiency of the Gauss points is complementary with the efficiency of the LHS for Lipschitz continuous functions. For high dimensions, independently of the accuracy needed, the MC experimental design is the most efficient one. Note that the (best) LHS, the (worst) LHS and the uniform experimental designs never really present any interest amongst the studied experimental designs, in the metric of figure 5.2.

5.3 Integration vs. Regression vs. Collocation vs. Kriging

In the literature, many authors do not rely on *integration* to estimate the gPC coefficients for a given polynomial order. In this section, we present regression methods, regression-gPC [35, 26, 270, 271, 15], collocation methods (Lagrange interpolation), collocation-gPC [175, 212, 302, 111, 176, 177, 115] and even kriging-gPC (simpler form of what is described in [262, 158, 258]) and compare them to integration-gPC¹³. We aim at showing that depending on the choice of the approximation basis and of the experimental design, the obtained discretisations may be equivalent or not. In the cases they are not, we aim at highlighting the more progressively possible their differences.

5.3.1 Regression-gPC approximations

Regression has been historically at the basis of many works in statistics for modeling [107, 15, 90, 117]. It is widely used and presents the advantage of being applicable in presence of noisy outputs, i.e.

¹³In the following sections, gPC and integration-gPC denote the same numerical process.

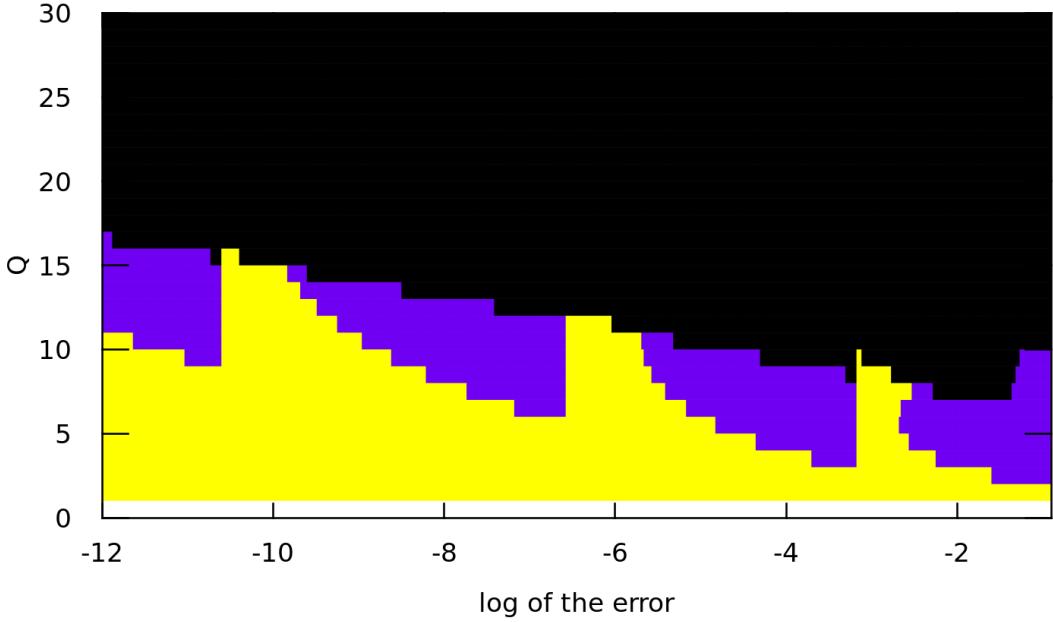


Figure 5.2: The figure presents the most efficient experimental design in dimension Q (y -axis) to reach a given accuracy (the x -axis displays $\log(\epsilon)$) assuming $\sigma_{MC} \approx \sigma_{unif} \approx \sigma_{LHS} \approx \sigma_{Gauss}$. Each color is for a type of experimental design: **Gauss quadrature** is the most efficient for low dimensions Q . The **(Lipschitz) LHS** experimental design has a very narrow area of efficiency. The MC method is by far the most efficient one above a certain dimension Q , independently of the desired accuracy. Note that the **uniform, (best) LHS** and the **(worst) LHS** are never represented.

experimental noise as well as numerical noise:

- Experimental noise refers to variability in the output at the different points of the experimental design *due to a finite accuracy of the measure instruments*: two identical experiments may give slightly different results. The differences could be made smaller with more accurate measurement devices or experimental settings less sensitive to external perturbations.
- Numerical noise refers to variability *due to the use of a stochastic resolution scheme* such as the MC ones described in part III. In this context, running twice the same simulation in identical configurations¹⁴ leads to fluctuations of the observables. Those fluctuations can be made smaller with finer discretisation parameters (typically by increasing N_{MC} , see part III for example).

In this section, we briefly detail the principles of regression together with its properties. For this, let us consider a sequence of linearly independent functions $F_P(x) = (f_0(x), \dots, f_P(x))^t$. Classically in regression approximation, the sequence $F_P(x) = (1, x, \dots, x^P)^t$ is often chosen, see [107, 15]. Let $X \rightarrow u(X)$ be our random variable of interest. The regression model $u_P^F(x)$, approximation of u , is defined as the vector product

$$u_P^F(X) = U_P^t F_P(X).$$

In the expression above, $U_P = (u_0, \dots, u_P)^t$ is defined by the vector of \mathbb{R}^P minimizing the least square error between u and u_P^F , i.e. such that

$$U_P = \underset{V \in \mathbb{R}^P}{\operatorname{Argmin}} [J(V_P)] = \underset{V \in \mathbb{R}^P}{\operatorname{Argmin}} \|u(X) - V_P^t F_P(X)\|_{L^2}^2. \quad (5.21)$$

Differentiating $J(V_P) = \|u(X) - V_P^t F_P(X)\|_{L^2}^2$ with respect to $V_P = (v_0, \dots, v_P)^t$ leads to

$$\nabla_{V_P} J(V_P) = 2\|(u(X) - V_P^t F_P(X))F_P(X)\|_{L^2}.$$

¹⁴but with different initial seeds for the random number generators.

Vector U_P is consequently the unique (due to the convexity of J) solution of

$$\nabla_{V_P} J(V_P) = 0 \iff \int u(X) F_P(X) d\mathcal{P}_X = \int V_P^t F_P(X) F_P(X) d\mathcal{P}_X. \quad (5.22)$$

In the particular case $F_P(x) = \Phi_P^X(x) = (\phi_0^X(x), \dots, \phi_P^X(x))^t$ and assuming *infinite integration accuracy*, the minimum of J is attained at the vector of gPC coefficients $(u_1^X, \dots, u_P^X)^t$.

Let us come back to the previous assumption: equations (5.21) and (5.22) were stated assuming *perfect* integration accuracy. The above expressions (5.21)–(5.22) are in practice discretised *via* the introduction of an experimental design $(X_i, w_i)_{i \in \{1, \dots, N\}}$. It results in the gathering of the set $(u(X_i), w_i)_{i \in \{0, \dots, P\}}$. Equation (5.21) is consequently replaced in practice by

$$U_P^N = \underset{V_P \in \mathbb{R}^P}{\operatorname{Argmin}} [J_N(V_P)] = \underset{V_P \in \mathbb{R}^P}{\operatorname{Argmin}} \sum_{i=1}^N w_i (u(X_i) - V_P^t F_P(X_i))^2. \quad (5.23)$$

In practice, we consequently look for the minimum of $J_N(V_P) \approx J(V_P)$. Functional J_N is minimum for $V_P \in \mathbb{R}^P$ satisfying

$$[(F_P^N)^t W_N F_P^N] V_P = [W_N F_P^N] \begin{pmatrix} u(X_1) \\ \dots \\ u(X_N) \end{pmatrix}. \quad (5.24)$$

In the above expression, we have $W_N = \operatorname{diag}(w_1, \dots, w_N)$ and

$$F_P^N = \begin{pmatrix} f_0(X_1) & \dots & f_0(X_N) \\ \dots & f_k(X_j) & \dots \\ f_P(X_1) & \dots & f_P(X_N) \end{pmatrix}.$$

To give an idea, the expressions of the two vector-matrix products in (5.24) read

$$[W_N F_P^N] = \begin{pmatrix} w_1 f_0(X_1) & \dots & w_N f_0(X_N) \\ \dots & w_j f_k(X_j) & \dots \\ w_1 f_P(X_1) & \dots & w_N f_P(X_N) \end{pmatrix}, \quad (5.25)$$

and

$$[(F_P^N)^t W_N F_P^N] = \begin{pmatrix} \sum_{i=1}^N w_i f_0^2(X_i) & \dots & \sum_{i=1}^N w_i f_0(X_i) f_P(X_i) \\ \dots & \sum_{i=1}^N w_i f_k(X_i) f_l(X_i) & \dots \\ \sum_{i=1}^N w_i f_0(X_i) f_P(X_i) & \dots & \sum_{i=1}^N w_i f_P^2(X_i) \end{pmatrix}. \quad (5.26)$$

Of course, independently of the choices of $(N, P) \in \mathbb{N}^2$, the matrix (5.26) is invertible¹⁵. The solution U_P^N satisfies the well-known (unbiased estimator see [107])

$$U_P^N = [(F_P^N)^t W_N F_P^N]^{-1} [W_N F_P^N] \begin{pmatrix} u(X_1) \\ \dots \\ u(X_N) \end{pmatrix}. \quad (5.27)$$

The conditioning of matrix (5.26) depends on both the choice of the basis $F_P(x)$ and of the experimental

¹⁵This will be emphasized in the few next lines

design $(X_i, w_i)_{i \in \{0, \dots, N\}}$. Suppose $F_P(x) = (1, x, \dots, x^P)^t$, then matrix (5.26) has expression

$$\left(\begin{array}{ccc} \sum_{i=1}^N w_i & \dots & \sum_{i=1}^N w_i X_i^P \\ \dots & \sum_{i=1}^N w_i X_i^{k+l} & \dots \\ \sum_{i=1}^N w_i X_i^P & \dots & \sum_{i=1}^N w_i X_i^{2P} \end{array} \right) \xrightarrow{N \rightarrow \infty} \left(\begin{array}{ccc} s_0^X & \dots & s_P^X \\ \dots & s_{k+l}^X & \dots \\ s_P^X & \dots & s_{2P}^X \end{array} \right). \quad (5.28)$$

It tends to the Hankel matrix of measure $d\mathcal{P}_X$, see (3.10), defined in section 3.4. Two situations may then occur:

- either measure $d\mathcal{P}_X$ is discrete and the Hankel matrices are invertible only up to a specific order (depending on the number of discrete states of $d\mathcal{P}_X$). In this case, the maximum size of (5.26) corresponds to the size of the last Hankel matrix with non-zero determinant.
- Or the Hankel matrices are invertible $\forall P \in \mathbb{N}$ but their determinants are known to tend to zero quickly as P grows (see section 3.4). It is consequently harder and harder to numerically inverse with P .

Nonetheless, in both cases in practice, (5.26) is always invertible. Figure 5.3 presents the results obtained

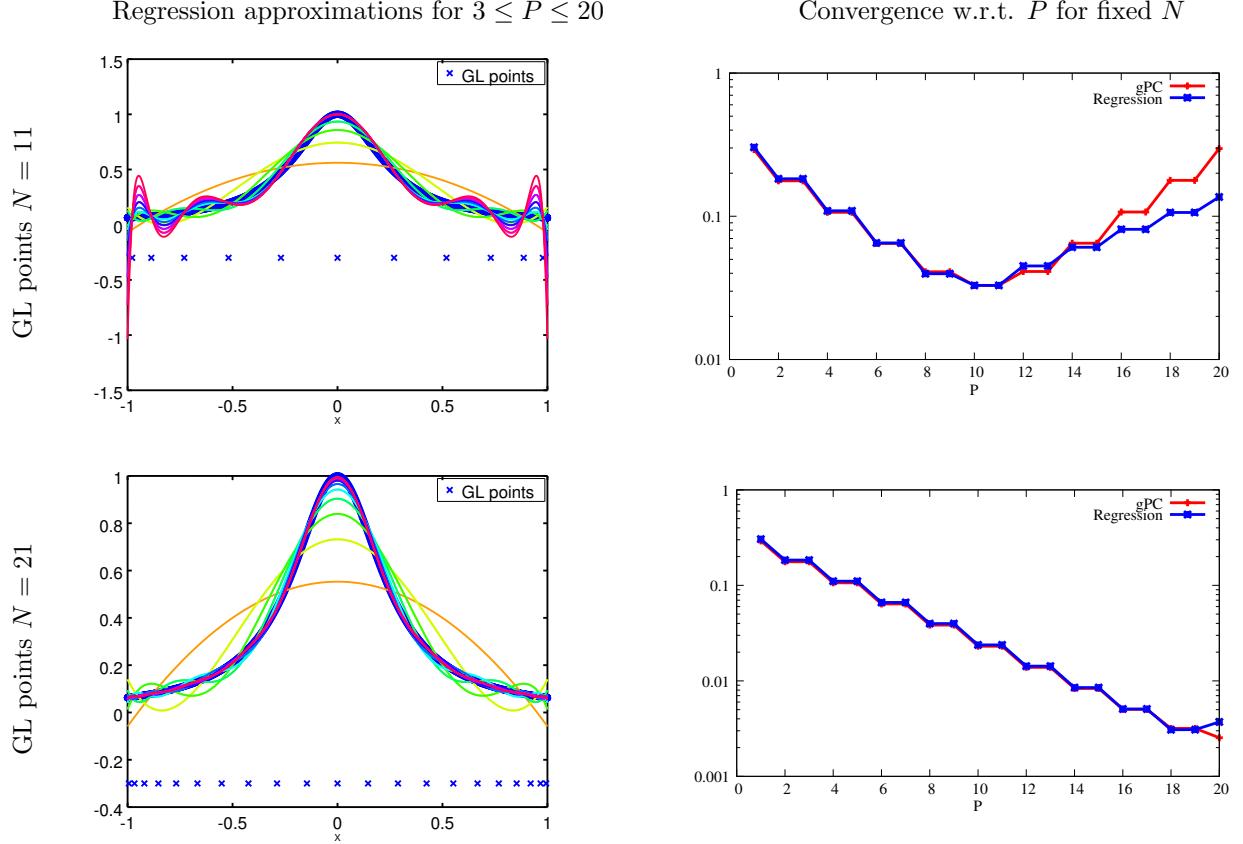


Figure 5.3: Application of Gauss-Legendre quadrature rule for *regression* for the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N .

with regression on the Runge function in exactly the same conditions as in section 5.2.3 together with the

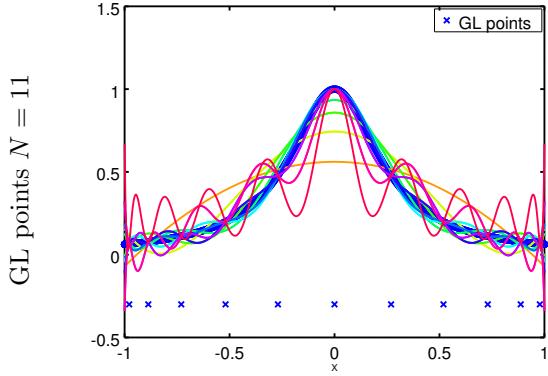
results obtained by integration-gPC. If we focus on the convergence studies (right column) with respect to P for fixed numbers of GL points $N = 11$ and $N = 21$, integration-gPC and regression have

- first, the same behaviour. For a small number of points of the experimental design, the polynomial order must be kept low. The error decreases then increases after $N = 11$ (as soon as P goes beyond $N = 11$).
- But for polynomial orders higher than $P = N = 11$, the error is a little bit more controlled with the regression approximation than with the integration one.

The shapes of the regression approximations for the different order for $N = 11$ (top left of figure 5.3) are quite different than the ones obtained with integration (top left of figure 5.1). When P is kept $P \leq N$, the results obtained with integration or regression are equivalent (up to the accuracy/cost of a matrix inversion). See for example the bottom pictures of figure 5.3. This is due to the fact that, by definition, the L^2 -minimization is invariant with a change of basis.

Now suppose a particular form for $F_P(x) = \Phi_P^X(x) = (\phi_0^X(x), \dots, \phi_P^X(x))^t$ with $(\phi_k^X)_{k \in \{0, \dots, P\}}$ the components of a chosen gPC basis. The obtained approximations are denoted *regression-gPC* ones in this document. Let us introduce $U_P^{\text{int}, N} = (u_0^{X, N}, \dots, u_P^{X, N})^t$ the vector of coefficients of the gPC approxima-

Regression-gPC approximations for $3 \leq P \leq 20$



Convergence w.r.t. P for fixed N

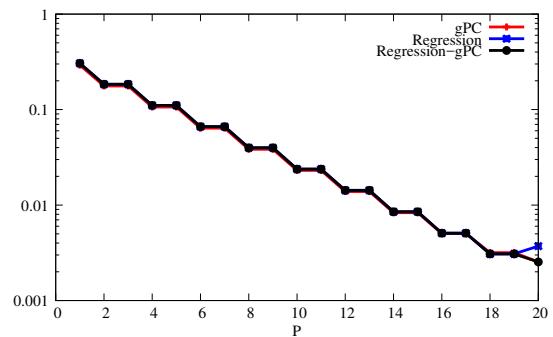
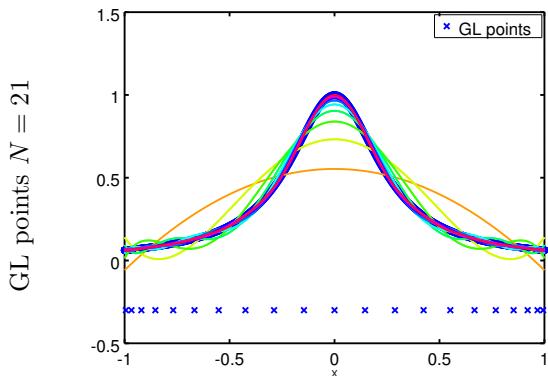
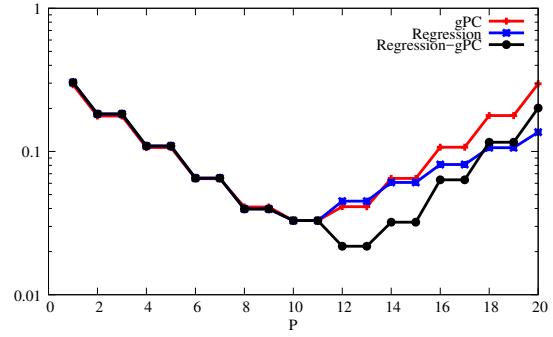


Figure 5.4: Application of Gauss-Legendre quadrature rule for *regression-gPC* of the gPC coefficients of the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N .

tion obtained by integration. From (5.24), it is easy noticing that, independently of the choice of the

experimental design, the regression solution U_P^N is related to the integration coefficients $U_P^{\text{int},N}$ by

$$\left(\begin{array}{ccc|c} \sum_{i=1}^N w_i (\phi_0^X(X_i))^2 & \dots & & \sum_{i=1}^N w_i \phi_0^X(X_i) \phi_P^X(X_i) \\ \dots & \sum_{i=1}^N w_i \phi_k^X(X_i) \phi_l^X(X_i) & \dots & \\ \sum_{i=1}^N w_i \phi_0^X(X_i) \phi_P^X(X_i) & \dots & & \sum_{i=1}^N w_i (\phi_P^X(X_i))^2 \end{array} \right) U_P^N = U_P^{\text{int},N}. \quad (5.29)$$

Suppose now $(X_i, w_i)_{i \in \{1, \dots, N\}}$ is a Gauss quadrature rule having the properties described in 5.2. If we furthermore assume $P \leq N$, it ensures the *exact* orthonormality of the gPC basis even with $N < \infty$ (i.e. even in a finite integration accuracy context). Consequently, with such choice, we have

$$U_P^N = U_P^{\text{int},N}.$$

Otherwise (i.e. if $N > P$), the regression-gPC coefficients and the integration ones differ from the fact the integration coefficients may not minimize the least squared error. This is emphasized in figure 5.4. For $P > N$, the regression-gPC approximations have a better control of the L^2 error than both classical regression and integration.

The main interest of regression-gPC is to be able to deal with any experimental design, independently of its integration accuracy, still ensuring a relatively good conditioning of matrix (5.26). It is particularly

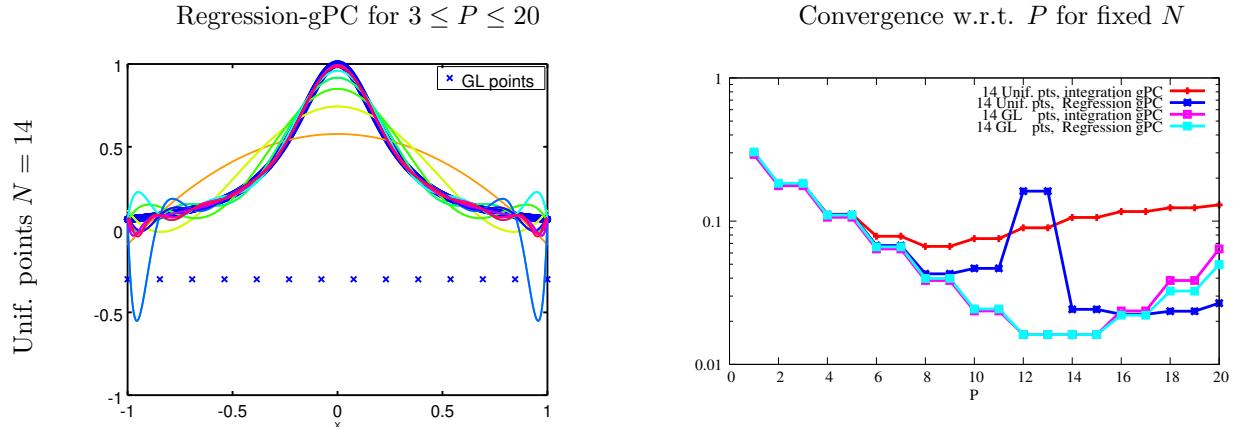


Figure 5.5: Application of Gauss-Legendre quadrature rule for *regression-gPC* of the gPC coefficients of the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N .

convenient for experimental settings (in opposition to numerical experiments) and has been originally designed [107, 15, 90] for its ability to take into account experimental noise.

The above property is emphasized in figure 5.5 in which we briefly investigate the sensitivity to the choice of the experimental design for integration and regression-gPC. Figure 5.5 (left) shows the results obtained with regression-gPC with $N = 14$ equispaced points (uniform experimental design). Figure 5.5 (right) allows comparing the results obtained with integration-gPC and regression-gPC for $N = 14$ GL points and $N = 14$ equispaced ones. For the GL points, the behaviour is similar to what was presented in figure 5.4 (except we have $N = 14$ instead of $N = 11$). With the equispaced experimental design, the integration accuracy is lower than with the GL points for fixed N (see section 5.2.5). Integration-gPC consequently gives less satisfactory results with such design: the two approaches have equivalent L^2 -performances only up to $P = 5$. For higher polynomial orders, the integration error becomes prepon-

derant with respect to the truncation one. With regression-gPC on another hand, the accuracy of the GL and equispaced experimental designs are comparable up to order $P = 9$.

5.3.2 Collocation-gPC approximation

Regressions are convenient especially when one has to deal with experimental/numerical noise. When considering *numerical* experiments, i.e. simulations, some resolution schemes may be *reproducible* in the sense two runs of the same configuration give exactly the same results. This is the case for example for the finite volume schemes used in the 'fil rouge' application. Dealing with reproducible simulation codes, one may demand the stochastic approximation method to be able to strictly recover the numerical results at the experimental design points, see [175, 212, 302, 111, 176, 177, 115]. This can be obtained using interpolation methods such as Lagrange interpolation or high-order splines for example.

Lagrange interpolation can be obtained applying formulae (5.27) in the particular case $N = P$. In practice, in order to avoid the inversion of a possibly badly conditioned matrix, the Lagrange formulae is applied

$$L_i(x) = \prod_{\substack{j=1 \\ i \neq j}}^N \frac{x - X_j}{X_i - X_j}, \forall i \in \{1, \dots, N\}. \quad (5.30)$$

The resulting collocation approximation is then given by

$$u_N^L(X) = \sum_{j=1}^N u(X_j) L_i(X). \quad (5.31)$$

In term of asymptotical error analysis, we have the following well-known property: suppose $u \in C^0([a, b])$, and an experimental design $(X_i)_{i \in \{1, \dots, N\}}$ with N distinct nodes, then there exists $\xi \in [a, b]$ such that

$$u(X) - u_N^L(X) = \frac{u^{(N)}(\xi)}{N + 1!} \prod_{i=1}^N (X - X_i). \quad (5.32)$$

Taking the L^∞ -norm in the above expression we obtain

$$\|u(X) - u_N^L(X)\|_{L^\infty} \leq \frac{1}{N + 1!} \max_{\xi \in [a, b]} |u^{(N)}(\xi)| \max_{x \in [a, b]} \left| \prod_{i=1}^N (x - X_i) \right|. \quad (5.33)$$

Note that the collocation approximation does not necessarily converge. Its converging behaviour is strongly correlated to the choice of the experimental design. Figure 5.6 (top) presents the collocation approximations of Runge function¹⁶ obtained for N going from 1 to 20 of a uniform experimental design. The collocation approximations diverge¹⁷ as N increases (figure 5.6 top-left).

The question arising now is: is it possible to choose an experimental design ensuring the convergence of the collocation approach. It is commonly known, see [225], that the term $\max_{x \in [a, b]} \left| \prod_{i=1}^N (x - X_i) \right|$ in (5.33) is $\mathcal{O}(\frac{1}{2^N})$ at the roots of Chebyshev's polynomials (CC points). This ensures uniform convergence at those points. The roots of Chebyshev polynomials are Gauss points. Some authors generalized the property of Chebyshev roots for collocation to arbitrary Gauss ones: the method is commonly called *stochastic collocation, or collocation-gPC* in the literature, see [175, 212, 302, 111, 176, 177, 115]. It refers to the use of Lagrange polynomials at the Gauss quadrature points associated to the probability measure $d\mathcal{P}_X$ of the input random variable X .

If derivatives of u are available at the experimental design points, *Hermite interpolation* can also be

¹⁶This example is well-known, I claim no originality here.

¹⁷It is even possible to prove it diverges in this case, see [64, 102] for example.

applied, generalizing Lagrange interpolation. It ensures

$$u(X) - u_N^H(X) = \frac{u^{(N)}(\xi)}{N + 1!} \prod_{i=1}^N (X - X_i)^{k_i},$$

where k_i corresponds to the number of orders available at point X_i . For example if one has access to $(u(X_i), u'(X_i)) \forall i \in \{1, \dots, N\}$, then $k_i = 2, \forall i \in \{1, \dots, N\}$.

Figure 5.6 presents the collocation-gPC approximations of Runge function (i.e. Lagrange polynomials at GL points) in the same conditions as in the previous sections. As testifies the convergence study of

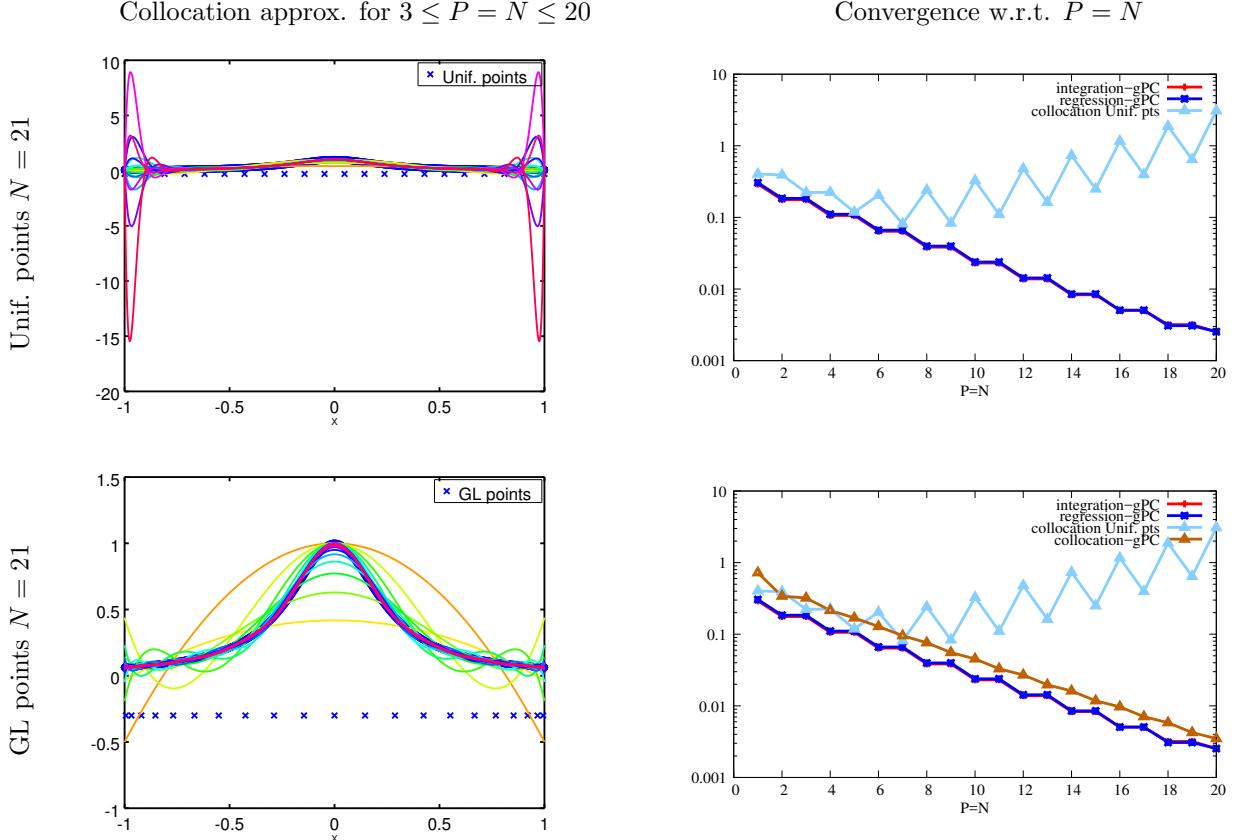


Figure 5.6: Application of Gauss-Legendre quadrature rule for *collocation-gPC* of the gPC coefficients of the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N .

figure 5.6 (right), collocation-gPC exhibits an exponential convergence behaviour. The global accuracy at each order/number of points $N = P$ remains higher than for integration and regression. This is mainly due to systematic oscillating behaviour between the points. Of course, for $P = 20$ in figure 5.6 (right), regression-gPC degenerates toward collocation-gPC.

Collocation-gPC has only one parameter as $N = P$: in figure 5.6, we compared collocation-gPC to integration-gPC and regression-gPC with $N = 21$. For the latters, the comparison may seem unfair, especially for high polynomial orders P . In figure 5.7, we perform some convergence studies keeping $P = 5$ fixed for integration-gPC and regression-gPC and compare them to the collocation-gPC approximations with increasing N . The convergence studies have two regimes for the integration-gPC and regression-gPC approximations, only one for collocation-gPC:

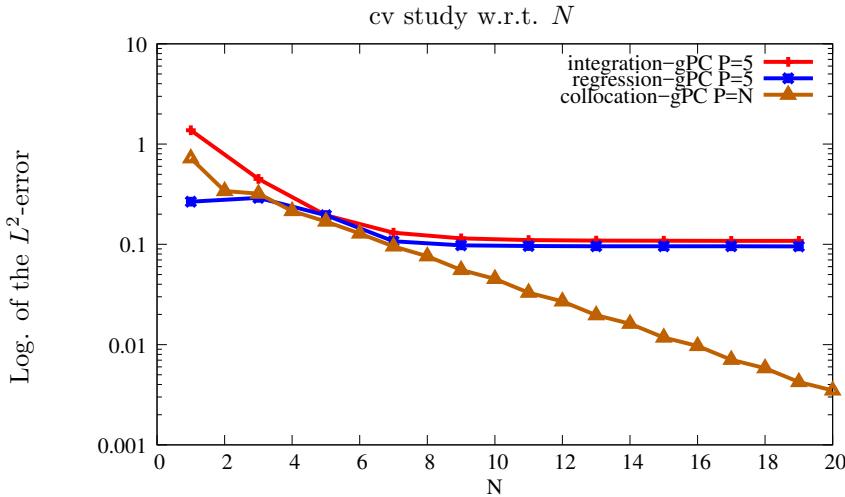


Figure 5.7: Convergence studies for integration-gPC, regression-gPC (both for fixed $P = 5$) and collocation-gPC with respect to N for approximating Runge function (5.17).

- in the first regime, defined by $N \leq 7$, the three methods exhibit an exponential convergence rate (log-scale). We recall we chose $P = 5$ for this example, implying a small numbers of points ($N \leq P$) leads to inaccurate approximations.
- Beyond $N = 7$ points, integration-gPC and regression-gPC’s accuracies stagnate: the stagnation plateau corresponds to the fact the truncation error (second term in (5.5)) becomes preponderant with respect to the integration one as N increases. This is emphasized by the fact this plateau is the same for the two approximation methods (as they share the same experimental design).

For collocation-gPC, the error keeps on decreasing exponentially with $N = P$. This is due to the fact that in this case the asymptotic error only depends on the spacing between the points of the experimental design, see (5.33).

From the two previous sections, one may wonder whether it is possible to take advantage of both methods. Regression-gPC allows taking into account noisy outputs and exhibits a fast convergence with respect to P . Collocation-gPC ensures recovering exactly the outputs at the experimental design points and exhibits a convergence rate mainly depending on the spacing between the points once the integration accuracy reached. A compromise between regression-gPC and collocation-gPC would be a penalized regression with the introduction of Lagrange multipliers to ensure, as constraints, the approximation is interpolatory with u at the design points. This idea is at the basis of *kriging*, briefly presented in the next section.

5.3.3 Kriging-gPC approximations

The reader interested in kriging may find different denominations such as *simple kriging*, *Ordinary kriging*, *Universal kriging* etc. Our starting point is kriging-gPC, which is here a less elaborated version¹⁸ of what can be found in [262, 158, 258]. We focus on this variant of kriging as it is the most general we know and, of course, is related to a gPC basis, central in this part II. Basically, kriging-gPC consists in choosing $F_P(X) = \Phi_P(X)$ as an approximation basis in a (universal¹⁹) kriging approximation: it confers to the resolution similar advantages as the ones emphasized in the section comparing regression and regression-gPC (good conditioning etc.). We insist the section is non-exhaustive regarding kriging techniques and we rely on very pedagogical publications dedicated to them [284, 21, 158, 258] for a complete state-of-the-art. This section mainly aims at comparing and understanding the differences

¹⁸less elaborated in the sense in [262, 158, 258], the authors provide an algorithm in order to choose automatically the gPC order P whereas in this document we choose it *a priori*.

¹⁹If I am not mistaking.

between kriging-gPC and gPC approximations and provide an original numerical analysis of the methods.

Kriging is also known as Gaussian process modeling. It assumes the output random variable of interest $u(X)$ is a realisation of a Gaussian random process. Let us sketch the idea behind the methodology in the next lines. Recall we aim at approximating the transformation of the known random variable X into the unknown one $u(X)$. The gPC approximation relied on a polynomial approximation, i.e. $u(X) \approx \sum_{k=0}^P u_k^X \phi_k^X(X)$ where $(\phi_k^X)_{k \in \{0, \dots, P\}}$ is the gPC basis and $(u_k^X)_{k \in \{0, \dots, P\}}$ the gPC coefficients. Kriging-gPC relies on the introduction of Z_P^X such that

$$u(X) = \sum_{k=0}^P u_k^X \phi_k^X(X) + Z_P^X(X).$$

In other words, it corresponds to a gPC development plus its residue Z_P^X in the P -truncated gPC associated to X . In a kriging context, $\sum_{k=0}^P u_k \phi_k^X(X)$ is commonly called the *trend*. Kriging introduces an *additional mathematical ingredient* whose aim is to approximate the random variable Z_P^X , the residue of the gPC development. The idea is to introduce an additional dimension $u \in \text{Supp}(X)$ and assume $Z_P^X(X) \sim \sigma^2(u)Z(u)$ is a zero-mean gaussian process of variance σ^2 independent of X . The gaussian process is fully characterised by its covariance kernel K defined by $K(u, v) = \mathbb{E}[Z(u)Z(v)]$ and such that $\sigma^2(u) = \mathbb{E}[Z^2(u)]$. Suppose K is known, then several constraints must be satisfied for Z to be a relevant gaussian process to approximate $u(X)$. Introduce $\mu_Z(u) = \mathbb{E}[Z(u)]$, then we must for example have

$$\left\{ \begin{array}{lcl} \mathbb{E}[u(X)] & = & u_0^X + \mathbb{E}[\mu_Z(X)] = u_0^X + 0 = u_0^X, \\ \mathbb{E}[u^2(X)] & = & \sum_{k=0}^P (u_k^X)^2 + \sum_{k=0}^P u_k^X \mathbb{E}[\phi_k^X(X)\mu_Z(X)] + \mathbb{E}[\mu_Z^2(X)], \\ \dots, \end{array} \right. \quad (5.34)$$

and so on. Ensuring the constraints are satisfied is directly linked to how K is chosen or built. In practice, kriging models:

- first generally assume a particular parameterized shape of the covariance function $K(u, v, \theta)$ where θ is an additional (set of) parameter(s).
- The second step consists in calibrating θ . This means looking for $\hat{\theta}$ minimizing differences with the above constraints in a norm which remains to be defined at this stage of the discussion. This can be done by various means (Maximum Likelihood, Cross Validation estimation, etc. see [262, 158, 258, 21, 19]).
- Once $\hat{\theta}$ obtained, we have access to the random variable $\mu(X, \hat{\theta})$ and its *predictive variance* $\sigma^2(X, \hat{\theta})$ with explicit matrix vector formulas (briefly detailed in the following lines).

In practice, K is often chosen homogeneous, i.e. such that $K(u, v, \theta) = K(u - v, \theta)$. For a given choice of K and θ we have (see [158])

$$\begin{aligned} \mu(X, \theta) &= U_P^N(\theta)^t \Phi_P(X) + k(X, \theta)^t W_N K^{-1}(\theta) \begin{pmatrix} u(X_1) - U_P^N(\theta)^t \Phi_P(X_1) \\ \dots \\ u(X_N) - U_P^N(\theta)^t \Phi_P(X_N) \end{pmatrix}, \\ \sigma^2(X, \theta) &= \sigma_K^2(\theta) \left(1 - [\Phi_P^t(X), k^t(X, \theta)] \begin{bmatrix} 0 & (W_N \Phi_P^N)^t \\ (W_N \Phi_P^N)^t & W_N K(\theta) \end{bmatrix} \begin{bmatrix} \Phi_P(X) \\ k(X, \theta) \end{bmatrix} \right). \end{aligned}$$

In the above expressions, $\mu(X, \theta), \sigma^2(X, \theta)$ are the mean and variance of the gaussian process approximating $u(X)$. The notations are almost the same as the ones of section 5.3.1: $(X_i)_{i \in \{1, \dots, N\}}$ are the points of the experimental design, $W_N = (w_1, \dots, w_N)^t$ the vector of their weights and

$$U_P^N(\theta) = \begin{pmatrix} u_0^X(\theta) \\ \dots \\ u_P^X(\theta) \end{pmatrix}, \Phi_P(X) = \begin{pmatrix} \phi_0^X(X) \\ \dots \\ \phi_P^X(X) \end{pmatrix}, \Phi_P^N = \begin{pmatrix} \phi_0^X(X_1) & \dots & \phi_0^X(X_N) \\ \dots & \Phi_k^X(X_j) & \dots \\ \phi_P(X_1) & \dots & \phi_P^X(X_N) \end{pmatrix}.$$

It additionally introduces $K(\theta)$, the matrix of general term $K_{i,j}(\theta) = K(X_j - X_i, \theta)$ and $k(X, \theta) = (k(X - X_1, \theta), \dots, k(X - X_N, \theta))^t$. Besides, the estimations of the coefficients of the development together with the variance parameter are given by

$$\begin{aligned} U_P^N(\theta) &= [(\Phi_P^N)^t W_N K^{-1}(\theta) \Phi_P^N]^{-1} W_N \Phi_P^N K^{-1}(\theta) \begin{pmatrix} u(X_1) \\ \dots \\ u(X_N) \end{pmatrix}, \\ \sigma_K^2(\theta) &= \left(\begin{pmatrix} u(X_1) \\ \dots \\ u(X_N) \end{pmatrix} - \Phi_P^N U_P^N(\theta) \right)^t W_N K^{-1}(\theta) \left(\begin{pmatrix} u(X_1) \\ \dots \\ u(X_N) \end{pmatrix} - \Phi_P^N U_P^N(\theta) \right). \end{aligned} \quad (5.35)$$

In the above expressions, θ remains to be chosen. In fact, equations (5.35) express the results of the minimization of the L^2 -norm (least square error, as for regression) between $u(X)$ and $\mu(X, \theta)$ for a fixed θ . As hinted at in [158], if $K = I_N$ ²⁰ where I_N is the identity of size N , then (5.35) degenerates toward (5.27) for the regression approximation.

The discussion about the relevant shape of the covariance function K or the way the parameter θ is tuned is beyond the scope of this document. We refer to [21, 19, 20] for the reader interested in deepening those considerations. The covariance kernel K can be evaluated but in general, it is chosen *a priori*. The most classical choices are gaussian, exponential or Matérn kernels²¹. Kriging provides σ^2 as a measure of precision. However this measure relies on the *correctness* of the covariance function, see [19, 21]. In other words, the term *predictive* variance may be strong and it, in general, reflects an assumption. If it does not hold, the error estimation might be bad and no error estimation properties are guaranteed. However, typically, still a good interpolation is achieved for the random variable $\mu(X, \theta)$, mean of the Gaussian process. We focus on it in the next numerical analysis and tests.

We first go through the numerical analysis of (the mean of) the kriging-gPC approximation. We aim at helping interpreting the numerical results and comparisons with the previously presented approximations (regression-gPC and collocation-gPC mainly) displayed in figure 5.8. For this, let us rewrite the mean of the kriging-gPC process under a more friendly form. By noticing that

$$\begin{aligned} \mu(X, \theta) &= \sum_{k=0} u_k^X(\theta) \phi_k^X(X) \\ &\quad + (k(X - X_1, \theta), \dots, k(X - X_N, \theta)) K^{-1}(\theta) \begin{pmatrix} u(X_1) - \sum_{k=0}^P u_k^X(\theta) \phi_k^X(X_1) \\ \dots \\ u(X_N) - \sum_{k=0}^P u_k^X(\theta) \phi_k^X(X_N) \end{pmatrix}, \end{aligned} \quad (5.36)$$

the expression can be recast as

$$\mu(X, \theta) = \sum_{k=0} u_k^X(\theta) \phi_k^X(X) + \sum_{i=1}^N a_i^P k(X - X_i, \theta). \quad (5.37)$$

Expression (5.37) may appear downgrading in comparison to (5.36) as many important properties of the approximation do not anymore explicitly appear in the coefficients $(a_i^P)_{i \in \{1, \dots, N\}}$. Still, it is enough for the following material. Let us introduce the functional F such that

$$F(X, \theta) = u(X) - \mu(X, \theta) - g(\bar{X}, \theta) \prod_{i=1}^N (X - X_i).$$

²⁰closely related to the particular choice $k(u, v) = \delta_u(v)$.

²¹which recovers continuously with a parameter both the gaussian and the exponential kernels

We define \bar{X} as an arbitrary point in $Supp(X)$ such that $\forall i, \bar{X} \neq X_i$ and

$$g(\bar{X}, \theta) = \frac{u(\bar{X}) - \mu(\bar{X}, \theta)}{\prod_{i=1}^N (\bar{X} - X_i)}. \quad (5.38)$$

The functional F has consequently $N + 1$ roots X_1, \dots, X_N, \bar{X} . Assume furthermore that $u(X)$ and $k(u, v, \theta)$ are C^{N+1} where, we recall, N is the number of points of the experimental design. Then according to Rolle's theorem, $\exists \xi_0$ such that $F^{(N+1)}(\xi_0, \theta) = 0$. Let us now consider two situations, we distinguish $g = g_{P \leq N}$ in the first case and $g = g_{P > N}$ in the second:

- first, suppose $P \leq N$ so that differentiating $N + 1$ times F resumes to

$$F^{(N+1)}(X, \theta) = u^{(N+1)}(X) - \sum_{i=1}^N a_i^P k^{(N+1)}(X - X_i, \theta) - g_{P \leq N}(\bar{X}, \theta)(N + 1!).$$

Using the fact that $F^{(N+1)}(\xi_0, \theta) = 0$ allows identifying $g_{P \leq N}$ as

$$g_{P \leq N}(\bar{X}, \theta) = \frac{1}{N + 1!} \left(u^{(N+1)}(\xi_0) - \sum_{i=1}^N a_i^P k^{(N+1)}(\xi_0 - X_i, \theta) \right).$$

Using the above expression of $g_{P \leq N}$ with respect to ξ_0 in (5.38) leads to the following error estimator:

$$u(X) - \mu(X, \theta) = \frac{1}{N + 1!} \left(u^{(N+1)}(\xi_0) - \sum_{i=1}^N a_i^P k^{(N+1)}(\xi_0 - X_i, \theta) \right) \prod_{i=1}^N (X - X_i). \quad (5.39)$$

Equation (5.39) can be compared to (5.32) for collocation by noticing that

$$u(X) - \mu(X, \theta) = \underbrace{\frac{u^{(N+1)}(\xi_0)}{N + 1!} \prod_{i=1}^N (X - X_i)}_{(*) \text{ recalls (5.32) for collocation}} - \underbrace{\frac{1}{N + 1!} \sum_{i=1}^N a_i^P k^{(N+1)}(\xi_0 - X_i, \theta) \prod_{i=1}^N (X - X_i)}_{(**)}$$

The first term $(*)$ in (5.40) recovers the collocation error term. The error depends on P only via the coefficients $(a_i^P)_{i \in \{1, \dots, N\}}$ in the second term $(**)$. Equation (5.40) testifies one can decrease the constant multiplying the convergence rate²² of the approximation method if K is sufficiently well suited. Note also that conversely, nothing prevents it from increasing it, with respect to collocation, if it is not.

- Suppose now $P > N$ and differentiate $N + 1$ times F . We obtain

$$F^{(N+1)}(X, \theta) = u^{(N+1)}(X) + \sum_{k=0}^P u_k^X(\theta) (\phi_k^X(X))^{(N+1)} - \sum_{i=1}^N a_i^P k^{(N+1)}(X - X_i) - g_{P > N}(\bar{X})(N + 1!).$$

To simplify the above expression, we can rewrite $\phi_k^X(X) = \Gamma_k^X \phi_k^{X,m}(X) = \Gamma_k^X \prod_{i=1}^k (X - \gamma_i^k)$, $\forall k \in \{0, \dots, P\}$ with $\phi_k^{X,m}$ the monic orthogonal polynomial relative to ϕ_k^X . In the latter expression, $(\gamma_i^k)_{i \in \{1, \dots, k\}}$ are the roots²³ of ϕ_k^X . For $k > N$, the $(N + 1)^{th}$ derivative of ϕ_k^X can be expressed as

$$(\phi_k^X(X))^{(N+1)} = \Gamma_k^X \sum_{i_1 + \dots + i_k = N+1} C_{i_1, \dots, i_k}^{N+1} \prod_{j=1}^k (X - \gamma_j^k)^{(i_j)}.$$

²²The convergence rate is here related to $\frac{1}{N+1!} \prod_{i=1}^N (X - X_i)$ together with a choice of norm but we here study the raw expression (5.40).

²³For a gPC basis, we know those roots are real, distinct in $Supp(X)$ but the material holds for an arbitrary choice of $F_P(X)$, with complex roots.

In the above expression, the multinomial coefficients are given by $C_{i_1, \dots, i_k}^{N+1} = \frac{N+1!}{\prod_{j=1}^k i_j!}$. Each $(X - \gamma_j^k)_{j \in \{1, \dots, k\}}$ are monomials so that $(i_j)_{j \in \{1, \dots, k\}} \in \{0, 1\}^k$. Otherwise, the above expression would be zero as $k < N + 1$. In other words the multinomial coefficients simplify to

$$(\phi_k^X(X))^{(N+1)} = (N+1)!\Gamma_k^X \sum_{i_1+\dots+i_k=N+1} \prod_{j=1}^k (X - \gamma_j^k)^{(i_j)}.$$

Now, once again using the fact that $F^{(N+1)}(\xi_0, \theta) = 0$ ensures

$$\begin{aligned} g_{P>N}(\bar{X}, \theta) &= g_{P \leq N}(\bar{X}, \theta) + \sum_{k=0}^P u_k^X(\theta) \Gamma_k^X \sum_{i_1+\dots+i_k=N+1} \prod_{j=1}^k (\xi_0 - \gamma_j^k)^{(i_j)}, \\ &\stackrel{N \text{ fixed}}{\underset{P \gg N}{\equiv}} g_{P \leq N}(\bar{X}, \theta) + \mathcal{O}(\Gamma_P^X). \end{aligned}$$

Using the above expression in (5.38) leads to the following error estimator for the mean $\mu(X, \theta)$:

$$\begin{aligned} u(X) - \mu(X, \theta) &= \\ &\frac{1}{N+1!} \left(\begin{array}{l} +u^{(N+1)}(\xi_0) - \sum_{i=1}^N a_i^P k^{(N+1)} (\xi_0 - X_i) \\ +(N+1)! \sum_{k=0}^P u_k^X(\theta) \Gamma_k^X \sum_{i_1+\dots+i_k=N+1} \prod_{j=1}^k (\xi_0 - \gamma_j^k)^{(i_j)} \end{array} \right) \prod_{i=1}^N (X - X_i). \end{aligned} \quad (5.41)$$

With (5.41), it is easy verifying for fixed N and $P \gg N$, $u(X) - \mu(X, \theta) \sim \Gamma_P^X$. This imples the approximation may diverge as²⁴. The analysis even shows that in this regime, the asymptotic behaviour is independent of the choice of $K(u, v, \theta)$. In other words, the same analysis allows explaining the behaviour of regression-gPC for $P > N$ in the examples of the previous figures (obtained with the particular choice $K(u, v, \theta) = \delta_u(v)$).

To illustrate the above material, we suggest going through the application of kriging-gPC to the Runge function in the same conditions as in the previous paragraphs. The covariance function is here chosen as an exponential one

$$K(u, v, \theta) = \theta \exp(-|u - v|\theta). \quad (5.42)$$

The parameter θ is calibrated performing a simple dichotomy to minimize the predictive variance $\sigma_K^2(\theta)$ as suggested in [262, 158, 258]. Figure 5.8 presents the kriging-gPC approximations together with the ones obtained by integration-gPC and regression-gPC in the same conditions (the comparisons with collocation-gPC will be tackled later on). Let us first comment on the qualitative results of the first column of figure 5.8: for a low number of quadrature points ($N = 11$), kriging-gPC behaves as the previous approximations and remain very oscillatory. As N increases to 21 (bottom-left picture), the oscillations are way more controlled and the kriging-gPC results are much less sensitive to the choice of the truncation order P than the other approximations. The quantitative convergence results of the right column of figure 5.8 show first that for $P \leq N$, the kriging-gPC approximations outperform the other ones. We recall kriging-gPC benefits an additional discretisation parameter K which here is very efficient in the sense it probably allows interesting compensations between term (*) and term (**) in expression (5.40). For higher polynomial order, i.e. in the top pictures of figure 5.8 with $N = 11$ and $P \geq N$, analysis (5.41) becomes more and more relevant and we recover experimentally that the L^2 -norm of the error $\|u(X) - \mu(X, \hat{\theta})\|_{L^2} = \mathcal{O}(\Gamma_P^X)$ grows fast with P . The figure even allows recovering the fact that in such conditions (N fixed and $P \gg N$) kriging-gPC and regression-gPC give equivalent results as expected by the error analysis of (5.41): in this regime the leading term is independent of the choice of K . If now P is kept lower than $N = 21$ as in the bottom right picture of figure 5.8, error analysis (5.40) applies. The kriging-gPC approximations then give very satisfactory results with a *flat* convergence curve testifying of a less sensitive behaviour with respect to the discretisation parameter P than other methods.

²⁴As for example, for Legendre polynomials $\Gamma_P^X = \sqrt{2P+1}$.

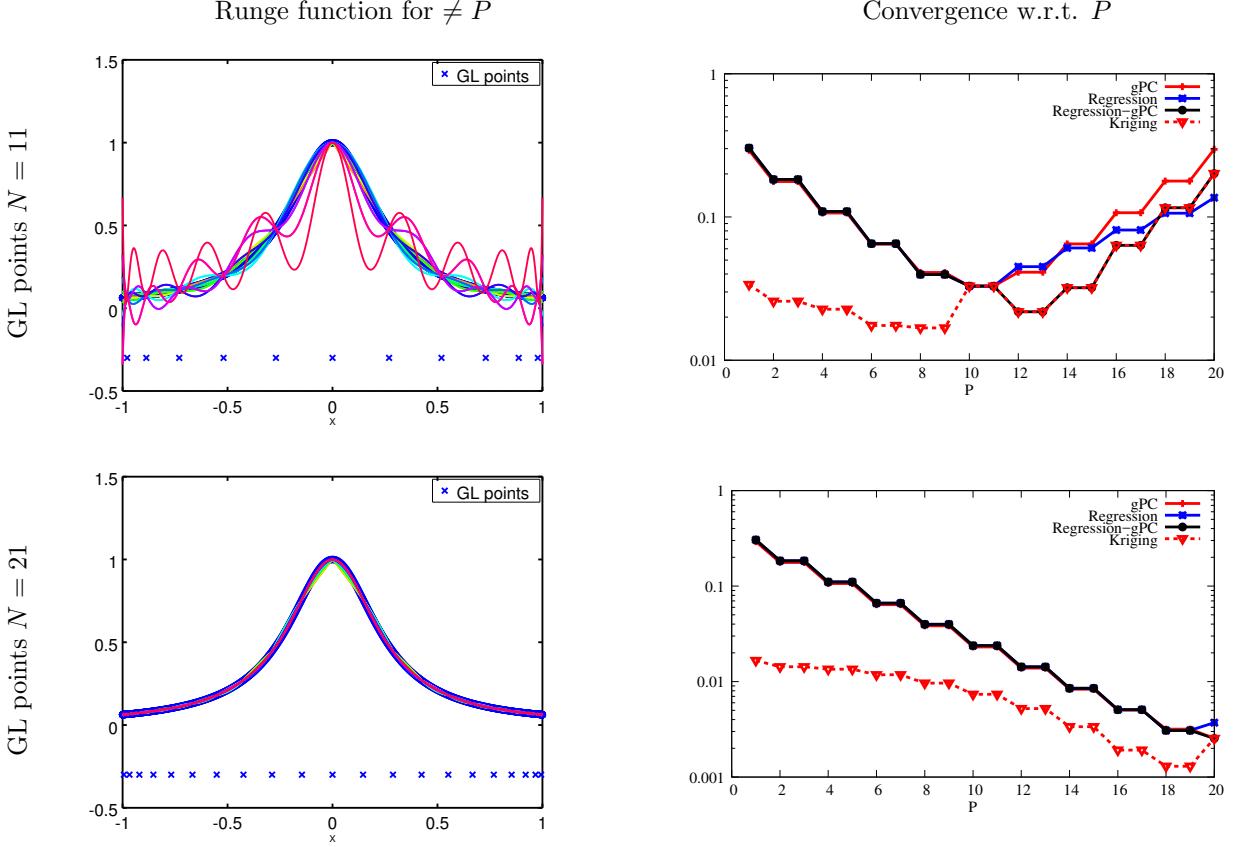


Figure 5.8: Application of Gauss-Legendre quadrature rule for *kriging-gPC* of the gPC coefficients of the transformation of a uniform random variable through the Runge function with $N = 11$ (top) and $N = 21$ (bottom). The left column present the polynomial approximations for $3 \leq P \leq 20$. The right column present the L^2 -norm of the error with respect to P for fixed N . The kriging kernel is chosen exponential (5.42) and a dichotomy is applied to calibrate θ .

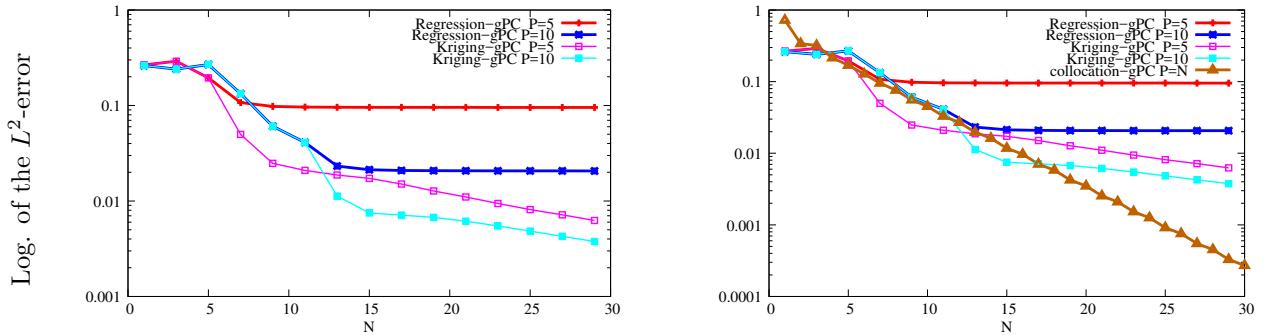


Figure 5.9: Convergence studies for regression-gPC, kriging-gPC (both for fixed $P = 5$ and $P = 10$) and collocation-gPC with respect to N for approximating Runge function (5.17).

To fully understand the influence of the covariance function, let us finally perform a convergence study with respect to N for fixed $P = 5$ and $P = 10$ (same conditions as in figure 5.7). The results are presented in figure 5.9: the left picture compares the convergence studies with respect to N obtained with regression-gPC as in section 5.3.1 and kriging-gPC for fixed $P = 5$ and $P = 10$. Both methods present two regimes:

- in the first one, characterised by $N \leq P$, regression-gPC and kriging-gPC gives *exactly the same* results in L^2 -norm in agreement with the previous analysis (5.41). It testifies of a relative independence of the choice of K of the approximation in such configuration.
- For $N > P$, regression-gPC approximations stagnate as the truncation error remains preponderant with respect to the integration one. On another hand, for kriging-gPC, the accuracy of the approximations continues to increase with N due to the covariance term in (5.40) ensuring a convergence driven by the spacing between the points of the experimental design in this regime.

The right picture of figure 5.9 presents the same curves together with the collocation-gPC one: for $N > P$ for which error analysis (5.40) applies, the kriging-gPC approximations do not systematically give better results than collocation-gPC. This is due to the fact the covariance function

- may be well-suited for some couples (N, P) (for example for $P = 5$ and $N \in \{5, \dots, 13\}$ or for $P = 10$ and $N \in \{10, \dots, 16\}$),
- and not so well in comparison to collocation for others (typically for large $N > 17$ in the example of figure 5.9 right).

Kriging-gPC, *via* the introduction of an additional discretisation tool (the covariance kernel K), ensures a second convergence regime for $N \geq P$ in comparison to integration-gPC or regression-gPC which can both lead to approximations of stagnating accuracy. The choice of the covariance kernel, in this regime, strongly affects the convergence rate of the approximation (slope of the L^2 -norm of the error with respect to N) and can lead to better approximations than integration/regression/collocation-gPC if K is well suited. Note that K *needs*, this was especially emphasized in (5.40), to depend on N to make sure kriging-gPC outperform them $\forall N \in \mathbb{N}$.

5.4 Few other applications of gPC

In this last section, we first come back to the 'fil rouge' problem of chapter 2 before comparing gPC, regression, regression-gPC, collocation-gPC and kriging-gPC on a discontinuous solution.

5.4.1 Application to the 'fil rouge' problem of chapter 2

Now the methodology for solving non-intrusively an arbitrary uncertainty propagation problem presented, it only remains to apply it to our favorite configuration ('fil rouge' of chapter 2). Let us give few details about the choices made (experimental design, integration...):

- concerning the approximation basis, the initial random variable being a uniform law, we select the Legendre basis (see table 3.1) in order to apply non-intrusive (integration-)gPC.
- The next step consists in choosing the experimental design $(X_i, w_i)_{i \in \{1, \dots, N\}}$. We here select the Gauss-Legendre points and weights. They are effective in low stochastic dimension ($Q = 1$ for our 'fil rouge' problem), for smooth solutions and give satisfactory enough results for discontinuous one [277]. In this section, we take $N = 15$ Gauss-Legendre points in order to discretise $(X, d\mathcal{P}_X)$.
- We run N times the black-box simulation code at $(X_i, w_i)_{i \in \{1, \dots, N\}}$ in order to obtain the output points $(u(X_i), w_i)_{i \in \{1, \dots, N\}}$. Note that when one has access to computation clusters, those N runs can be launched simultaneously as they are independent. The methodology is said *embarrassingly parallel* as it does not require communication between the N processes. The computational efficiency is 100%, the communications being made only for the postprocessing step, assumed negligible here in comparison to one of the N runs. Concerning those N deterministic runs, they are performed so that we can consider the relative accuracy for each run with respect to spatial discretisation Δx on the observables of interest is about 10^{-4} .
- The postprocessing step consists in computing the $(u_k^{X, N})_{k \in \{0, \dots, P\}}$, building the gPC approximation $u_P^{X, N}(X)$ and recovering approximations of the mean profile, the variance one and the pdfs of the mass density at 3 points of interest at time $t = 0.14$ as in chapter 2.

The chosen polynomial order in the following computations is $P = 3$, kept voluntarily low (to make sure integration accuracy remains smaller than the truncation one). Figure 5.10 presents the results obtained in term of spatial profiles for three realisations (top pictures), the mean and the variance at $t = 0$ and $t = 0.14$ (bottom pictures). For the profiles of figure 5.10, the results are very interesting: indeed, with

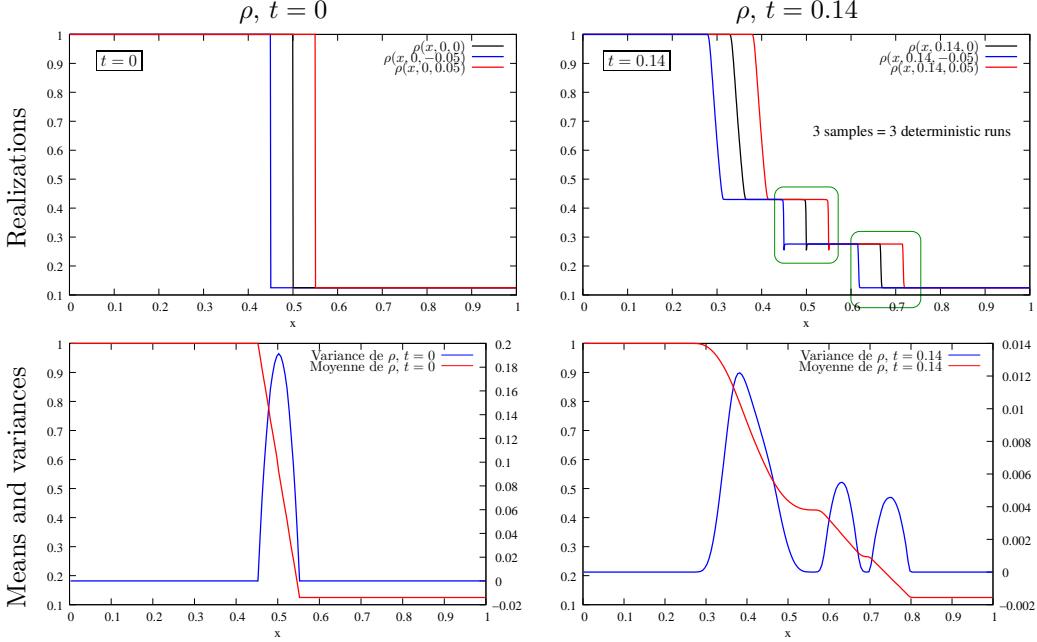


Figure 5.10: Application of non-intrusive gPC, mass density spatial profiles for some realisations, the mean and the variance at $t = 0$ and $t = 0.14$.

only $N = 15$ runs and $P = 3$, we are able to recover the mean and variance profiles with a good accuracy. In fact, the curves are not distinguishable from the ones obtained from the reference Monte-Carlo ones for these observables. This can be explained by the high accuracy of the Gauss quadrature rule.

Figure 5.11 presents the results in term of pdf and functional representation of the mass density in the vicinities of the rarefaction fan, the interface and the shock. The approximation in the vicinity of the rarefaction fan is accurate: it is not distinguishable from the reference Monte-Carlo computations. This is due to the smoothness of the solution in this area. The high accuracy in the vicinity of the rarefaction fan can even be more easily observed on the functional representation of $\rho(x = 0.38, t = 0.14, X)$ on figure 5.11 (top right). In this context, the non-intrusive gPC approach is very efficient even for such low polynomial order $P = 3$ as the gain in computational time for the same accuracy with respect to the MC approach is given by $\times \frac{N_{MC}}{N} = \frac{1000}{15} \approx 66.6$.

Concerning the approximations in the vicinities of the interface and the shock, we unfortunately do not observe the same gain. The discrete behaviour of the random variables is not captured by the non-intrusive gPC approximations. We do not detail more these results, they are very close to the ones encountered in the example of chapter 3, section 3.5.2 and we kind of expected such behaviour.

5.4.2 Integration vs. Regression vs. Collocation vs. Kriging vs. discontinuity

The first question arising after the previous study would be: if gPC fails to recover discontinuous solution, does any of the resolution schemes presented in section 5.3 allows dealing with it? To answer this question we suggest applying gPC, regression, regression-gPC, collocation-gPC and kriging-gPC to a discontinuous function $X \rightarrow \mathbf{1}_{]-\infty, \frac{3}{10}]}(X)$ with $X \sim \mathcal{U}([-1, 1])$.

Figure 5.12 presents the results obtained with gPC, regression, regression-gPC, collocation-gPC and kriging-gPC on the latter function. The presentation is slightly different than previously as the left column now displays the best approximations obtained with every methods with $N = 11$ (top) and $N = 21$ (bottom). The right column shows convergence studies with respect to P in the same conditions

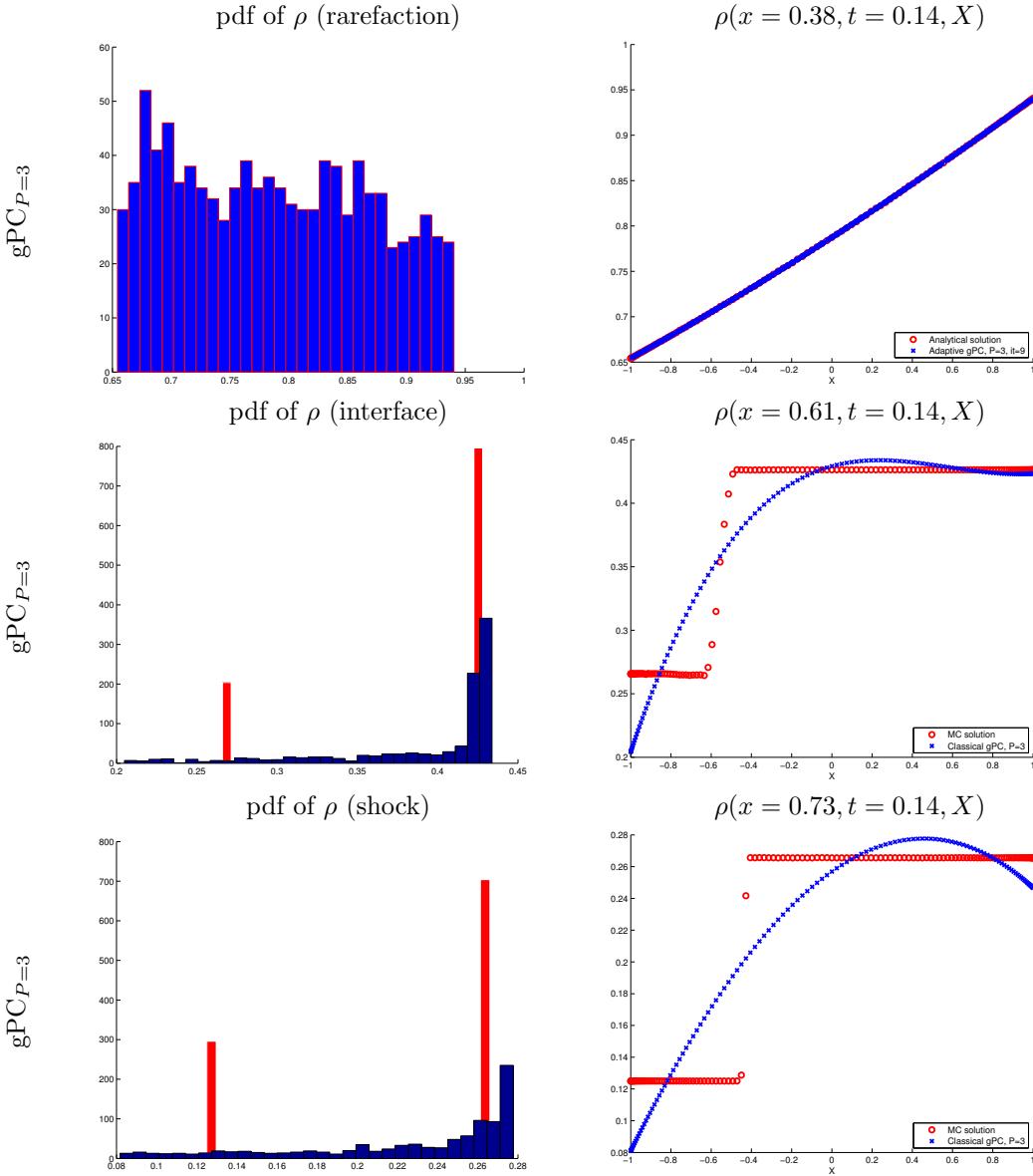


Figure 5.11: Pdfs (left) and functional representation in the random space of the mass density at $t = 0.14$ in the vicinities of the rarefaction fan ($x = 0.38$), the interface ($x = 0.61$) and the shock ($x = 0.73$) with the $gPC_{P=3}$ approximation.

with $N = 11$ (top) and $N = 21$ (bottom). Note that the convergence study obtained with collocation-gPC is not presented in the top right picture as it implies much more points (as $P = N$) than $N = 11$.

We suggest beginning by commenting the convergence studies of the right column of figure 5.12. For $N = 11$ (top right), gPC, regression, regression-gPC have the same behaviour: the error slightly decreases before exploding. The one of kriging-gPC is more singular: the error is way lower than for the other approximations for small P but increases as fast²⁵ as regression-gPC as soon as $P > N$. The top left picture of figure 5.12 presents the best approximations obtained with every methods with $N = 11$. For gPC, regression, regression-gPC, it corresponds to $P = 5$. For collocation-gPC, it corresponds to $N = P = 11$. For kriging-gPC, it corresponds to $P = 1$. First, the gPC and regression approximations perfectly match and, as expected, poorly recover the function of interest. Regression-gPC presents a

²⁵We once again recover numerically the fact that for $P \gg N$, the explosion rate is independent of the choice of the covariance kernel, see (5.41). Having the same behaviour for the Runge function *and* for the discontinuous one of this section, we also recover it is independent of u .

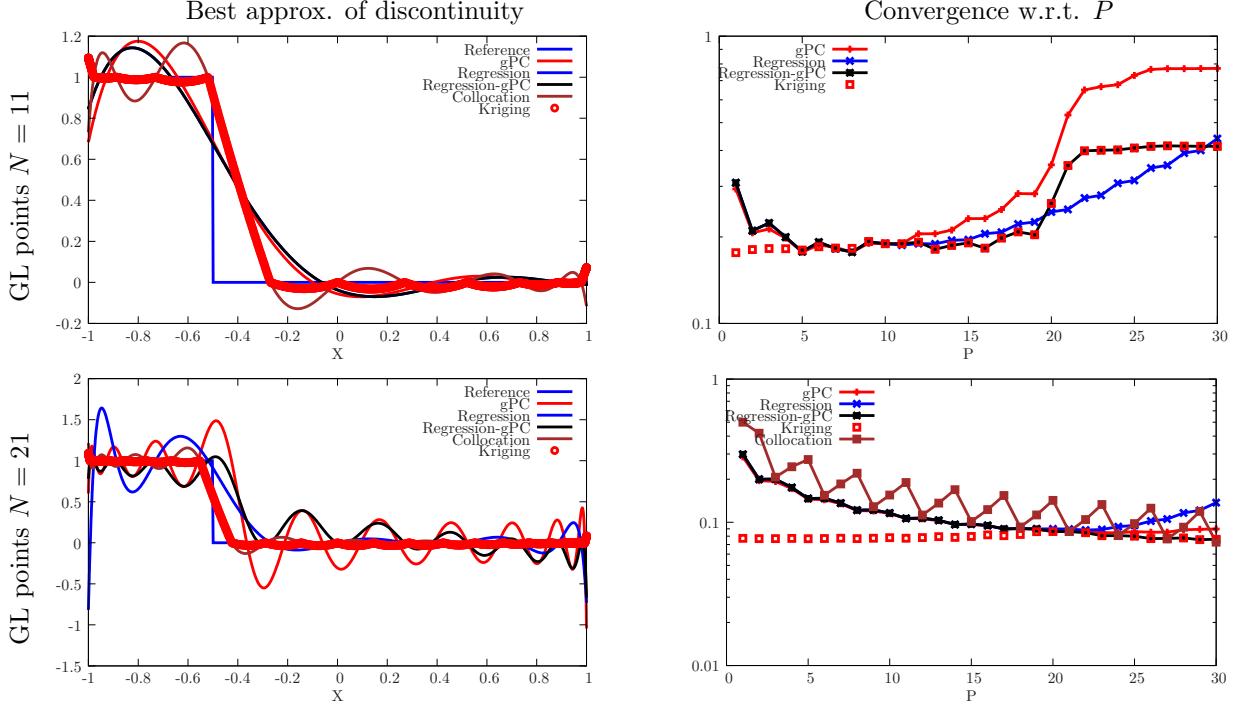


Figure 5.12: Application of Gauss-Legendre quadrature rule for *gPC*, *regression*, *regression-gPC*, *collocation-gPC* and *kriging-gPC* for the approximation of the transformation of a uniform random variable through a discontinuous function. The experimental designs have with $N = 11$ (top) and $N = 21$ (bottom). The left column present the best approximations obtained with every of the previous methods. The right column present the L^2 -norm of the error with respect to P for fixed N . The kriging kernel is chosen exponential (5.42) and a dichotomy is applied to calibrate θ .

similar behaviour as *gPC* and *regression*. *Collocation-gPC* and *kriging-gPC* coincide at the $N = 11$ points of the experimental design. *Kriging-gPC* is way less oscillatory but interpolates almost linearly between the points of the experimental design on both sides of the discontinuity.

For $N = 21$, the convergence studies of the resolution schemes is displayed on the bottom right picture of figure 5.12. First, the non-monotonous convergence of *collocation-gPC* is singular. Once again, *gPC* and *regression* present similar results with an increase of the error as P becomes greater than N . The error for *regression-gPC* is more controlled. The *kriging-gPC* error, even if the lowest amongst every methods, keeps increasing from $P = 1$ to $P = N$ before equaling the one for *regression-gPC* for $P > N$. The fact the error is the lowest for $P = 1$ testifies the covariance kernel ensures the accuracy for this function. This is interesting as it precisely relies on a smoothness hypothesis. Figure 5.12 bottom left presents the best results obtained with every methods: it corresponds to $P = 20$ for *gPC*, *regression* and *regression-gPC*, to $P = N = 21$ for *collocation-gPC* and to $P = 1$ for *kriging-gPC*. First, even if *every approximation has a quantitatively better accuracy (L^2 -norm) than in the previous case, qualitatively, the coarser approximations (for $N = 11$) seem better*. The *gPC*, *regression* and *regression-gPC* approximations are very oscillatory, especially in the vicinities of the boundaries of $[-1, 1]$. Surprisingly, *collocation-gPC* is less oscillatory than the three previous methods. The best approximation remains the one obtained with *kriging-gPC*, even if almost linearly interpolating between the points on each side of the discontinuity. The most accurate *kriging-gPC* approximation is once again obtained for $P = 1$: this implies the covariance kernel is responsible for this accuracy, more than the polynomial trend, even if relying on smoothness hypothesis of the solution.

5.5 Summary for non-intrusive gPC for systems of conservation laws

Let us sum up the sections concerning the choice of the experimental design and the gPC based post-processing to approximate non-intrusively random variable $u(X)$. We insist on two main points:

- first, on the complementarity of MC methods and Gauss quadrature rules for integration. The first one has a *slow but independent of the smoothness of the solution u and of the dimension of the uncertain variable X* convergence rate. The second one has a *fast but smoothness dependent and sensitive to the dimension* convergence rate. For these reasons, both are interesting but in complementary contexts. In this part II, as already explained in chapter 2, we deal with a small number of uncertain parameters and rely consequently more on Gauss quadratures than on MC points. Important dimension problems are tackled in part III with the resolution of the linear Boltzmann equation (in a Monte-Carlo context).
- Besides, regarding the choice of the postprocessed approximation, we performed the numerical analysis and compared integration-gPC, regression-gPC, collocation-gPC and kriging-gPC in the same conditions. We think these (analysis and comparisons) are original even if gPC and kriging-gPC approximations have already been experimentally compared [217, 253, 158, 258, 262] on many different statistical observables. The numerical analysis mainly aimed at identifying more easily under which conditions the strategies differ, are equivalent or are efficient. We also insist on the fact the comparisons have mainly been made in the L^2 -norm. Whether the different approximations bear interesting properties with respect to other norms is beyond the scope of this document but will probably be tackled in further researchs.

In practice, for the different studies we present in this document (mainly in the following chapters) and in view of the previous results, we systematically take $P \leq N$ in every stochastic directions and use Gauss quadrature rules. In such context, integration-gPC is equivalent to regression-gPC and even to collocation-gPC if $N = P$. This ensures fast postprocessings for the gPC-reconstruction of random variable $u(X)$ especially interesting when many outputs of interest must be approximated (in chapter 7 for example, we need a gPC approximation in every cell of a spatial discretisation at every time steps). Integration-gPC may be less flexible than regression-gPC with which it is easy *a posteriori* taking into account additional points of the experimental design. Kriging-gPC presents the advantages of both regression-gPC and collocation-gPC and may avoid stagnating approximations as $P \leq N$. This property is ensured thanks to the introduction of an additional discretisation tool (covariance kernel K) up to the cost of a more computational posttreatment. Such mathematical tool is obviously of great interest for industrial application but the numerical analysis and experiments of this section especially showed the starting points of the convergence rates of the approximations *for smooth solutions* is mainly dictated by the choice of the trend, i.e. the gPC basis. For *discontinuous solutions* the (exponential) covariance kernel plays a stronger role than expected. For these reasons, in the next chapters, we focus on the *polynomial* ingredient, at the basis of every of these methods and kickstart of every convergence curves: still, care will taken to design new numerical methods which will remain compatible with the possibility to enrich them with regression/collocation/kriging-gPC.

On the ‘fil rouge’ problem, the intrusive gPC approach (section 4.1) and the non-intrusive one do not have the same behaviour at all (in the vicinities of discontinuous solutions mainly). For the first one, we were able to detect an abnormal behaviour of the reduced model quite soon: the simulation code for the reduced model crashed at the first iteration and lead us to study and analyse more in detail what happened. Here, there are no robustness difficulties. This can be a pro but also a con as it can be hard determining *a posteriori* if the oscillatory²⁶ behaviour is physical or numerical. The method can lead to results which are not workable nor interpretable. On another hand, having at hand a black-box code solving the Euler equations, the non-intrusive approach is easy and fast to apply, especially if one has a computing cluster at disposition. It is probably the best way to tackle a punctual uncertainty quantification study, if care is taken to focus on smooth observables (an example of application is given in chapter 7). For systematic uncertainty quantification studies, the investment in intrusive

²⁶Note that with small orders P , the behaviour is not even really oscillatory for some approximations.

methods can become relevant: it forces a more deepened analysis of the interplay between the physics of interest, its numerical solvers and the uncertain parameters. An example of efficient intrusive application (more efficient than a non-intrusive application) of gPC is given in part III, chapter 9, section 9.11.2.

Regarding the results of the different methods on discontinuous solutions, we were obviously not satisfied with their respective behaviours on the 'fil rouge' application. This is all the more frustrating as the discontinuities are very localized, the solutions being smooth almost everywhere else. But they are often the locations of interests. At this stage, there is still one interesting advantage of gPC which has not yet been exploited (not even in the literature to our knowledge): in chapter 3, we insisted on the fact that some basis where more efficient than others in order to approximate an output random variable. In the next section, we show how we tried to integrate this *a priori* knowledge into a new approximation method.

Chapter 6

The non-intrusive iterative gPC (i-gPC) approach

Let's play with orthonormal basis

Contents

6.1	The main idea behind iterative-gPC (i-gPC)	103
6.1.1	A particular change of variable $Z(X)$ ensuring a gain	106
6.1.2	Description of the i-gPC approximation algorithm	106
6.1.3	Weak contraction of the i-gPC approximation	107
6.2	Application of i-gPC on two simple test-problems	108
6.2.1	Discontinuous output random variable	108
6.2.2	Smooth output random variable	109
6.3	Numerical analysis of i-gPC in finite integration context (stopping criterion)	111
6.3.1	Convergence behaviour of i-gPC under finite numerical integration accuracy	113
6.3.2	Strategy for adaptive approximation truncation	114
6.4	Few other applications of i-gPC	117
6.4.1	Application to the 'fil rouge' configuration	117
6.4.2	Integration vs. Regression vs. Collocation vs. Kriging vs. i-gPC	118
6.5	Summary for non-intrusive gPC and i-gPC approximations	119

As explained in the previous section, the application of non-intrusive gPC to our 'fil rouge' configuration was not satisfactory enough, mainly due to the (lack of) accuracy in the vicinities of the discontinuities. The polynomial order was kept low, $P = 3$, but increasing P implies both aliasing problems with respect to the construction of the orthonormal polynomial basis (see section 3.4) and an unaffordable amount of points $(X_i, w_i)_{i \in \{1, \dots, N\}}$ in order to accurately estimate the coefficients¹ $(u_k^{X,N})_{k \in \{0, \dots, P\}}$. In the next paragraphs, we detail our main contribution to non-intrusive gPC. More details can be found in [238, 242] and in [30] in collaboration with A. Birolleau, former PhD student.

6.1 The main idea behind iterative-gPC (i-gPC)

The main idea of the following material comes from the observation that for a given problem, different basis perform differently. In table 6.1, we recalled the results of section 3.2.2 in which we applied the PC approximation, the gPC one and added the Chebyshev approximation for the problem of transforming a

¹remember we have to keep $P \leq N$, see section 5.2.

uniform random variable $X \sim \mathcal{U}[0, 1] \rightarrow u(X) = \sin(2\pi X) \sim \mathcal{A}$ into an Arcsinus one. On this example, gPC performs better than PC only by using the basis orthonormal with respect to the input random variable X instead of the Hermite basis. With the Chebyshev basis (chosen according to table 3.1 and

input r.v.	$X \sim \mathcal{U}([0, 1])$		
transformations	$u(X) = \sin(2\pi X)$		
Corresponding laws	$u(X) \sim \mathcal{A}$ with pdf $f(x) = \frac{1}{\pi\sqrt{1-x^2}}$		
exact mean	0		
exact variance	$\frac{1}{2}$		
r.v./approximation basis	Gauss./Hermite (PC)	Unif./Legendre (gPC)	Arcsin/Chebyshev (opt.)
gPC	(order $P = 9$)	(order $P = 9$)	(order $P = 1$)
approximated mean	0	0	0
approximated variances	0.47769	0.5000000001	$\frac{1}{2}$

Table 6.1: Reminder of the results of section 3.3 regarding the performances of the PC basis, the gPC one and the optimal one on the same test-case (Case 2).

because here we know the output distribution), the results are optimal: with $P = 1$, the approximation is exact. Table 6.1 presents some quantitative results on the same transformation $X \rightarrow \sin(2\pi X)$, applying three different approximation basis: the Hermite, Legendre and Chebyshev ones. The gPC coefficients are analytically computed, presenting what can asymptotically be obtained with an infinitely accurate integration (it only remains the truncation error cf. (5.5)). With the Chebyshev basis, the convergence rate is optimal in the sense the development only needs two coefficients (or $P = 1$) to be analytical. Such behaviour is in fact more than a simple observation applying only for the Arcsinus law:

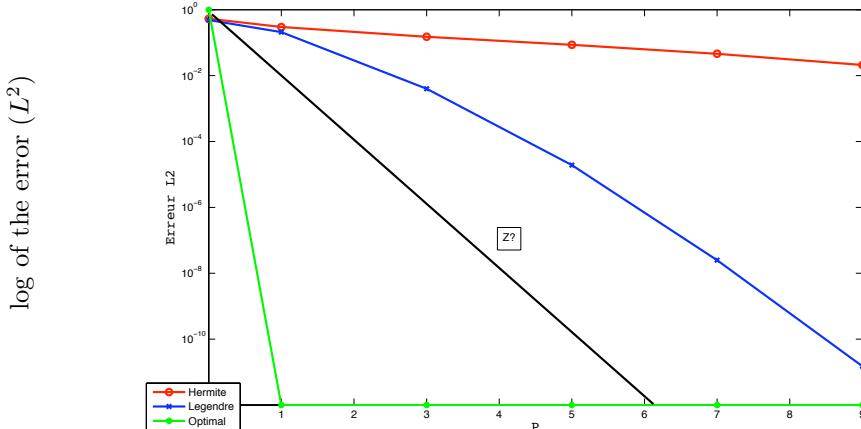


Figure 6.1: Spectral convergence of the PC basis, the gPC one, the optimal one and the question of building an intermediary basis giving at least better results than the gPC one if it can not recover the optimal one.

it occurs to any output random variable with finite second moment, see [291, 305, 104]. There exists an optimal basis for every transformation from any random variable X to any L^2 bounded one $u(X)$. The existence of such basis allows trying to find it, or at least approach it. The idea is summed up in figure 6.1 where the convergence rate of PC is exponential but slow whereas the gPC one is better. The optimal basis is represented but we can imagine there is a kind of continuum (?) of basis intermediary between the gPC one and the optimal one. The question now is *how can we find and build some of them?* In the following lines, we suggest one way to answer this question.

We aim at improving the accuracy of the approximation in a new gPC basis with respect to the gPC

one associated to the input random variable X . In term of numerical analysis, this implies considering two gPC approximations:

$$u(X) \simeq u_P^X(X) = \sum_{k=0}^P u_k^X \phi_k^X(X), \quad u_k^X = \mathbb{E}[u(X)\phi_k^X(X)], \quad (6.1)$$

$$u(X) \simeq u_P^Z(Z) = \sum_{k=0}^P u_k^Z \phi_k^Z(Z), \quad u_k^Z = \mathbb{E}[u(X)\phi_k^Z(Z)]. \quad (6.2)$$

The first basis (6.1) is the classical gPC one, orthonormal with respect to the inner product of the *a priori known* probability measure $d\mathcal{P}_X$ of the input random variable. The second one (6.2) is also a gPC basis in the sense it is orthonormal with respect to an inner product defined by a probability measure $d\mathcal{P}_Z$ of a random variable Z . At this stage, Z remains to be determined. We aim at studying the difference of the L^2 errors in the two gPC basis. For this, (the computations are detailed in [238]) we classically introduce $u_P^X(X)$ into the L^2 -norm of the error in the gPC basis associated to Z

$$\|u(X(Z)) - u_P^Z(Z)\|_2^2 = \|u(X(Z)) - u_P^X(X(Z)) + u_P^X(X(Z)) - u_P^Z(Z)\|_2^2.$$

We now expand the previous expression to make

$$\Delta_{Z,X}^P = \|u(X(Z)) - u_P^Z(Z)\|_2^2 - \|u(X) - u_P^X(X)\|_2^2,$$

appear. In the above expressions, the difference of errors has been expressed with respect to $\Delta_{Z,X}^P$ which, once expanded leads to

$$\begin{aligned} \Delta_{Z,X}^P &= \mathbb{E} \left[(u_P^X(X(Z)) - u_P^Z(Z))^2 \right] + 2\mathbb{E} \left[(u_P^X(X(Z)) - u_P^Z(Z)) (u(X(Z)) - u_P^X(X(Z))) \right] \\ &= \underbrace{\mathbb{E} \left[(u_P^X(X))^2 \right]}_{e_1} - \underbrace{2\mathbb{E} \left[u_P^X(X(Z)) u_P^Z(Z) \right]}_{2e_2} + \underbrace{\mathbb{E} \left[(u_P^Z(Z))^2 \right]}_{e_3} - 2e_1 + 2e_2 - 2e_4 + 2e_5. \end{aligned}$$

We intensively use the orthonormality of both basis to show that for any arbitrary mapping $X(Z)$:

$$\begin{aligned} - e_1 &= \mathbb{E} \left[(u_P^X(X))^2 \right] = \sum_{k=0}^P (u_k^X)^2, \\ - e_3 &= \mathbb{E} \left[(u_P^Z(Z))^2 \right] = \sum_{k=0}^P (u_k^Z)^2, \\ - e_5 &= \mathbb{E} \left[u_P^X(X) u(X) \right] = \sum_{k=0}^P u_k^X \underbrace{\mathbb{E} \left[\phi_k^X(X) u(X) \right]}_{u_k^X} = e_1, \\ - e_4 &= \mathbb{E} \left[u_P^Z(Z) u(X(Z)) \right] = \sum_{k=0}^P u_k^Z \underbrace{\mathbb{E} \left[\phi_k^Z(Z) u(X(Z)) \right]}_{u_k^Z} = e_3. \end{aligned}$$

With the above results, the difference of errors in L^2 -norm leads to

$$\|u(X) - u_P^Z(Z)\|_2^2 - \|u(X) - u_P^X(X)\|_2^2 = \sum_{k=1}^P (u_k^X)^2 - \sum_{k=1}^P (u_k^Z)^2. \quad (6.3)$$

The above expression is valid for an *arbitrary* choice of the random variable Z . In (6.3), we used the fact that the mean in the Z -basis equals the mean in the X -one independently of the choice of Z (i.e. $u_0^Z = u_0^X$). Result (6.3) is singular and, to our knowledge, original. The question now is, is it possible to wisely choose Z with respect to X to make sure the approximation in the new basis gives better performances than in the initial one? The answer is the purpose of the next section.

6.1.1 A particular change of variable $Z(X)$ ensuring a gain

The question now is can we choose $Z(X)$ such that

$$\|u(X) - u_P^Z(Z)\|_2^2 - \|u(X) - u_P^X(X)\|_2^2 = \sum_{k=1}^P (u_k^X)^2 - \sum_{k=1}^P (u_k^Z)^2 \leq 0? \quad (6.4)$$

In [238, 242], we put forward one possible change of variable $Z(X)$ having the desired effect. Taking $Z(X) = u_P^X(X)$ ensures a gain. Let us explain how:

- first, let us introduce $s_k^Z = \int z^k d\mathcal{P}_Z(z)$ the statistical moments of Z . Then the orthonormal basis $(\phi_k^Z)_{k \in \{0, \dots, P\}}$ has general term (Christoffel)

$$\forall k \in \{0, \dots, P\}, \phi_k^Z(z) = \frac{1}{\sqrt{\underline{H}_{2(k-1)}^Z \underline{H}_{2k}^Z}} \begin{vmatrix} s_0^Z & s_1^Z & \dots & s_k^Z \\ \dots & \dots & \dots & \dots \\ s_n^Z & s_{n+1}^Z & \dots & s_{n+k}^Z \\ \dots & \dots & \dots & \dots \\ 1 & z^1 & \dots & z^k \end{vmatrix},$$

where $\underline{H}_{2k}^Z(s_0^Z, \dots, s_{2k}^Z)$ are the Hankel determinants, see section 3.4.

- With the choice $Z(X) = u_P^X(X) = \sum_{k=0}^P u_k^X \phi_k^X(X)$, the first three moments of Z are given by

$$s_0^Z = 1, \quad s_1^Z = u_0^X, \quad \text{and} \quad s_2^Z = \sum_{k=0}^P (u_k^X)^2.$$

- It implies the second component of the new basis $(\phi_k^Z)_{k \in \{0, \dots, P\}}$ can be expressed as

$$\phi_1^Z(z) = \frac{1}{\sqrt{s_0^Z \begin{vmatrix} s_0^Z & s_1^Z \\ s_1^Z & s_2^Z \end{vmatrix}}} \begin{vmatrix} s_0^Z & s_1^Z \\ 1 & z \end{vmatrix} = \frac{1}{\sqrt{\sum_{k=1}^P (u_k^X)^2}} (z - u_0^X).$$

- The second gPC coefficient in the new basis consequently reads

$$u_1^Z = \mathbb{E} \left[u(X) \phi_1^Z \left(\sum_{k=0}^P u_k^X \phi_k^X(X) \right) \right] = \frac{1}{\sqrt{\sum_{k=1}^P (u_k^X)^2}} \sum_{k=1}^P u_k^X \mathbb{E} [u(X) \phi_k^X(X)] = \sqrt{\sum_{k=1}^P (u_k^X)^2}.$$

If we now use this particular choice in (6.3), we get

$$\|u(X) - u_P^Z(Z)\|_2^2 - \|u(X) - u_P^X(X)\|_2^2 = \sum_{k=1}^P (u_k^X)^2 - (u_1^Z)^2 - \sum_{k=2}^P (u_k^Z)^2 = - \sum_{k=2}^P (u_k^Z)^2 \leq 0, \quad (6.5)$$

independently of any regularity assumptions on u . The inequality has been set assuming perfect numerical accuracy here but the numerical analysis in a finite integration context has been deepened in [242]. We insist some other choices $Z(X)$ may be better and this is ongoing research on the topic. In the following section, we suggest one way to exploit inequality (6.5) in a new algorithm.

6.1.2 Description of the i-gPC approximation algorithm

Equation (6.5) naturally represents the first step of an iterative gPC (i-gPC) algorithm whose iteration can be described in 5 points:

- first, build the classical non-intrusive gPC approximation: $u(X) \approx u_P^X(X) = \sum_{k=0}^P u_k \phi_k^X(X)$.

- Secondly, introduce the new random variable $Z^1 = u_P^X(X) = \sum_{k=0}^P u_k \phi_k^X(X)$, and build the gPC basis orthonormal with respect to the probability measure $d\mathcal{P}_{Z^1}$ of Z^1 , i.e. such that

$$\int \phi_k^{Z^1} \phi_t^{Z^1} d\mathcal{P}_{Z^1} = \delta_{k,t}, \forall (k,t) \in \{0, \dots, P\}^2.$$

We recall we do not need to explicitly estimate $d\mathcal{P}_{Z^1}$ in order to build this orthonormal basis $(\phi_k^{Z^1})_{k \in \{0, \dots, P\}}$ (see section 3.4). In fact, we estimate the moments of Z^1 thanks to the chosen experimental design and apply the algorithms of section 3.4 in order to build the new basis.

- Once the basis built, we need to estimate the coefficients of the random variable $u(X)$ in the new basis $(\phi_k^{Z^1})_{k \in \{0, \dots, P\}}$. By definition, they are given by:

$$u_k^{Z^1} = \int u(X(Z^1)) \phi_k^{Z^1}(Z^1) d\mathcal{P}_{Z^1}.$$

A change of variable allows expressing the coefficients $(u_k^{Z^1})_{k \in \{0, \dots, P\}}$ with respect to the probability measure $d\mathcal{P}_X$ of the input X :

$$u_k^{Z^1} = \int u(X(Z^1)) \phi_k^{Z^1}(Z^1) d\mathcal{P}_{Z^1} \stackrel{[107]}{=} \int u(X) \phi_k^{Z^1}(Z^1(X)) d\mathcal{P}_X.$$

We insist this change of variable is exact and is not induced by any assumption or approximation, see [107]. This implies we can estimate the coefficients in the new basis from the same experimental design $(X_i, w_i)_{i \in \{1, \dots, N\}}$ used for the initial gPC approximation during the first step. We consequently have

$$u_k^{Z^1} = \int u(X) \phi_k^{Z^1}(u_P^X(X)) d\mathcal{P}_X \approx \sum_{i=1}^N \textcolor{blue}{u}(\textcolor{blue}{X}_i) \phi_k^{Z^1}(\textcolor{magenta}{Z}^1(\textcolor{violet}{X}_i)) \textcolor{blue}{w}_i = u_k^{Z^1,N}. \quad (6.6)$$

The possibility to reuse the same points and weights at each iteration is important in practice as it implies the new iterative algorithm is still a postprocessing of the initial experimental design (see $(\textcolor{blue}{w}_i, \textcolor{blue}{u}(\textcolor{blue}{X}_i))_{i \in \{1, \dots, N\}}$ in (6.6)). It does not need anymore runs of the simulation code. Expression (6.6) also allows highlighting the convenience of having an explicit and fast to estimate expression of $X \rightarrow Z^1(X)$ (see $\textcolor{magenta}{Z}^1(\textcolor{violet}{X}_i)$) to perform the computations of the new coefficients. Note that we never need the inverse of $X \rightarrow Z^1(X)$ in practice.

- The last step of the iterative process implies the construction of the new approximation in the new basis $(\phi_k^{Z^1})_{k \in \{0, \dots, P\}}$ with the approximated coefficients $(u_k^{Z^1,N})_{k \in \{0, \dots, P\}}$:

$$u(X) \approx u_{P,N}^{Z^1}(Z^1) = \sum_{k=0}^P u_k^{Z^1,N} \phi_k^{Z^1}(Z^1).$$

- The rest resumes to looping on the Z^j to iterate on the different gPC basis.

Inequality (6.5) together with the above process leads to what we call an *i-gPC approximation at iteration j*, denoted by $u_{Z^j}^P(Z^j(X))$ in the following section.

6.1.3 Weak contraction of the i-gPC approximation

The previously presented *i-gPC approximation at iteration j*, denoted by $u_{Z^j}^P(Z^j(X))$, has the property (see [238]) of a weak contraction. With the previous notations, in the corresponding basis, we have at the j^{th} iteration²:

$$\|u(X) - u_{Z^j}^P(Z^j)\|_{L^2(\Omega)} \leq \|u(X) - u_{Z^{j-1}}^P(Z^{j-1})\|_{L^2(\Omega)}. \quad (6.7)$$

²still assuming perfect integration accuracy.

Expression (6.7) does not depend on any (additional) regularity hypothesis on $X \rightarrow u(X) \in L^2(\Omega)$. It ensures that at the j^{th} iteration, the approximation in the new basis is better or at least not worse than the approximation at the $(j - 1)^{th}$ iteration.

Based on the described algorithm and (6.7), we suggest studying and analysing the non-intrusive i-gPC approximation. For this, we once again adopt a step-by-step analysis. As in chapter 3, we first assume the coefficients are very well approximated (analytically whenever it is possible). This analysis will show what can be asymptotically obtained if the coefficients are accurate enough. Note that we just described an iterative algorithm *without* any stopping criterion: we will spend some time on it later, as it will closely intertwined with the integration error in the coefficients $(u_k^{Z^j, N})_{k \in \{0, \dots, P\}}$.

6.2 Application of i-gPC on two simple test-problems

The first test-problem we consider is the same as in section 3.7. It has been introduced to illustrate the sensitivity of gPC to Gibbs phenomenon. The second one corresponds to the worst case scenario we experimentally encountered when applying i-gPC.

6.2.1 Discontinuous output random variable

We consider the transformation of a uniform random variable $X \sim \mathcal{U}([-1, 1]) \rightarrow u(X)$ into a binomial one, where u is defined as

$$u(x) = \mathbf{1}_{[-\infty, -\frac{1}{2}]}(x) = \begin{cases} 1 & \text{if } x \leq -\frac{1}{2}, \\ 0 & \text{else.} \end{cases} \quad (6.8)$$

Figure 6.2 recalls the results obtained with gPC. It corresponds to the first step of our iterative algorithm. At the first iteration, the gPC approximation poorly captures the discrete behaviour of the output random variable. Figure 6.3 shows, in functional representation, how the successive i-gPC approximations behave

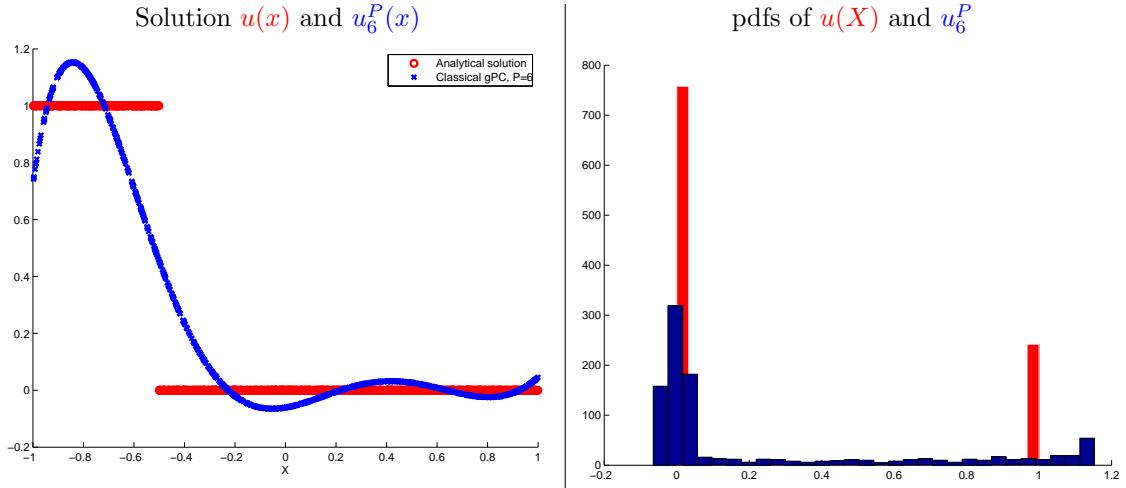


Figure 6.2: Application of the first step of i-gPC (=gPC) on the transformation of a uniform law into a binomial one for $P = 6$. The gPC_6 approximation is very sensitive to the Gibbs phenomenon

iteration after iteration on this discontinuous problem: the two states of the discontinuity are better and better captured. The i-gPC approximations do not respect the maximum principle, the amplitude of the oscillations are more important at iteration #2 than iteration #1 but the discontinuity location is more and more accurate. At iteration #4, there are only 5 points between the two continuous states 1 and 0. Figure 6.4 (left) presents the histogram of the reference solution obtained with an MC method (reference) together with the one obtained with the $\text{i-gPC}_{P=6}^{Z^{k=5}}$ approximation: the Dirac masses are very accurately captured, both their masses and their locations. We insist these results were obtained in

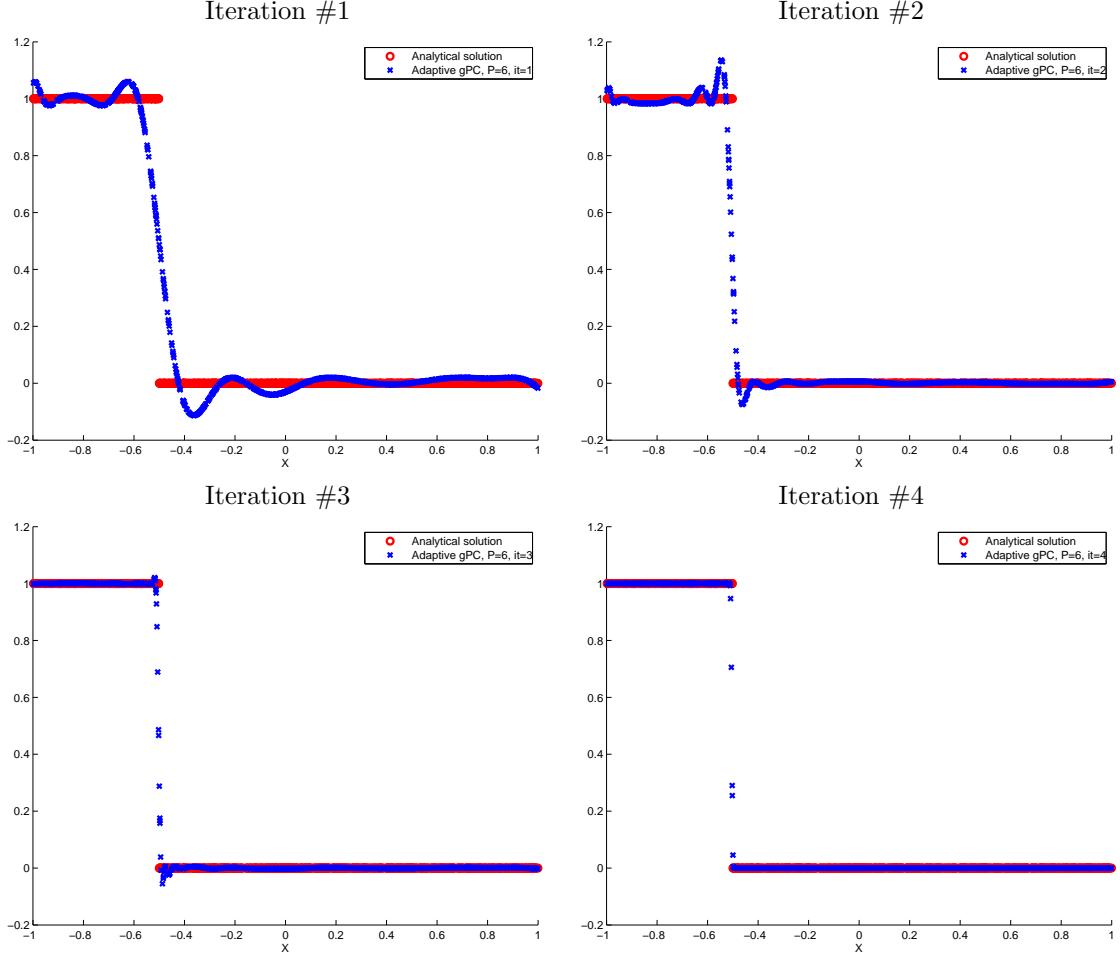


Figure 6.3: Functional representation of the i-gPC approximation through iterations 1, 2, 3 and 4.

exactly the same condition as the gPC approximation of figure 6.2 in term of polynomial order $P = 6$ and integration points N (even if we took $N \gg 1$ here). Figure 6.4 (right) presents convergence results in the logarithm of the L^1 -norm of the error with respect to the number of iterations k for fixed orders $P = 3$ up to $P = 10$. On this picture, one can notice that every approximations, for every truncation orders P , converge toward the same error once given enough iterations. For $P = 3$, it needs more iterations (12) than for $P = 6$ (for which it needs only 7 iterations) or $P = 10$ (for which only 5 iterations are enough). In fact, in [242], we showed numerically that the error at the end of the iterative process is independent of the truncation order P . For transformation (6.8), the ideal situation with a strict inequality (6.7) for every performed iterations is encountered, the inequality is a strong contraction:

$$\|u(X) - u_{Z^j}^P(Z^j)\|_{L^2(\Omega)} < \|u(X) - u_{Z^{j-1}}^P(Z^{j-1})\|_{L^2(\Omega)}. \quad (6.9)$$

If the above observation holds for every iteration j , increasing the number of iteration j ensures the convergence of the approximation $\|u(X) - u_{Z^j}^P(Z^j)\|_{L^2(\Omega)} \xrightarrow{j \rightarrow \infty} 0$ for the fixed truncation order P . In [242], we showed that on this specific test problem, the iterative process allows recovering the Krawtchouk basis, optimal with respect to the Binomial law of the considered output $u(X)$, see table 3.1.

6.2.2 Smooth output random variable

The second test-problem we want to tackle in this section corresponds to the worst case scenario we encountered. For this, we consider the transformation of a uniform random variable $X \sim \mathcal{U}([-1, 1]) \rightarrow$

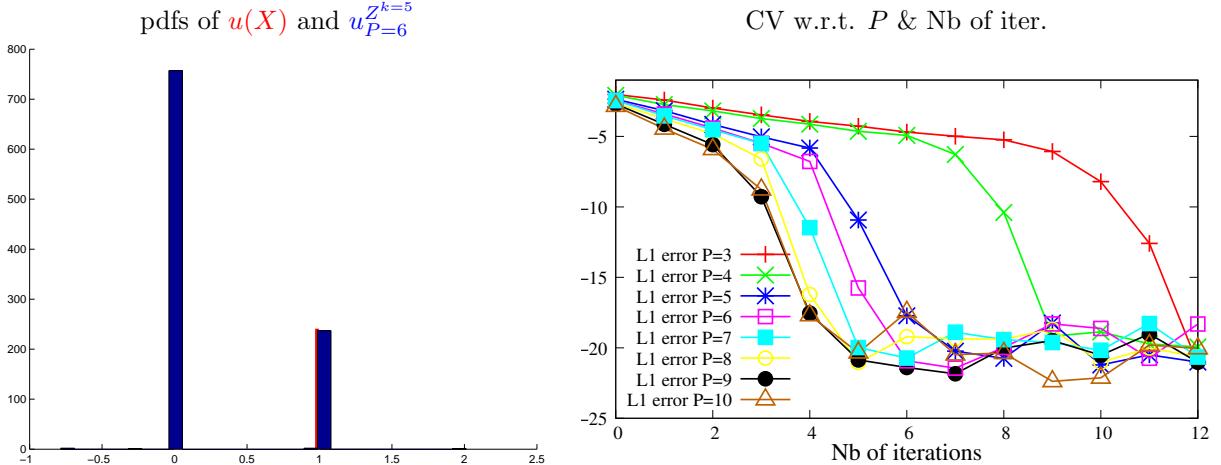


Figure 6.4: Left: histogram of the reference solution obtained with an MC method and of the i-gPC $_{P=6}^{Z^{k=5}}$. Right: Convergence plot showing the L^1 error w.r.t. the number of iterations.

$u(X)$ into a polynomial of order 10. It is given by

$$u(X) = \phi_0^L(X) + \phi_3^L(X) + \phi_{10}^L(X). \quad (6.10)$$

In (6.10), $(\phi_k^L(X))_{k \in \{0, \dots, P\}}$ denotes the Legendre polynomials. We apply gPC and then i-gPC for a truncation order $P = 8$. Note that if $P = 10$, both gPC and i-gPC (at the first iteration) give the analytical solution on this test-problem. We choose, on purpose, to truncate the basis earlier, for $P = 8$. Figure 6.5 presents the results in term of functional representation and histogram obtained with the

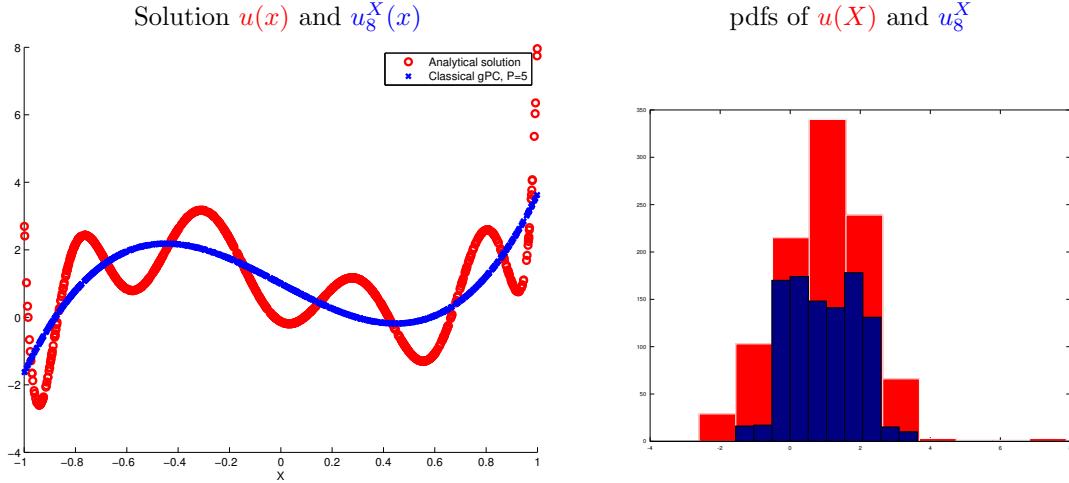


Figure 6.5: Left: functional representations of the analytical solution (6.10) and its gPC approximation of order $P = 5$. Right: histograms of the analytical solution (6.10) and its gPC approximation of order $P = 5$.

MC reference method and with the gPC $_{P=8}$ approximation. As observed in figure 6.5 (left) for the functional representation, the gPC $_{P=8}$ approximation is equivalent to the gPC $_{P=3}$ one. The terms of orders between 4 and 8 are orthogonal to the solutions and the coefficients are zero in this basis for $k \in \{4, \dots, 8\}$. Figure 6.6 (left) presents the same results in term of functional representation as figure 6.5 (left) obtained with i-gPC $_{P=8}^{Z^{k=10}}$ in the same conditions. Qualitatively, we observe a gain with respect to gPC. It is observable mainly in the vicinity of the boundaries of the random space $[-1, 1]$. Now figure 6.6 (right) presents quantitative results. It is a convergence study displaying the logarithm of the L^1 -norm of the error with respect to the number of iterations from $P = 5$ up to $P = 9$. We can see that the gain observable on figure 6.6 (left) for $P = 8$ is obtained at the first iteration and the next ones did not

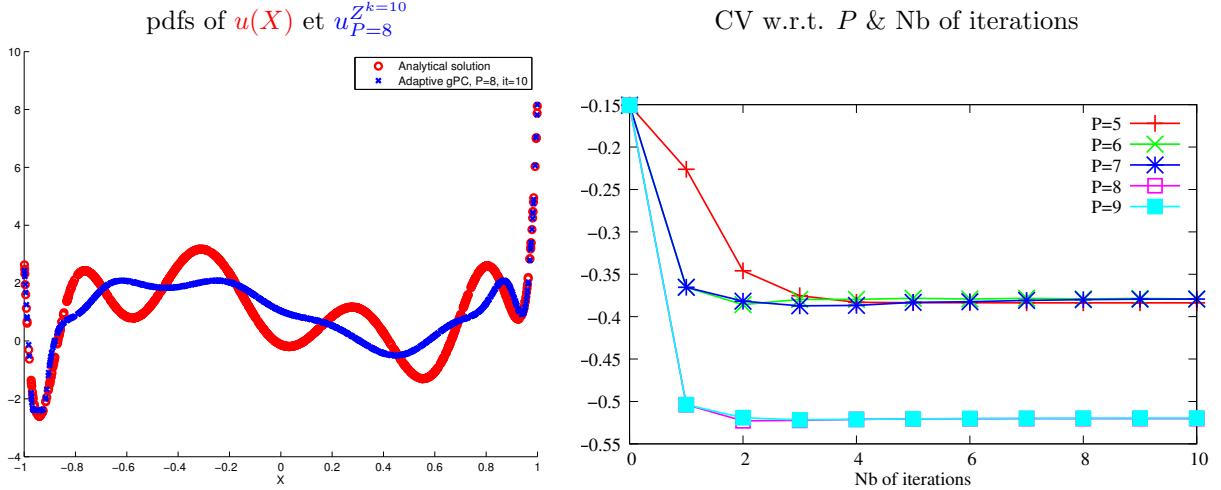


Figure 6.6: Left: functional representations of the analytical solution (6.10) and its i-gPC approximation of order $P = 5$ after 10 iterations. Right: convergence results with respect to the number of iterations k for fixed orders $P = 5, \dots, 9$.

improve the accuracy of the approximation. In this case, i-gPC allows a small gain after the first iteration then stagnates. Nonetheless it corresponds to the worst case scenario in the sense the gain is small and the stagnation occurs for $k_0 = 1$. In this situation, we can still rely on the convergence property with respect to P of the gPC approximation. The iterative algorithm has been designed in order to ensure that at each step, we rely on a gPC approximation (at iteration j , the approximation is still a gPC one). We consequently can still rely on Cameron-Martin's theorem.

Of course, in this section, we depicted two very opposite situations:

- the first one corresponds to the ideal one, the i-gPC procedure ensures recovering the optimal basis, see [242] for more details.
- The second one corresponds to the case where the iterative procedure stagnates after the first iteration with a very small increase of the accuracy.

There exists some intermediary states to these situations (some are presented in [238, 242] and others in chapter 8). But we wanted to focus on the two previous ones especially because of the respective regularity of the transformations: the iterative process does not necessarily performs better on smooth transformations. To sum up, with the strong hypothesis of having a very accurate numerical integration method, we are able to emphasize that i-gPC

- gives (very) satisfactory results on discontinuous solutions,
- allows a gain even on smooth ones,
- is only a postprocess and do not need additional points with respect of a gPC approximation.

These (asymptotical) results are, to our opinion, very interesting and motivate us to continue the study of the iterative approach in the context of a finite numerical integration accuracy.

6.3 Numerical analysis of i-gPC in finite integration context (stopping criterion)

In this section, we revisit inequality (6.7) in the context of finite numerical integration accuracy. The study has been carried on in [30], [242]. In order to understand the need for the following numerical analysis, we consider the same test-cases as in section 6.2 but with a more reasonable number of points

N for our experimental design $(X_i, w_i)_{i \in \{1, \dots, N\}}$. Let us first consider the smoother case, i.e. the one where u is defined as

$$X \sim \mathcal{U}[-1, 1] \longrightarrow u(X) = \phi_0^L(X) + \phi_3^L(X) + \phi_{10}^L(X).$$

We now apply the same algorithm but with only $N = 17$ Clenshaw Curtis (CC) points for the experimental design (see section 5.2) and $P = 5$. Figure 6.7 displays the functional representation of the analytical solution, the gPC $_{P=5}$ approximation and the i-gPC $_{P=5}^{Z^k}$ for $k \in \{1, \dots, 6\}$. The quadrature points are also

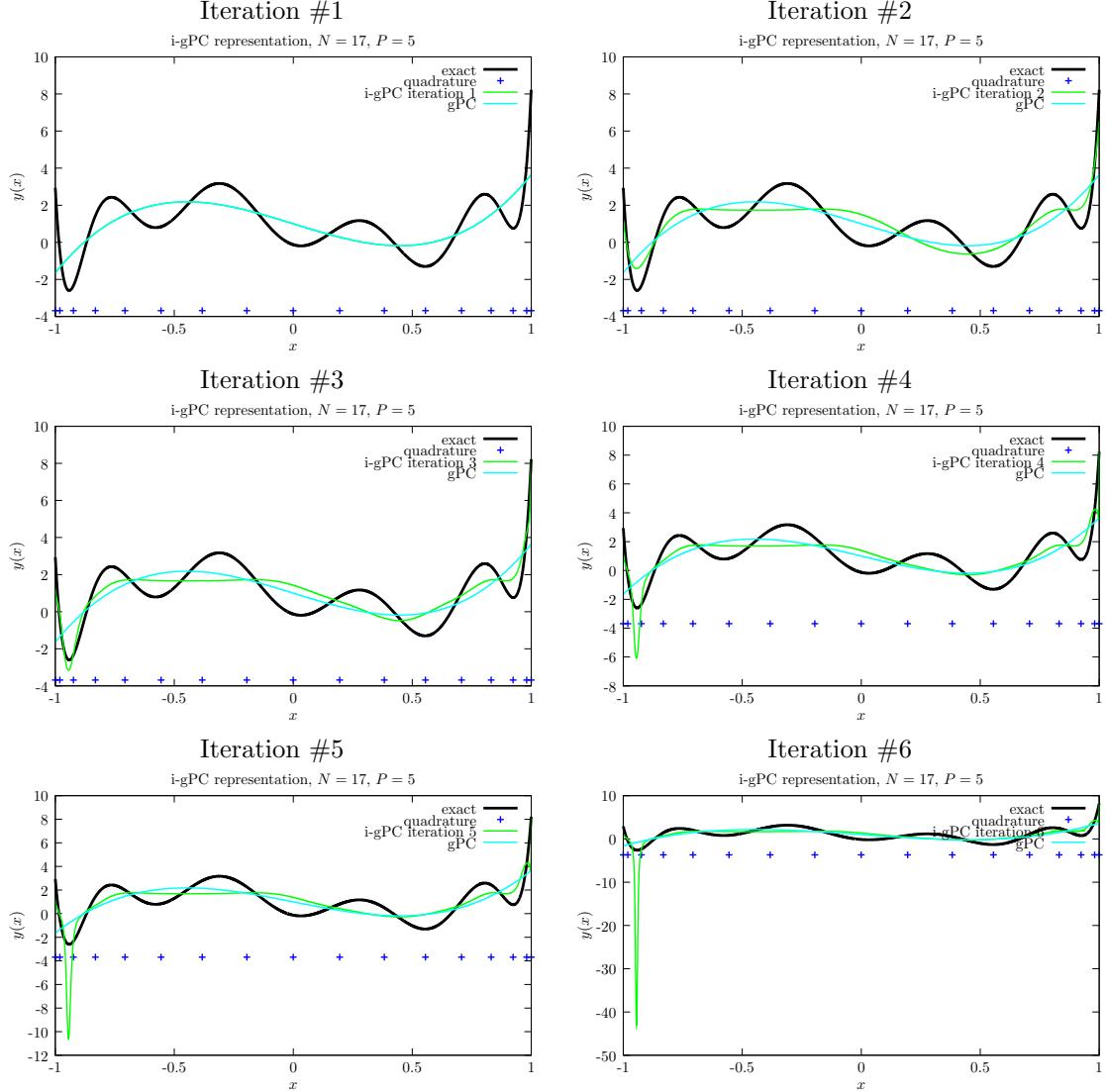


Figure 6.7: Functional representation of the i-gPC approximation through iterations 1, 2, 3, 4, 5 and 6 for the smooth problem $u(X) = \phi_0^L(X) + \phi_3^L(X) + \phi_{10}^L(X)$ with finite integration accuracy (without proper stopping criterion).

displayed on the pictures (blue crosses at the bottom of each pictures). At the first iteration, the gPC and i-gPC approximations are the same. After iterations #2 and #3, we can observe a gain of the i-gPC approximation with respect to the gPC one, especially on the boundaries of the uncertain domain $[-1, 1]$ (as in the case of infinite integration accuracy in section 6.2). But at iteration #4, in the vicinity of $X \approx -1$, the i-gPC solution does not anymore get closer to the target solution. And after iterations #5 and #6, a numerical instability can easily be identified, developing from the small initial perturbation appearing at iteration #4. In fact, we observe here that the error can first decrease before increasing.

This behaviour was not predicted by inequality (6.7). We consequently need to introduce the integration accuracy of the experimental design $(X_i, w_i)_{i \in \{1, \dots, N\}}$ in our numerical analysis to understand and analyse the previous behaviour.

6.3.1 Convergence behaviour of i-gPC under finite numerical integration accuracy

According to Cameron Martin's theorem and its generalization to arbitrary pdfs/polynomial basis [55, 291] together with the hypothesis of convergence of the chosen experimental design, the following approximations

$$\begin{aligned} u(X) \approx u_X^{P,N}(X) &= \sum_{k=0}^P u_k^{X,N} \phi_k^X(X) = Z(X), \\ u(Z) \approx u_Z^{P,N}(Z) &= \sum_{k=0}^P u_k^{Z,N} \phi_k^Z(Z), \end{aligned} \quad (6.11)$$

converges and their errors can be decomposed into an integration and a truncation one:

$$\begin{aligned} (e_X^N)^2 &= \|u(X) - u_X^{P,N}(X)\|_{L^2}^2 = \underbrace{\sum_{k=0}^P (u_k^X - u_k^{X,N})^2}_{(e_{\text{int},P}^{X,N})^2} + \sum_{k=P+1}^{\infty} (u_k^X)^2, \\ (e_Z^N)^2 &= \|u(Z) - u_Z^{P,N}(Z)\|_{L^2}^2 = \underbrace{\sum_{k=0}^P (u_k^Z - u_k^{Z,N})^2}_{(e_{\text{int},P}^{Z,N})^2} + \sum_{k=P+1}^{\infty} (u_k^Z)^2. \end{aligned} \quad (6.12)$$

In [242], we obtained the finite integration accuracy counterpart of inequality (6.5) which can be stated as follow: with the previous notations, in the context of finite numerical integration accuracy, we have

$$(e_Z^N)^2 - (e_X^N)^2 = (e_{\text{int},P}^{Z,N})^2 - (e_{\text{int},P}^{X,N})^2 - \sum_{k=2}^P (u_k^Z)^2. \quad (6.13)$$

The proof is in [242]. According to the above inequality, the accuracy of the approximation can increase if we do not have

$$(e_{\text{int},P}^{Z^N,N})^2 - (e_{\text{int},P}^{X,N})^2 - \sum_{k=2}^P (u_k^{Z^N})^2 \leq 0. \quad (6.14)$$

The above condition is not always ensured: the integration error $(e_{\text{int},P}^{Z,N})^2$ can become preponderant in comparison to $-(e_{\text{int},P}^{X,N})^2 - \sum_{k=2}^P (u_k^{Z^N})^2$. This is the case when numerical integration accuracy is greater than the residue in the new basis. Consequently, if we do not want the approximation error to potentially increase after some iteration, we have to stop it before the projection error becomes preponderant in (6.14). In order to understand what happens more precisely, we need to go back to the definition of the gPC coefficients in their respective basis.

As in section 3.4, we introduce the truncated family of statistical moments $(s_k)_{k \in \{0, \dots, 2P\}}$ and truncated family of polynomials $\phi_k^X(X)$ orthonormal with respect to $d\mathcal{P}_X$. We recall the latter orthonormal polynomial family is directly linked with the moments of X [7, 156] as $\forall n \in \{0, \dots, P\}$:

$$\phi_n^X(x) = \frac{1}{\sqrt{H_{2(n-1)} H_{2n}}} \begin{vmatrix} s_0 & s_1 & \dots & s_n \\ \dots & \dots & \dots & \dots \\ s_k & s_{k+1} & \dots & s_{n+k} \\ \dots & \dots & \dots & \dots \\ 1 & x^1 & \dots & x^n \end{vmatrix}. \quad (6.15)$$

Once again suppose the moments $(s_k)_{k \in \{0, \dots, 2P\}}$ are not accurately computed, suppose one has

$$(s_k^\varepsilon)_{k \in \{0, \dots, 2P\}} = (s_k + \varepsilon_k)_{k \in \{0, \dots, 2P\}} \approx (s_k)_{k \in \{0, \dots, 2P\}}.$$

Assume a particular form for the perturbation $\varepsilon = (\varepsilon_0, \dots, \varepsilon_{2P}) = (0, \dots, 0, \delta)$ of the moments. This implies every moments of order $n \in \{0, \dots, 2P - 1\}$ are accurately computed whereas the last one s_{2P} is perturbed by δ such that $(s_k^\varepsilon)_{k \in \{0, \dots, 2P\}}$ is still in the moment space. One can compute the polynomials orthonormal with respect to the perturbed moments: they coincide with the one of X up to order $P - 1$ and we have

$$\phi_P^\varepsilon(x) = \frac{1}{\sqrt{\underline{H}_{2(P-1)} \underline{H}_{2P}^\varepsilon}} \begin{vmatrix} s_0 & s_1 & \dots & s_n \\ \dots & \dots & \dots & \dots \\ s_k & s_{k+1} & \dots & s_{n+k} \\ \dots & \dots & \dots & \dots \\ 1 & x^1 & \dots & x^n \end{vmatrix}. \quad (6.16)$$

The perturbation ε only perturbs the last Hankel determinant $\underline{H}_{2P}^\varepsilon$. Now, from the definition of $\underline{H}_{2P}^\varepsilon$ and by a development of the last line of the determinant, we have

$$\underline{H}_{2P}^\varepsilon = \underline{H}_{2P} + \delta \underline{H}_{2(P-1)} \quad \text{so that} \quad \phi_P^\varepsilon(x) = \frac{1}{\sqrt{1 + \delta \frac{\underline{H}_{2(P-1)}}{\underline{H}_{2P}}}} \phi_P^X(x). \quad (6.17)$$

Consequently, when considering the gPC coefficients of the transformation $u(X)$ in the perturbed basis, one has that $\forall n \in \{0, \dots, P - 1\}, u_n^\varepsilon = u_n^X$ and

$$u_P^\varepsilon = u_P^X \frac{1}{\sqrt{1 + \delta \frac{\underline{H}_{2(P-1)}}{\underline{H}_{2P}}}} \approx u_P^X - \frac{1}{2} \frac{\underline{H}_{2(P-1)}}{\underline{H}_{2P}} \delta u_P^X + \mathcal{O}(\delta^2), \quad (6.18)$$

for the last coefficient. It is known in the litterature [14, 118, 7, 156] that in the previous conditions $\forall n \in \{0, \dots, P\}$

$$\frac{\underline{H}_{2n}}{\underline{H}_{2(n-1)}} \leq 2^{-(4n+2)} \quad \text{leading to} \quad \left| \frac{du_P^X}{d\delta} \right| \geq 2^{4P+1},$$

which testifies for a higher and higher sensivity of the gPC coefficient with respect to a small inaccuracy in the statistical moments (δ) as the polynomial order P increases. In the next section, we present how we control *a posteriori* the integration error $(e_{\text{int},P}^{Z,N})^2$ (which is the only positive term in (6.14)). The process is inspired by the idea of the numerical ‘‘admissibility’’ of the moment data of [14, 118].

6.3.2 Strategy for adaptive approximation truncation

If the existence of $(\phi_k^Z)_{k \in \mathbb{N}}$ is straightforward (see [273, 117]), the existence of $(\phi_k^{Z^N})_{k \in \{0, \dots, P\}}$, the basis obtained with finite accuracy on the coefficients $(u_k^{X,N})_{k \in \{0, \dots, P\}}$ clearly depends on the quality of the approximation of the truncated statistical moments $(s_i^Z)_{i \in \{0, \dots, 2P+1\}}$ defined as:

$$s_i^Z = \mathbb{E}[Z^i] = \int x^i d\mathcal{P}^Z(x) \approx s_i^{Z^N} = \sum_{l=1}^N w_l \left(u_X^{P,N}(X_l) \right)^i, \forall i \in \mathbb{N}. \quad (6.19)$$

They are central in the construction of the gPC basis orthonormal with respect to $d\mathcal{P}_{Z^N}$. This problem is very sensitive and if the statistical moments $(s_i^{Z^N})_{i \in \mathbb{N}}$ are not accurately enough calculated, one can encounter the problem of *realisability* of the moment problem, see [198, 63, 155, 129, 172, 14, 7, 156]. Once the truncated statistical moments $(s_i^{Z^N})_{i \in \{0, \dots, 2P+1\}}$ built, the $2P+2$ -sized vector may or may not represent the first moments of a random variable due to a discrepancy in their calculations. If it is, this implies the existence of a basis $(\phi_k^{Z^N})_{k \in \{0, \dots, P\}}$ associated to the constraints $(s_i^{Z^N})_{i \in \{0, \dots, 2P+1\}}$ [14, 117].

Let us here assume the distribution has a bounded³ support $[a, b]$ and consider the related truncated moment problem (Hausdorff). Given that any interval may be linearly mapped, we now consider the normalized truncated moments $(\bar{s}_i^{Z^N})_{i \in \{0, \dots, 2P+1\}}$ in $[0, 1]$. Their existence is characterised in term of the Hankel determinants (Hd)

$$\begin{aligned} \underline{H}_{2P} &= \begin{vmatrix} \bar{s}_0^{Z^N} & \cdots & \bar{s}_P^{Z^N} \\ \vdots & & \vdots \\ \bar{s}_P^{Z^N} & \cdots & \bar{s}_{2P}^{Z^N} \end{vmatrix}, & \underline{H}_{2P+1} &= \begin{vmatrix} \bar{s}_1^{Z^N} & \cdots & \bar{s}_{P+1}^{Z^N} \\ \vdots & & \vdots \\ \bar{s}_{P+1}^{Z^N} & \cdots & \bar{s}_{2P+1}^{Z^N} \end{vmatrix}, \\ \overline{H}_{2P} &= \begin{vmatrix} \bar{s}_1^{Z^N} - \bar{s}_2^{Z^N} & \cdots & \bar{s}_P^{Z^N} - \bar{s}_{P+1}^{Z^N} \\ \vdots & & \vdots \\ \bar{s}_P^{Z^N} - \bar{s}_{P+1}^{Z^N} & \cdots & \bar{s}_{2P-1}^{Z^N} - \bar{s}_{2P}^{Z^N} \end{vmatrix}, & \overline{H}_{2P+1} &= \begin{vmatrix} \bar{s}_0^{Z^N} - \bar{s}_1^{Z^N} & \cdots & \bar{s}_P^{Z^N} - \bar{s}_{P+1}^{Z^N} \\ \vdots & & \vdots \\ \bar{s}_P^{Z^N} - \bar{s}_{P+1}^{Z^N} & \cdots & \bar{s}_{2P}^{Z^N} - \bar{s}_{2P+1}^{Z^N} \end{vmatrix}, \end{aligned} \quad (6.20)$$

via the following theorem [7, 156]:

Theorem 6.1 [7, 156]

- The truncated moment problem is well posed iff:
 $\underline{H}_n \geq 0$ and $\overline{H}_n \geq 0$, $\forall n \in \{0, \dots, 2P+1\}$.
- If $\underline{H}_n > 0$ and $\overline{H}_n > 0$ $\forall n \in \{1, \dots, 2P+1\} \rightarrow$ the polynomial basis orthonormal with respect to the pdf having the first moments $(s_i^{Z^N})_{i \in \{0, \dots, 2P+1\}}$ exists and is dense in $L^2(\Omega, \mathcal{A}, \mathcal{P})$.
- Finally, if $\exists k \in \{1, \dots, 2P+1\}$ such that $\underline{H}_l > 0$ and $\overline{H}_l > 0 \forall l \in \{1, \dots, k-1\}$ and such that $\underline{H}_k = 0$ or $\overline{H}_k = 0$, then there exists a polynomial basis orthonormal with respect to the $(\bar{s}_i^{Z^N})_{i \in \{0, \dots, k\}}$, this basis is not dense in $L^2(\Omega, \mathcal{A}, \mathcal{P})$.

Similarly to the work introduced in [14], we construct the stopping criterion of our algorithm based on theorem 6.1. For each i-gPC iteration, we compute the Hankel determinants to check wellposedness (the ε threshold refers to machine accuracy) :

- [a.] If the first point of theorem 6.1 is not numerically satisfied , i.e. $\exists n_0 \in \{0, \dots, 2P+1\}$ such that $\underline{H}_n < -\varepsilon$ or $\overline{H}_n < -\varepsilon \rightarrow$ we substitute $P \leftarrow \lfloor \frac{n_0}{2} \rfloor$ and test if a smaller truncation order is realizable.
- [b.] If it is and in particular even the second point, i.e. $\forall n \in \{0, \dots, 2P+1\}$, $\underline{H}_n > \varepsilon$ and $\overline{H}_n > \varepsilon \rightarrow$ there exists a dense polynomial basis \rightarrow we perform another iteration.
- [c.] If the third condition of theorem 6.1 is satisfied, i.e. $\exists k \in \{1, \dots, 2P+1\}$ such that $\underline{H}_l > \varepsilon$ and $\overline{H}_l > \varepsilon \forall l \in \{1, \dots, k-1\}$ and such that $|\underline{H}_k| \leq \varepsilon$ or $|\overline{H}_k| \leq \varepsilon \rightarrow$ there exists a polynomial basis (not dense) depending only on the k first moments \rightarrow we truncate the moment problem to the k^{th} term and modify the order of the i-gPC representation for this step \rightarrow we stop the iterations.

Once again, even in a finite integration accuracy context, two situations may occur:

- The moments are realizable, the numerical integration accuracy is “good enough”, i.e. (6.14) is fulfilled, and the iterative procedure converges (stagnates) according to the inequality (equality) (6.7); equality may be reached after a certain step.
- The moments for the considered order P are not realizable, the numerical integration accuracy is too low, i.e. $\exists k_0$, such that $(e_{\text{int}, P}^{Z^{k_0}, N})^2$ is important with respect to the negative terms in (6.14). In order to reduce the integration error in the new basis, we lower the polynomial order $P \leftarrow \lfloor \frac{n_0}{2} \rfloor$ in the new basis at iteration k_0 . Practically, this corresponds to making sure the integration error with $\lfloor \frac{n_0}{2} \rfloor$ is lower than with P : basically we make sure $e_{\text{int}, \lfloor \frac{n_0}{2} \rfloor}^{Z^{n_0}, N} < e_{\text{int}, P}^{Z^{n_0}, N}$ so that the positive term in (6.13) becomes less important.

³Note that the generalization to the truncated Hamburger/Stieljes moment problem is not straightforward. In this paper, we restrict ourselves to transformation of arbitrary random variables to random variables with bounded support. This restriction still allows dealing with many physical applications.

Let us briefly apply the above material to a discontinuous test-problem slightly different for the one of section 6.2 in a finite integration accuracy context. More benchmarks are considered in [242] together with interesting features of the new iterative algorithm. Let X be a uniform random variable on $[-1, 1]$. Let u be a step function equal to 1 on $[-1, -0.4]$ and 0 on $[-0.4, 1]$. The function is sampled only once at N points. Our choice here is to use a numerical quadrature with 14 Gauss-Legendre (GL) points. We apply non-intrusive gPC (with truncation degree $P = 3$ and $P = 5$) and i-gPC (with truncation degree $P = 3$ and $P = 5$) together with a spectral collocation approximation (for collocation, $N = 14 \Rightarrow P = 14$). Figure 6.8 shows the latter approximations together with the analytical solution. The gPC and collocation polynomial approximations are oscillating whereas the i-gPC approximations do not and avoid the Gibbs' phenomenon, see [130]. Table 6.2 (left) presents the L^1 and L^2 norm of

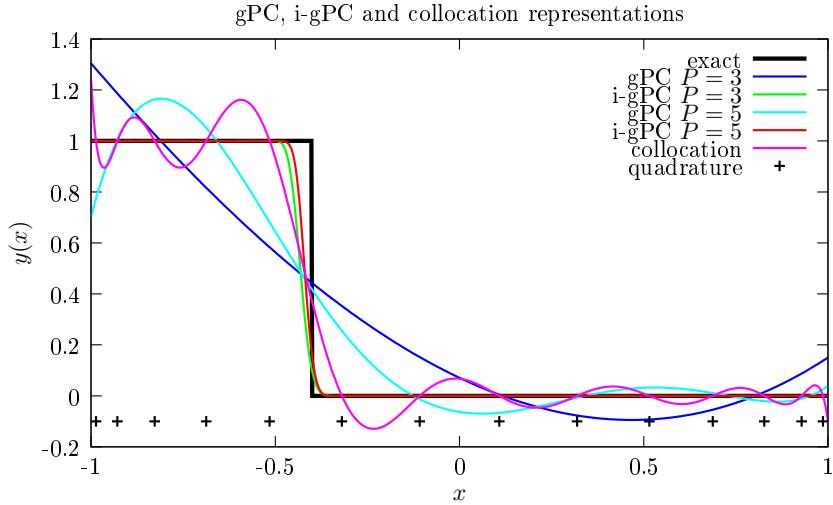


Figure 6.8: gPC ($P = 3$ and $P = 5$), i-gPC ($P = 3$ and $P = 5$) and collocation ($P = 14$) comparison

the errors of these approximations. The table shows that i-gPC does better than gPC and collocation both in L^1 and L^2 norms. In practice, we notice that numerical integration accuracy is reached by the

	L^1 -error	L^2 -error	CC Quad. level	gPC ($P = 5$)	i-gPC ($P = 5$)
gPC $P = 3$	0.15	0.20	$l = 5$	0.095623	0.0075440
i-gPC $P = 3$	0.016	0.099	$l = 6$	0.094358	0.0070233
gPC $P = 5$	0.097	0.16	$l = 7$	0.094625	0.0030701
i-gPC $P = 5$	0.011	0.077	$l = 8$	0.094402	0.0018716
collocation	0.067	0.11	$l = 9$	0.094490	1.3055e-04

Table 6.2: Left: gPC, i-gPC and collocation errors comparison. Right: Convergence with respect to the number of Clenshaw-Curtis (CC) quadrature level for gPC and i-gPC representations for fixed P : numerical integration accuracy is reached.

approximation in the sense that it is here the limiting factor. This is emphasized in the numerical study summed up in table 6.2 (right). The results are the L^1 -norm of the error on the same test problem with $P = 5$ and for different level l of Clenshaw-Curtis (CC) quadrature rule ($l = 5, l = 6, l = 7, l = 8$ and $l = 9$ ⁴). The accuracy of the gPC approximation is only slightly affected by the increase of level of the quadrature rule. Consequently, the truncation error is in this case predominant compared to the integration error. The behaviour of i-gPC is different: as the number of quadrature points increases, the approximation gains in quality. It shows that it is driven by the numerical integration accuracy, and not the truncation order P .

⁴The level l of the CC rule ensures the quadrature has $k = 2^l - 1$ points nested by levels.

6.4 Few other applications of i-gPC

In this last section, we first come back to the 'fil rouge' problem of chapter 2 before comparing gPC, regression, regression-gPC, collocation-gPC, kriging-gPC and i-gPC on Runge function and on a discontinuous solution.

6.4.1 Application to the 'fil rouge' configuration

Now the iterative process i-gPC fully described, we go back to our 'fil rouge' configuration with our new numerical approximation scheme i-gPC. Note that we do not anymore systematically monitor the number of iterations as it is automated *via* the stopping criterion described in the previous sections. The new algorithm i-gPC being only a postprocess, we can reuse the same $N = 15$ runs of the Euler code with the previous Gauss-Legendre points with i-gPC in the same conditions ($P = 3$). In term of mean and variance of the mass density profiles, the results with i-gPC are very close to the ones obtained with gPC. In fact, they are exactly the same as mean and variance mainly depend on the integration

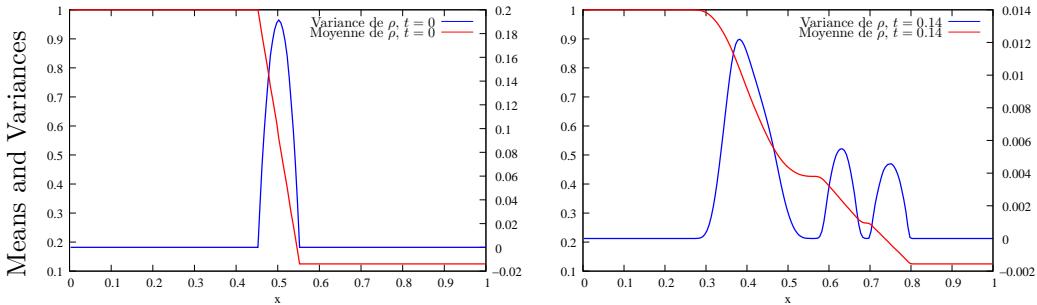


Figure 6.9: Application of non-intrusive i-gPC, mass density spatial profiles for some realisations, the mean and the variance at $t = 0$ and $t = 0.14$.

scheme, hence the experimental design which, here, is common to both approaches. The results are displayed in figure 6.9 even if we do not spend more time on them and focus on figure 6.10 presenting the histograms at the particular locations and time of interest obtained with i-gPC. First, the new iterative process allows recovering the same results as gPC in the vicinity of the rarefaction fan, see figure 6.10 (top). The algorithm has been built in order not to degrade the accuracy of the gPC approximation. For both other waves, in the vicinities of the interface and of the shock, figure 6.10 (bottom pictures), the gain is qualitatively very important. For the histograms, the discrete behaviour is quite well captured with two Dirac masses at the correct location and a quite good estimation of their masses. On the functional representations, i-gPC allows recovering the steep gradients of the solutions with respect to the random variable. We present some other quantitative results on the 'fil rouge' configuration in table 6.3. It displays the L^1 and L^2 norms for the random variable 'mass density' in the vicinities of the rarefaction fan, the interface and the shock. As tackled before, the chosen spatial discretisation for every $N_{MC} = 1000$ runs of the reference MC solutions and the $N = 15$ runs of the i-gPC approximation was such that one can rely on an accuracy of about $10^{-4} \approx \Delta x$. Based on this, we notice on the 'rarefaction' lines of table 6.3 that the accuracy in the vicinity of the rarefaction fan is about 10^{-4} for both gPC and i-gPC. For this wave, the stochastic counterpart of the solver do not limit the accuracy of the results as they are close to the spatial accuracy ($\Delta x = 10^{-4}$). On another hand, if we consider the norms of the errors for gPC in the vicinities of the shock and the interface, the accuracy drops to 10^{-2} and is two decades greater than the spatial one. For these waves, in a sense, computational resources are lost as the stochastic solver does not allow recovering the accuracy of the *a priori* chosen discretisation for the black-box runs. Regarding i-gPC in the same conditions for the interface and the shock, we recover the accuracy of the deterministic solver for the Euler equations at the N points as the L^1 and L^2 norms of the errors are about 10^{-4} . In this context, i-gPC allows an important gain with respect to gPC as testifies the last column of table 6.3 with a factor $\times 250$ of gain in the vicinity of the shock and $\times 30$ in the vicinity of the interface. More than a gain, it allows avoiding a waste of computational resources allocated for the N runs of accuracy $\Delta x = 10^{-4}$. We tried to obtain such accuracy with gPC, i.e. about 10^{-4} for the random variable 'mass density' in the vicinities of the interface and the shock but we were

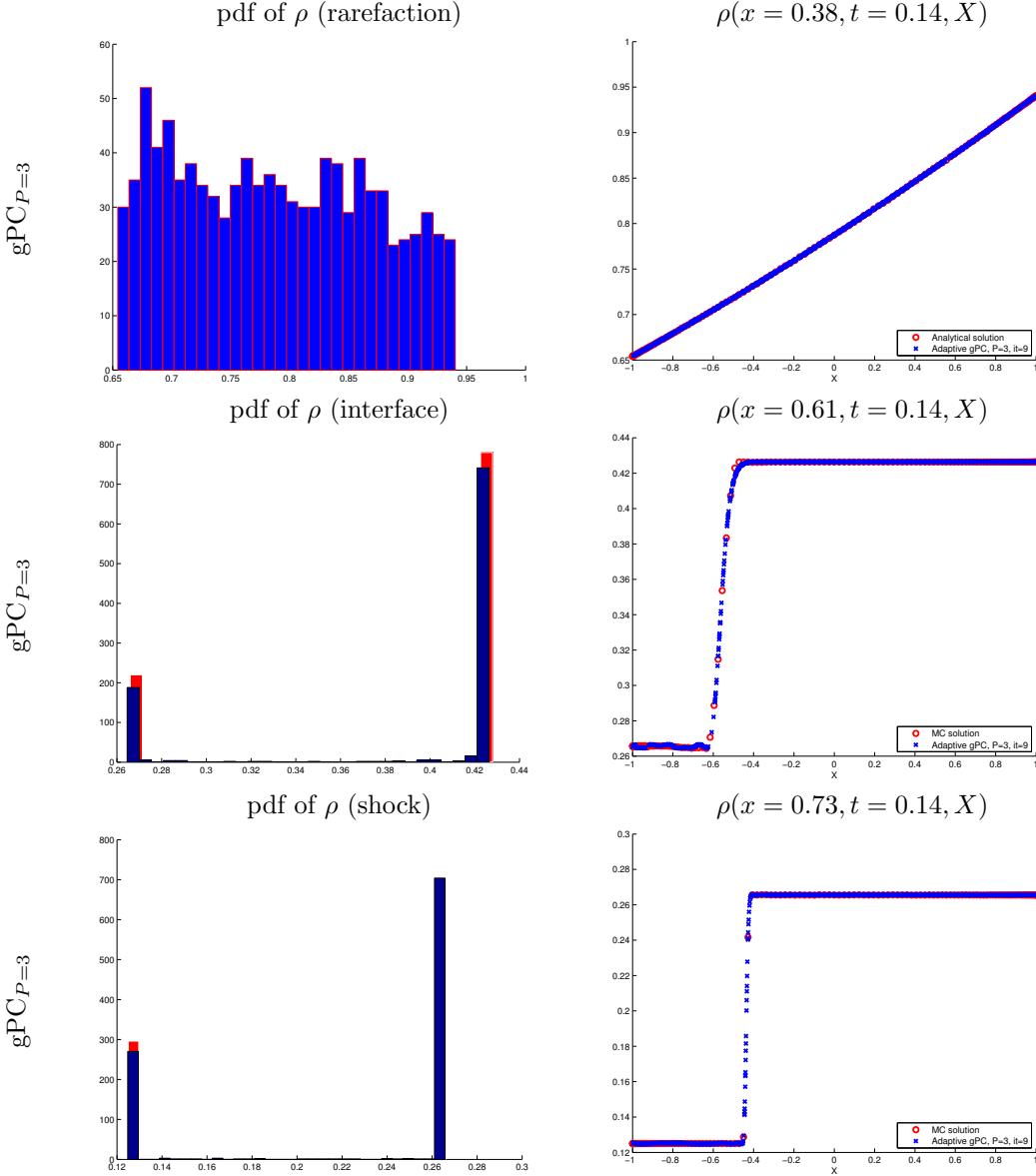


Figure 6.10: Pdfs (left) and functional representation in the random space of the mass density at $t = 0.14$ in the vicinities of the rarefaction fan ($x = 0.38$), the interface ($x = 0.61$) and the shock ($x = 0.73$) with the i-gPC $P=3$ approximation.

not able to reach it with $P = 27$ and more than thousand runs of Gauss-Legendre quadrature points. We suspect important aliasing errors during the construction of the gPC basis for such high orders, see section 3.4.

6.4.2 Integration vs. Regression vs. Collocation vs. Kriging vs. i-gPC

The natural question arising now is how does i-gPC perform in comparison with the different approximations of section 5.3? In this section, we apply integration-gPC, regression, regression-gPC, collocation-gPC, kriging-gPC and i-gPC to the same functions studied in sections 5.3 and 5.4.2 (discontinuity and Runge function).

Let us begin with function $X \rightarrow \mathbf{1}_{[-\infty, \frac{3}{10}]}(X)$ with $X \sim \mathcal{U}([-1, 1])$, intensively studied in section 5.4.2. Figure 6.11 is in the same vein as figure 5.12: we only added the curves obtained with i-gPC.

	$gPC_{P=3}$, L^1 -norm	$i-gPC_{P=3}$, L^1 -norm	$\frac{\text{error}_{gPC}}{\text{error}_{i-gPC}}$
rarefaction	1.817e-04	1.817e-04	1
interface	1.633e-02	3.861e-04	42.29
shock	2.053e-02	7.863e-05	261.09
	$gPC_{P=3}$, L^2 -norm	$i-gPC_{P=3}$, L^2 -norm	$\frac{\text{error}_{gPC}}{\text{error}_{i-gPC}}$
rarefaction	2.196e-04	2.196e-04	1
interface	2.523e-02	8.406e-04	30.01
shock	2.765e-02	1.133e-04	244.04

Table 6.3: Quantitative comparison of $gPC_{P=3}^{N=15}$ and $i-gPC_{P=3}^{N=15}$ with respect to an MC reference solution with $N_{MC} = 1000$ points in term of approximations of the random variable 'mass density' in the vicinities of the rarefaction fan, the interface and the shock.

Let us begin by commenting on the convergence study with respect to P of figure 6.11 (top-right): the general behaviour of i-gPC is comparable to the ones of the other approximation methods. The error first decreases than explodes as soon as $P > N$. With such low number of points of the experimental design, it is complex controlling term (6.14) even with the stopping criterion described above: for some polynomial orders, the i-gPC error is slightly more important than the gPC one, which should not happen with a good numerical integration. This is the case for $P = 4, 5, 6$ for example. For those polynomial orders, the stopping criterion should have stopped the algorithm one iteration earlier. Still, the i-gPC approximations remain controlled and of comparable qualities as the ones obtained with the other methods. Figure 6.11 (top-left) presents the best approximations obtained with every methods: the curves are the same as figure 5.12 described in section 5.4.2 except we added the i-gPC one. The best i-gPC approximation is obtained with $P = 3$. It is much less oscillatory than the other methods and the discontinuous behaviour of the solution is already captured.

The bottom pictures of figure 6.11 (bottom) present the same studies with $N = 21$ GL points. With such an accurate numerical integration, i-gPC gives the best results as soon as $P > 2$. The convergence curve for i-gPC is always below the other ones. The best i-gPC approximation is displayed figure 6.11 (bottom-left) together with the best ones of the other methods: it is obtained for $P = 3$, it is the less oscillatory of every approximations and it captures the discontinuous behaviour of the solution.

To complete the comparison of i-gPC with the other methods of the litterature, we perform the same study to Runge function. First, once again, for a small number of points of the experimental design, the behaviour of i-gPC is similar to the ones of the other methods. A decrease of the error before an explosion: the smoothness of Runge function does not improve this point. Furthermore, as before, the small number of points makes the control of the integration error in (6.14) difficult and i-gPC does not always perform better than gPC: if P is kept lower than N in another hand (see $P = 2, 3$ in figure 6.12 (top-right)), the results obtained with i-gPC are intermediary to the ones obtained with gPC, regression, regression-gPC (upper bound) and kriging-gPC (lower bound). Qualitatively, see figure 6.12 (top-left) the i-gPC approximation gives equivalent results as the other methods. Now, if $N = 21$, the integration is accurate and i-gPC gives better results than gPC, regression, regression-gPC, up to $P = N = 20$. Kriging-gPC remains more efficient on such smooth output function even if qualitatively, every best approximations are indistinguishable (see figure 6.12 bottom-left).

6.5 Summary for non-intrusive gPC and i-gPC approximations

In this chapter, we studied a new non-intrusive application of gPC for uncertainty quantification. It intensively uses the possibility to apply gPC non-intrusively together with the exploitation of the existence of more or less adapted basis. The independence of the points makes them *embarrassingly parallel* and quickly efficient when one has at hand a robust simulation code and access to a computing cluster. With inequalities (6.3)–(6.5) we put forward the possibility to increase, during a postprocessing step, the quality of the approximation basis without *prior* assumption on the regularity of output random

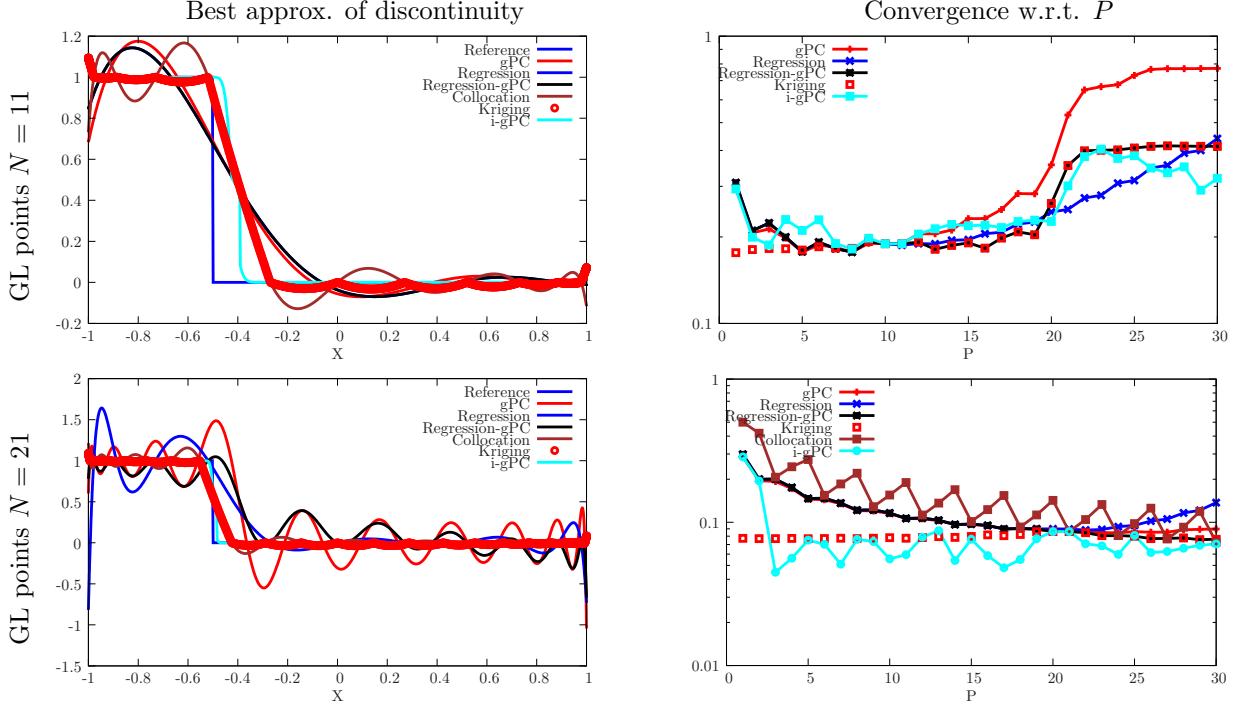


Figure 6.11: This figure is the same as figure 5.12 but we added the curves obtained with i-gPC. Application of Gauss-Legendre quadrature rule for gPC, regression, regression-gPC, collocation-gPC, kriging-gPC and i-gPC for the approximation of the transformation of a uniform random variable through a discontinuous function. The experimental designs have with $N = 11$ (top) and $N = 21$ (bottom). The left column present the best approximations obtained with every of the previous methods. The right column present the L^2 -norm of the error with respect to P for fixed N . The kriging kernel is chosen exponential (5.42) and a dichotomy is applied to calibrate θ .

variable. We even suggested an algorithm based on (6.5) to build this new basis from the initial one, associated to X . In this sense, from a practical point of view, the non-intrusive counterparts (MC, gPC or i-gPC) only differ from the number of runs N they need and the cost of the postprocessing step at the end of every runs of the black-box code. If we now focus on the i-gPC algorithm presented in section 6.1.2, which corresponds to our contribution in term of non-intrusive resolution scheme, we designed a new iterative method based on gPC and moment theory allowing

- improvements only thanks to a post-processing step,
- important gains on discontinuous solutions, smaller ones on continuous ones,
- recovering the optimal basis with respect to the output in certain cases (strong contraction).
- In the case of a *weak contraction* we worked on another iterative algorithm presented in chapter 8. It uses the fact that the initial gPC basis and the final one (i-gPC at the last iteration) are not necessarily orthonormal and aims at improving the accuracy of the approximation by working on the residue in the new basis. It, in a sense, consists on a reinterpretation of Cameron-Martin's theorem.

Every results for i-gPC have been presented in 1D stochastic dimension but its application to higher ones is straightforward and some cases are presented in [238, 242]. Note that the iterative approach has not been designed to deal with more stochastic dimensions than gPC, only to improve its accuracy.

We finally insist other authors noticed the importance of increasing the accuracy of the approximation basis prior to introducing any other numerical ingredients (such as the one introduced in kriging-gPC, see

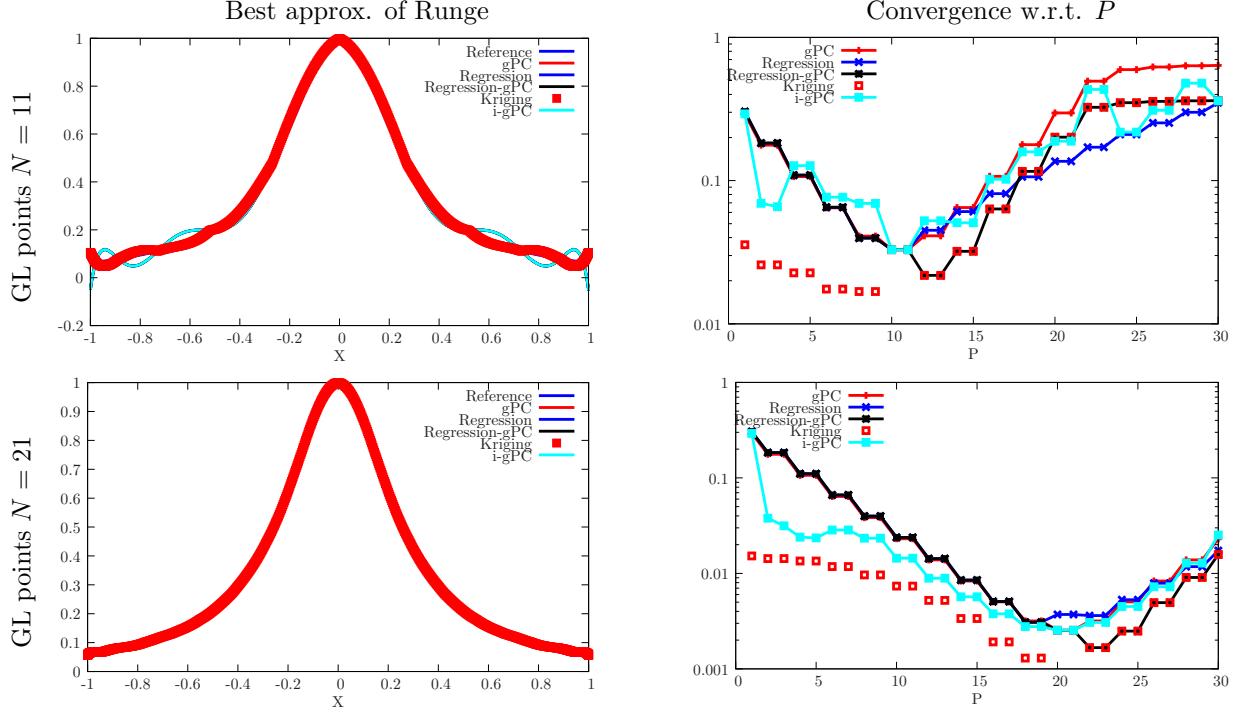


Figure 6.12: This figure is in the same vein as figure 6.11 except we consider Runge function instead of a discontinuous one. Application of Gauss-Legendre quadrature rule for *g*PC, *regression*, *regression-g*PC, *collocation-g*PC, *kriging-g*PC and *i-g*PC for the approximation of the transformation of a uniform random variable through Runge function. The experimental designs have with $N = 11$ (top) and $N = 21$ (bottom). The left column present the best approximations obtained with every of the previous methods. The right column present the L^2 -norm of the error with respect to P for fixed N . The kriging kernel is chosen exponential (5.42) and a dichotomy is applied to calibrate θ .

section 5.3.3). In [82] for example, the author builds a new family of polynomials ensuring the respect of bounds on the approximation of the output (maximum principle). Such new approximation basis will probably allow efficient resolutions in an uncertainty quantification context but, more generally, for any numerical methods implying polynomials (P_n, M_n models, high-order schemes for the deterministic resolution of hyperbolic system, reconstruction of positive pdfs etc.).

Chapter 7

Non intrusive gPC for Direct Numerical Simulation (DNS) acceleration

A well-known physical problem revisited as an uncertainty quantification one

An attempt to exploit the ergodicity of the gPC approximation (see section 3.1.2)

Contents

7.1	Perturbation reduced models as a limit of the gPC one	123
7.1.1	Perturbation reduced model of a system of conservation laws	123
7.1.2	gPC reduced model of a system of conservation laws	125
7.1.3	The perturbative reduced model as a limit of the gPC one	125
7.2	Direct Numerical Simulation (DNS) acceleration via gPC	128
7.2.1	The shock tube experiments and their initial conditions	129
7.2.2	Stochastic dimension reduction for the initial Uncertain Interface Position	131
7.2.3	The Multimaterial 2D Euler system	132
7.2.4	Observable of interest, Simulations and Comparisons with Experimental Results	133
7.3	Conclusion for the gPC application to chaotic flows	137

The study of linearly perturbed flows has been given a lot of interest in computational fluid dynamics (CFD) over the last decades (see [151, 149, 150, 68, 206, 274, 276, 275, 39] for example). Especially, as we are interested in this chapter, in order to predict the growth rate of instabilities with respect to time in very sensitive flows. On another hand, perturbation methods are also well-known in uncertainty quantification (UQ) – see [271] for example– and have been recently generalized in this field (see [218] for example). Spectral methods such as generalized Polynomial Chaos (gPC) have proven to be efficient for solving Stochastic Partial Differential Equations (SPDE) and contain perturbation methods – in the sense solutions obtained with perturbation methods can be recovered with gPC. Section 7.1 of this chapter aims at recalling these results (already hinted at in [231, 232, 243]): perturbation methods are presented as a limit, a reduced model, of gPC methods in a regime which will be identified. In section 7.2, just as gPC generalizes perturbation methods in uncertainty quantification, we aim at generalizing the study of complex flows thanks to gPC: we apply gPC in order to predict the growth rate of instabilities in *nonlinear* regimes in the same condition perturbation methods are commonly applied. We directly follow the methodology described in [237] and deepen the study by reinterpreting shock tube experiments from [244, 286] as uncertainty driven experiments. They corresponds to Richtmyer-Meshkov (RM) ones and

are recalled in section 7.2.

The original work presented in this chapter is published in [243] in a more concise form. We here give more details regarding perturbative methods (section 7.1) and the resolution of the stochastic inverse problem built in section 7.2.

7.1 Perturbation reduced models as a limit of the gPC one

In this section, we compare perturbative methods to gPC ones. The object is to formally show that perturbative reduced models from a conservation law of interest are the limit of the gPC reduced model of these same conservation law under certain conditions on the input uncertain parameters. The results are demonstrated for any system of conservation laws in $1D^1$ even if we will focus on the Euler system later on. We recall the general form for uncertain system of conservation laws is given by

$$\partial_t u + \partial_x f(u) = 0, \text{ with } u(x, t, X) \in \mathcal{D}_u \subset \mathbb{R}^d. \quad (7.1)$$

Random variable X has probability measure $d\mathcal{P}_X$ and models the uncertainty in, for example here, initial conditions.

7.1.1 Perturbation reduced model of a system of conservation laws

In this first section, we recall the general form of the perturbative reduced model of a system of conservation laws. Perturbative methods consists in considering small perturbations of the solution of a system of conservation laws around its mean value. The vector of unknowns u , solution of the system (7.1) is approximated by its Taylor development with respect to the random variable X , truncated at order P :

$$u(X) \approx \bar{u}^P(X) = \bar{u}_0 + \sum_{i=1}^P \frac{\bar{u}_i X^i}{i!}. \quad (7.2)$$

The validity of the perturbative approach directly depends on the characteristics of the random variable of interest X . Typically, if the probability measure of X ensures small fluctuations around μ the mean of X . Let us focus on the set of equations obtained from reducing the system of conservation laws of interest into a perturbed one. For this, we look at the system of equations satisfied by $\bar{U} = (\bar{u}_0, \dots, \bar{u}_P)^t$, in which formally we have $\bar{u}_k = \frac{\partial^k u}{\partial X^k}(x, t, \mu), \forall k \in \{0, \dots, P\}$. The perturbative reduced model of (7.1) is obtained plugging the Taylor development (7.2) in system (7.1) and identifying the multiplicators of the components of $(1, X, X^2, \dots, X^P)$. The steps are exactly the same as the one for the Hilbert expansion of chapter 1 to identify physical regimes of interest. Let us focus on the expression of the k^{th} equation of the obtained reduced model and more particularly on the expression of its flux:

$$\begin{aligned} f(\bar{u}^P(X)) &= f\left(\bar{u}_0 + \sum_{i=1}^P \frac{\bar{u}_i X^i}{i!}\right), \\ f(\bar{u}^P(X)) &= f(\bar{u}_0) + \sum_{j=1}^P \frac{1}{j!} f^{(j)}(\bar{u}_0) \left(\sum_{i=1}^P \frac{\bar{u}_i X^i}{i!}\right)^j. \end{aligned} \quad (7.3)$$

¹The results in higher dimensions are straightforward but only make the computations heavier.

Using Newton's Multinomial formulae², the flux expression becomes

$$\begin{aligned}
f(\bar{u}^P(X)) &= f(\bar{u}_0) + \sum_{j=1}^P \sum_{|\vec{k}|=j} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i X^i}{i!} \right)^{k_i}, \\
f(\bar{u}^P(X)) &= f(\bar{u}_0) + \sum_{j=1}^P \sum_{|\vec{k}|=j} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i} X^{Pj-(P-1)k_1-(P-2)k_2-\dots-k_{P-1}}, \\
f(\bar{u}^P(X)) &= f(\bar{u}_0) + \sum_{t=1}^{P^2} X^t \underbrace{\sum_{j=1}^P \sum_{|\vec{k}|=j}}_{j, \vec{k}/t = Pj - (P-1)k_1 - (P-2)k_2 - \dots - k_{P-1}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}.
\end{aligned} \tag{7.5}$$

After identification of the coefficients of $(X^n)_{n \in \{0, \dots, P\}}$, we deduce the components of the flux of the truncated reduced model obtained by perturbative methods $\forall n \in \{0, \dots, P\}$ are given by

$$f_n(\bar{u}_0, \dots, \bar{u}_P) = \underbrace{\sum_{j=1}^P \sum_{|\vec{k}|=j}}_{j, \vec{k}/n = Pj - (P-1)k_1 - (P-2)k_2 - \dots - k_{P-1}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}. \tag{7.6}$$

The obtained reduced model reads

$$\left\{
\begin{aligned}
&\partial_t \bar{u}_0 + \partial_x f(\bar{u}_0) = 0, \\
&\dots \\
&\partial_t \frac{\bar{u}_n}{n!} + \partial_x \left(\underbrace{\sum_{j=1}^P \sum_{|\vec{k}|=j}}_{j, \vec{k}/n = Pj - (P-1)k_1 - (P-2)k_2 - \dots - k_{P-1}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i} \right) = 0, \\
&\dots \\
&\partial_t \frac{\bar{u}_P}{P!} + \partial_x \left(\underbrace{\sum_{j=1}^P \sum_{|\vec{k}|=j}}_{j, \vec{k}/P = Pj - (P-1)k_1 - (P-2)k_2 - \dots - k_{P-1}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i} \right) = 0.
\end{aligned} \right. \tag{7.7}$$

It is weakly coupled in the sense the first equation on \bar{u}_0 does not depend on the higher terms $(\bar{u}_k)_{k \in \{1, \dots, P\}}$, the second one only depends on \bar{u}_0, \bar{u}_1 etc. Such reduced models are known to be *weakly hyperbolic*. This is very easy to verify by diagonalizing the Jacobian of the flux on simple (scalar) conservations laws at order $P = 1$. One would find out that the basis of eigenvectors is not complete. Weakly hyperbolic systems give satisfactory results at early times but fail for long term simulations: the solutions become unphysical as nothing prevents them from going (linearly with time) to infinity, see [81].

In the following sections, we briefly recall the construction of the gPC reduced model (even if already tackled in chapter 4).

²Newton's Multinomial formulae:

$$(x_1 + \dots + x_P)^j = \sum_{|\vec{k}|=j} C_{\vec{k}}^j \prod_{i=1}^P x_i^{k_i}, \tag{7.4}$$

where $\vec{k} = (k_1, \dots, k_P)$ is the vector of powers of x_i , $|\vec{k}| = \sum_{i=1}^P k_i$ and $C_{\vec{k}}^j = \frac{n!}{\prod_{i=1}^P k_i!}$.

7.1.2 gPC reduced model of a system of conservation laws

The gPC reduced model of system (7.1) obtained from sG-gPC, see section 4.1, is given by

$$\left\{ \begin{array}{l} \partial_t u_0 + \partial_x \int f \left(\sum_{i=0}^P u_i \phi_i^\delta \right) \phi_0^\delta d\mathcal{P}_X = 0, \\ \dots \\ \partial_t u_k + \partial_x \int f \left(\sum_{i=0}^P u_i \phi_i^\delta \right) \phi_k^\delta d\mathcal{P}_X = 0, \\ \dots \\ \partial_t u_P + \partial_x \int f \left(\sum_{i=0}^P u_i \phi_i^\delta \right) \phi_P^\delta d\mathcal{P}_X = 0. \end{array} \right. \quad (7.8)$$

Note that we slightly changed our notations, especially concerning the gPC basis, denoted by $(\phi_i^\delta)_{i \in \{0..P\}}$, where a parameter δ now explicitly appears. This new parameter denotes the variance (up to a multiplicative factor) of the random variable X . For example, if X is a uniform random variable, δ is such that $X = \mu + \delta \mathcal{U}[-1, 1]$ so that $X - \mu \sim \mathcal{U}[-\delta, \delta]$. In the following lines, we conserve the uniformity assumption but we insist this is without loss of generality. System (7.8) is built with sG-gPC, i.e. with the P_n -like closure of section 4.1. It is known to fail regarding hyperbolicity/wellposedness in certain cases detailed in chapter 4. Rigorously speaking, the next computations should have been carried on with the reduced model obtained by entropy closure of section 4.2.3 ensuring wellposedness. We insist the results would be same at the price of even more complex calculations³. In the following developments, we suppose system (7.8) is hyperbolic.

7.1.3 The perturbative reduced model as a limit of the gPC one

We begin by stating the following property:

Property 7.1 *With the previous notations, the solutions $(\bar{u}_k)_{k \in \{0, \dots, P\}}$ of system (7.7) are limits of the solutions $(u_k^\delta)_{k \in \{0, \dots, P\}}$ of system (7.8) as $\delta \rightarrow 0$.*

In term of uncertainty quantification problem, this means perturbation methods are equivalent to gPC ones in the limit of a small variance of the input initial random variable X . From a modeling point of view, it means gPC models asymptotically preserve the perturbative regime characterised by $\delta \rightarrow 0$. Stated as above, the results may seem obvious. They are much less comparing expressions (7.7) and (7.8). The following proof allows understanding how one can switch from one reduced model to the other.

Proof Let us introduce the matrix

$$\mathbb{A} = \begin{pmatrix} a_{0,0}^\delta & 0 & 0 & \dots & 0 \\ a_{1,0}^\delta & a_{1,1}^\delta & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{P,0}^\delta & a_{P,1}^\delta & \dots & a_{P,P-1}^\delta & a_{P,P}^\delta \end{pmatrix},$$

matrix of the coefficients of the gPC basis, $(\phi_i^\delta)_{i \in \{0, \dots, P\}}$, in the canonical one $(X^i)_{i \in \{0, \dots, P\}}$:

$$\forall i \in \{0, \dots, P\}, \text{ we have } \phi_i^\delta(X) = \sum_{j=0}^i a_{i,j}^\delta X^j \text{ (triangular matrix).}$$

We denote by \mathbb{B} the inverse of \mathbb{A} ⁴, we then have $\forall (i, j) \in \{0, \dots, P\}^2$, $\sum_{k=0}^P a_{i,k} b_{k,j} = \delta_{i,j}$.

³We would only have to take into account one degree of nonlinearity in order to have $u = \nabla s(v)$ where v is developed on the gPC basis.

⁴This inverse \mathbb{B} exists as \mathbb{A} is matrix allowing to go from one basis to another.

If $(\phi_i^\delta(X))_{i \in \{0, \dots, P\}}$ is the gPC basis associated to X and $(\phi_i(X))_{i \in \{0, \dots, P\}}$ the basis associated to $\frac{X-\mu}{\delta}$ (the reduced centered random variable of X), then we have that $(\phi_i^\delta(X) = \phi_i(X\delta))_{i \in \{0, \dots, P\}}$. This is easy to understand for example if we consider uniform random variables and the Legendre basis orthonormal with respect to $d\mathcal{P}_X(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)dx$. In such conditions, $(\phi_i^\delta(X) = \phi_i(X\delta))_{i \in \{0, \dots, P\}}$ is the basis orthonormal with respect to X uniformly distributed on $[-\delta, \delta]$. Let us perform the change of variable in (7.8) noticing that, from equality

$$\bar{u}_0 + \sum_{i=1}^P \frac{\bar{u}_i X^i}{i!} = \sum_{i=0}^P u_i \phi_i^\delta(X), \quad (7.9)$$

it is possible to obtain

$$\forall i \in \{0, \dots, P\}, \quad \frac{\bar{u}_i}{i!} = \sum_{j=0}^P a_{j,i} u_j, \quad (7.10)$$

which can also be rewritten

$$\forall i \in \{0, \dots, P\}, \quad u_i = \sum_{j=0}^P b_{j,i} \frac{\bar{u}_j}{j!}. \quad (7.11)$$

Then, introducing the expressions of (7.11) into (7.8), we get $\forall n \in \{0, \dots, P\}$:

$$\begin{aligned} \partial_t \left(\sum_{j=0}^P b_{n,j} \frac{\bar{u}_j}{j!} \right) + \partial_x \int f \left(\sum_{i=0}^P \sum_{j=0}^P b_{j,i} \frac{\bar{u}_j}{j!} \sum_{l=0}^P a_{i,l} (X\delta)^l \right) \phi_n^\delta(X) d\mathcal{P}(X) &= 0, \\ \partial_t \left(\sum_{j=0}^P b_{n,j} \frac{\bar{u}_j}{j!} \right) + \partial_x \int f \left(\sum_{j=0}^P \sum_{l=0}^P \frac{\bar{u}_j (X\delta)^l}{j!} \underbrace{\sum_{i=0}^P b_{j,i} a_{i,l}}_{\delta_{j,l}} \right) \phi_n^\delta(X) d\mathcal{P}(X) &= 0, \\ \partial_t \left(\sum_{j=0}^P b_{n,j} \frac{\bar{u}_j}{j!} \right) + \partial_x \int f \left(\underbrace{\sum_{j=0}^P \frac{\bar{u}_j (X\delta)^j}{j!}}_{\bar{u}^0 + \mathcal{O}(d)} \right) \phi_n^\delta(X) d\mathcal{P}(X) &= 0. \end{aligned} \quad (7.12)$$

After a limited development at order P , assuming $\delta \rightarrow 0$ and the use of Newton's Multinomial formulae (7.4), we get $\forall n \in \{0, \dots, P\}$:

$$\begin{aligned} \partial_t \left(\sum_{j=0}^P b_{n,j} \frac{\bar{u}_j}{j!} \right) + \partial_x \int \left(f(\bar{u}_0) + \sum_{t=1}^{P^2} (X\delta)^t \underbrace{\sum_{j=1}^P \sum_{|\vec{k}|=j}_{j, \vec{k}/t = Pj - \dots - k_{P-1}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}} \right) \phi_n^\delta(X) d\mathcal{P}(X) &= 0. \end{aligned} \quad (7.13)$$

Expressing each monomials $(X^t)_{t \in \{0, \dots, P\}}$ in the gPC basis $(\phi_t)_{t \in \{0, \dots, P\}}$ writes

$$\forall t \in \{0, \dots, P\}, \quad (\delta X)^t = \sum_{l=0}^P b_{l,t} \phi_l^\delta(X).$$

Its use in expression (7.13) leads to $\forall n \in \{0, \dots, P\}$:

$$\partial_t \left(\sum_{j=0}^P b_{n,j} \frac{\bar{u}_j}{j!} \right) + \partial_x \left(f(\bar{u}_0) + \sum_{t=1}^{P^2} b_{n,t} \underbrace{\sum_{j=1}^P \sum_{\substack{|\vec{k}|=j \\ j, \vec{k}/t = Pj - \dots - k_{P-1}}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}} \right) = 0. \quad (7.14)$$

To conclude, it is enough noticing that $b_{i,j} = 0$ if $i > j$ and that $b_{i,j} = \frac{\tilde{b}_{i,j}}{\delta^j}$ if $i \leq j$; the first assertion comes from the fact that matrices \mathbb{A}^t together with its inverse \mathbb{B}^t are triangular. The second assertion demands the study of the basis $(\phi_i(X\delta))_{i \in \{0, \dots, P\}}$. Simple calculations show that $a_{i,j} = \tilde{a}_{i,j} \delta^j$ and that $a_{i,i} = \delta^i = \frac{1}{\tilde{b}_{i,i}} \neq 0 \forall i \in \{0, \dots, P\}$. The results are obtained inverting \mathbb{A}^t and studying the coefficients one after another. The system becomes $\forall n \in \{0, \dots, P\}$:

$$\begin{aligned} & \partial_t \left(\tilde{b}_{n,n} \frac{1}{\delta^n} \frac{\bar{u}_n}{n!} + \sum_{j=0}^{n-1} \tilde{b}_{n,j} \frac{1}{\delta^j} \frac{\bar{u}_j}{j!} \right) \\ & + \partial_x \left(f(\bar{u}_0) + \sum_{t=0}^n \tilde{b}_{n,t} \frac{1}{\delta^t} \underbrace{\sum_{j=1}^P \sum_{\substack{|\vec{k}|=j \\ j, \vec{k}/t = Pj - \dots - k_{P-1}}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}} \right) = 0. \end{aligned}$$

Factorizing by $\tilde{b}_{n,n} \frac{1}{\delta^n} \neq 0$ in the $P+1$ equations of (7.15) and letting $\delta \rightarrow 0$ we recover system

$$\forall n \in \{0, \dots, P\}, \quad \partial_t \frac{\bar{u}_n}{n!} + \partial_x \left(\underbrace{\sum_{j=1}^P \sum_{\substack{|\vec{k}|=j \\ j, \vec{k}/n = Pj - \dots - k_{P-1}}} \frac{1}{j!} f^{(j)}(\bar{u}_0) C_{\vec{k}}^j \prod_{i=1}^P \left(\frac{\bar{u}_i}{i!} \right)^{k_i}} \right) = 0, \quad (7.15)$$

corresponding to the reduced model obtained by perturbation methods. ■

The above proof presents several interests:

- it establishes a link between the flow obtained by perturbative methods and the one obtained by the gPC reduced model. Results obtained by perturbative methods can be recovered from the one obtained by gPC under condition $\delta \rightarrow 0$. The link between the approaches is made through changes of variables between the two basis, i.e. relations (7.10) and (7.11). Both approaches are briefly compared numerically in figure 7.1 for Euler equations (mass density at orders 0 and 1) in a 1D Richtmyer-Meshkov-like configuration where a shock hits a perturbed interface initially at rest (more details are given in [232] but also in the next section 7.2).
- The above calculations are also interesting from a modeling points of view. As briefly highlighted before, the previous developments are very similar to the Hilbert ones of chapter 1 used to identify relevant regimes and obtain new models. Consider for example uniform random inputs, the interval $[-\delta, \delta]$ corresponds to the support of the initial perturbation and looking for higher order terms (terms in $\mathcal{O}(\delta)$) may allow obtaining corrective models to perturbation reduced models (called saturation model in the literature). This aspect will not be further developed in this document but may be the object of future works.
- Perturbative methods are also cheaper than gPC based ones. This property makes them a relevant approach for calibration. In [31] and [30], Bayesian inference is applied to calibrate models thanks

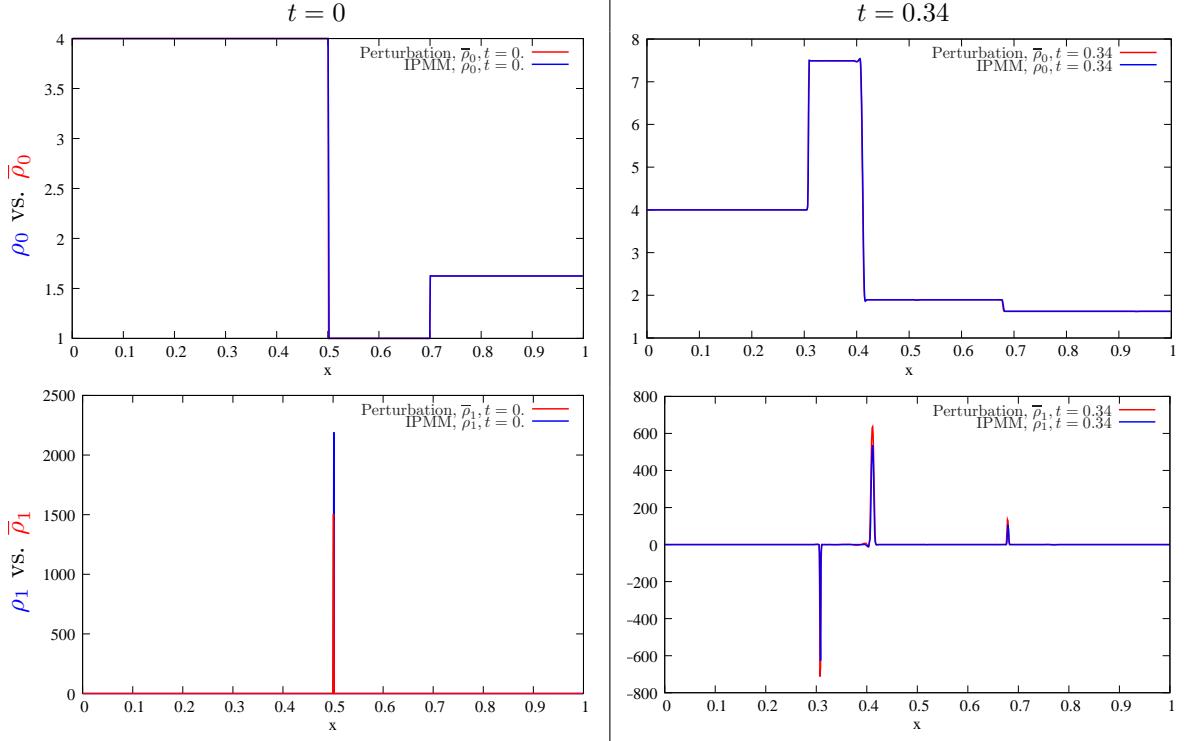


Figure 7.1: Comparison of perturbative and gPC methods on a Richtmyer-Meshkov-like 1D problem (see [232] for the initialization details). For the computations, we chose $\sigma = \delta = 10^{-6}$ with X a uniform random variable. The top pictures corresponds to the mean flows with both methods, the bottom ones for the first order ones.

to comparisons to experiments. Perturbative methods can be used for a fast calibration of the early simulation times (as they allow recovering the gPC results in the linear regime). Once the perturbative model calibrated, its parameters can then be mapped, thanks to the previous developments, into the ones of the gPC model, more relevant but also more costly. This will be emphasized in the next section 7.2.

- Finally, we insist we showed that in the limit $\delta \rightarrow 0$, gPC models recover the results obtained by perturbative methods which are known to build *weakly hyperbolic* models from hyperbolic ones. We insist this does not mean the gPC limit is weakly hyperbolic. It only means in the limit in which perturbative methods are relevant (i.e. the linear regime with small perturbations) the gPC models allow recovering the same observables as perturbative ones.

In this section, we showed that gPC reduced models contain perturbation ones. Perturbation models being used in order to study chaotic flows, we suggest trying to apply the gPC one in the same context. This has been the object of paper [237] in an intrusive manner and of paper [243] in a non-intrusive one. In the next chapter, we suggest revisiting the results published in [243].

7.2 Direct Numerical Simulation (DNS) acceleration *via* gPC

A Direct Numerical Simulation is a simulation in CFD in which the fluid model of interest, usually the Navier-Stokes equation for incompressible fluid, the Euler equations with viscous tensor for compressible ones, is numerically solved without turbulence model. The space and time scales must consequently be explicitly resolved by the discretisation mesh. DNS corresponds to the branch of CFD ensuring the highest fidelity solutions. Turbulence being inherently three dimensional, the computational cost may be prohibitive in certain configurations, the highest the Reynolds number, the finer the grid. For steady flows, intensive use of the ergodicity assumption of the solutions allows performing statistical

treatment over time integration on one numerical experiment but the study of unsteady ones may imply the resolution of a potentially high number of configurations (with different initial conditions). In this sense, DNS for unsteady flows are very similar to the uncertainty quantification studies we considered in this document.

We here aim at practically understanding the similarities and differences between DNS for turbulence and uncertainty quantification: remember (section 3.1.2) Wiener in [295] explicitly built Homogeneous Chaos in order to deal with turbulence problems, to build mathematical tools for which, by construction, ergodicity is ensured. This chapter is only a very first step toward this understanding, probably even the term DNS should not appear (accuse me of baiting the reader here... Maybe) as the performed simulations we consider here are unsteady flows in 2D space dimensions (and not 3D) without viscous effects. The viscous effects not being considered, we will focus on macroscale observable. We will never conclude anything in term of energy dissipation within the mixing zones we will consider. The typical observable will be the size the developing mixing zone/the growth of the initial perturbation. In fact, our hypothesis here are not stronger than the one made applying perturbative methods to the same configurations (see [151, 149, 150, 68, 206, 274, 276, 275, 39]). On the contrary, as testifies the previous section, those hypothesis are even lighter than with perturbative theory as the reduced model is valid also in the nonlinear regime (not only for a small initial perturbation δ , see section 7.1).

With the two above *nonetheless strong* hypothesis, we will compare calibrated results obtained with the non-intrusive gPC approximations to experimental results (and to the ones obtained from a perturbative reduced model).

In the next section, we begin by presenting the configurations and experiments of interest. They correspond to Richtmyer-Meshkov shock tubes and are much more complex than the ‘fil rouge’ problem considered all along the previous chapters. An intrusive gPC study of such experiments in cylindrical configurations has been published in [237]. We here focus on their planar counterparts with a non-intrusive gPC (and i-gPC) application. Regarding the initial uncertain interface position, it will be modeled by a stochastic process. Stochastic process often implies dimensionality issues (one random variable per cell of the simulation, i.e. an increasing number of random variable with spatial discretisation). We will present briefly the Karhunen-Loeve development allowing reducing the dimensionality of the uncertainty quantification problem. We will also spend some time describing the physical model of interest (multi-material Euler system) and defining the statistical observables we aim at recovering with our gPC approximation. In [243], the calibration step and the resolution of a stochastic inverse problem has been made by brute force methods. In [30],[31], we worked on Bayesian inference and were able to recover the results of [243] with automated procedures. In this document, we do not detail Bayesian Inference and the algorithm in order to solve a stochastic inverse problem, for them we refer to [30],[31].

7.2.1 The shock tube experiments and their initial conditions

Let us first describe the main general characteristics of a Richtmyer-Meshkov (RM) shock tube. Figure 7.2 presents its features: two fluids, one at rest (fluid 1 of figure 7.2), the other is shocked (fluid 0 of figure 7.2), are initially separated by a perturbed interface which we suggest to model by a stochastic process. When the shock hits the interface between the fluids, hydrodynamic instabilities are developing creating

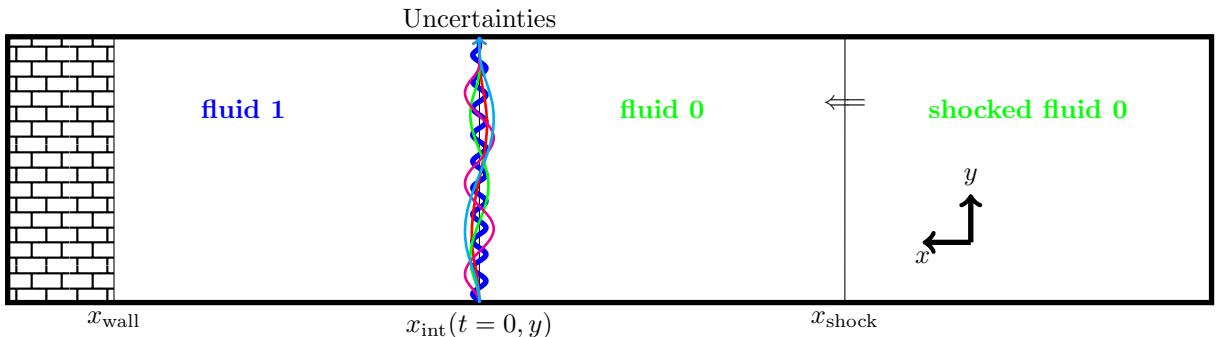
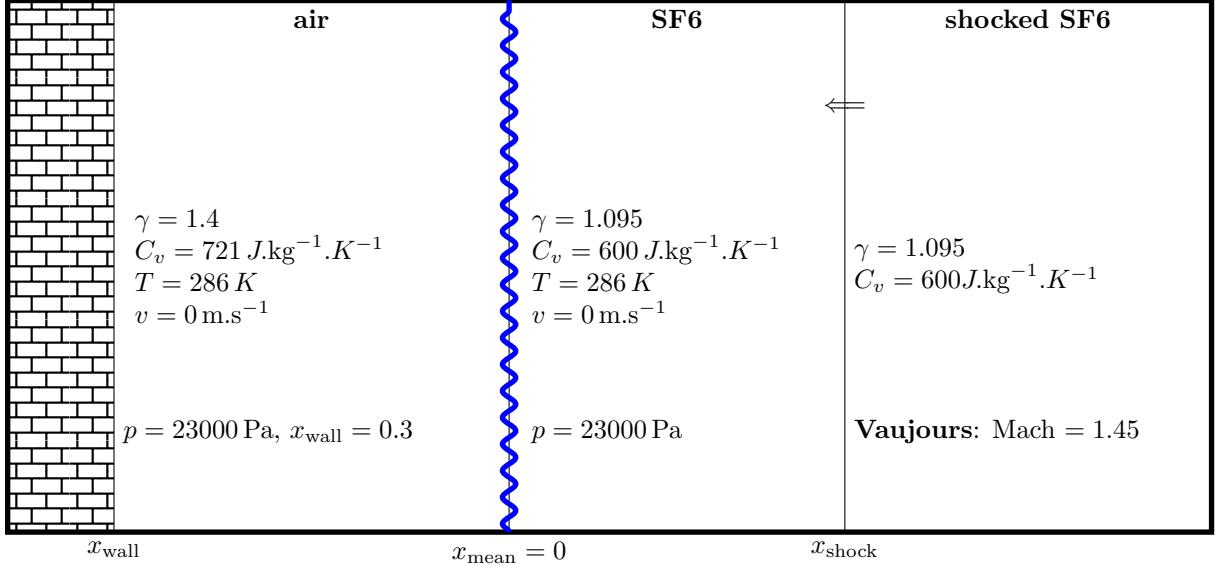


Figure 7.2: General scheme for the initialization of the RM shock tube.

a mixing zone. In this paper, we aim at computing the growth rates of this mixing zone with respect to time and at comparing the numerical results to the experimental ones. Figure 7.3 presents more

Vaujours

$$\text{Atwood number} = \frac{\rho_0 - \rho_1}{\rho_0 + \rho_1} = 0.67$$



CalTech VS

$$\text{Atwood number} = \frac{\rho_0 - \rho_1}{\rho_0 + \rho_1} = 0.67$$

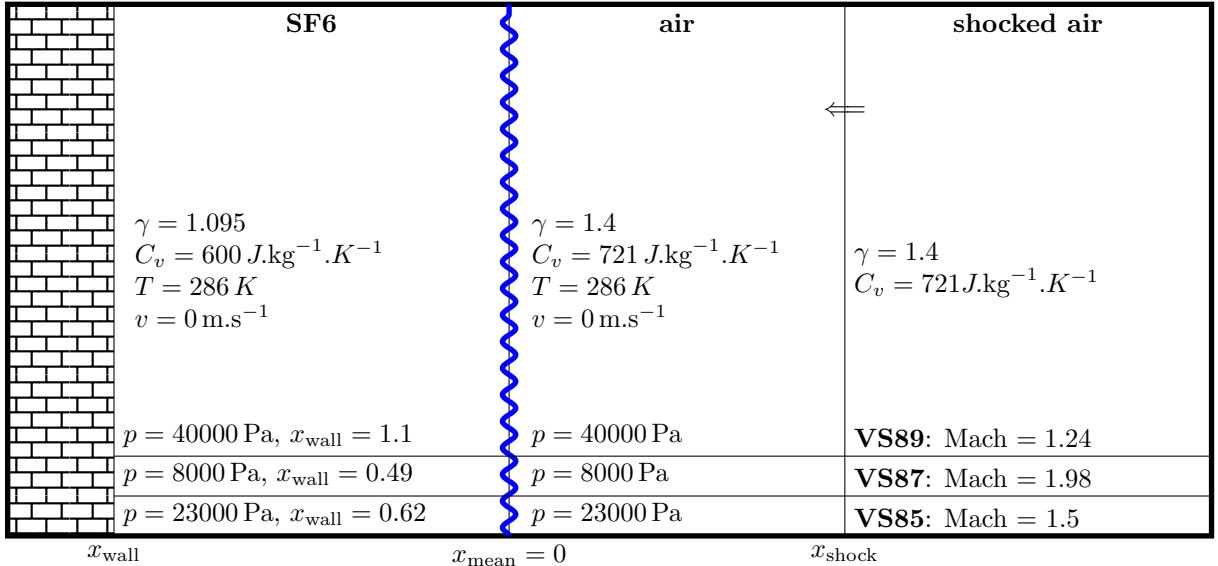


Figure 7.3: Initial conditions for the four RM experiments from [244] (Vaujours) and [286] (Vetter Sturtevant (VS)).

specifically the initial conditions of the shock tubes given in papers [244] (top) and [286] (bottom). Note that in [286], the results are obtained with the same experimental device in three different configurations: the same kind of fluids are used but for different volumes between the interface and the wall and different Mach numbers, see figures 7.3.

In practice, some knowledge of the modeling of the uncertain interface should be gained from experimental data in order to statistically characterise the initial stochastic process. This knowledge is not

provided in the considered papers [244, 286] as the authors probably did not imagine one would revisit the experiments as an uncertainty quantification case. Consequently, we here face a calibration problem: we have to find the stochastic process modeling the initial uncertain interface position recovering the experiments. This has been done using brute force methods in [243] and more subtle ways, intensively applying Bayesian inference, in [30], [31].

In the next section, we introduce a new way (in the document, not new in the literature) to represent a stochastic process, called **Karhunen-Loève** (KL) expansions. They are very convenient especially when one has *a priori* information on the stochastic process to represent, i.e. when one knows its covariance kernel. We suppose it is the case for the initially uncertain interface position as in [237]. Practically, it allows reducing the dimensionality of our uncertainty quantification problem ensuring an efficient use of gPC in order to approximate the observables of interest for times $t > 0$. In a second section, we briefly describe the multimaterial Euler system, recall its properties and present the chosen resolution scheme. The stochastic counterpart of the system is solved non-intrusively with gPC⁵. In the last section, we solve our stochastic model in RM configurations and compare our numerical results to experimental ones.

7.2.2 Stochastic dimension reduction for the initial Uncertain Interface Position

The initially uncertain interface position is modeled by a stochastic process. A stochastic process is a collection of random variables indexed by a parameter $y \in I \subset \mathbb{R}$, $\{F_y(\omega), y \in I, \omega \in \Omega\}$ and such that for fixed y , $F_y(\omega)$ is a random variable. Here, y refers to the vertical component of the interface's spatial position (see figure 7.2).

If we apply a cartesian mesh on the configuration of interest on figure 7.2, especially in the vicinity of the uncertain interface as in figure 7.4, we see that the realisations of the stochastic process cross several cells. This produces as many random variables as cells crossed by the stochastic process. In practice,

The stochastic process has:

- mean $\mu(y) = \mu$ in red on the picture. It is constant along the y -axis in our configurations.
- Variance σ^2 describing the horizontal fluctuations around the mean.
- Covariance $C(y, z)$ describing how points in the domain of fluctuations are correlated.

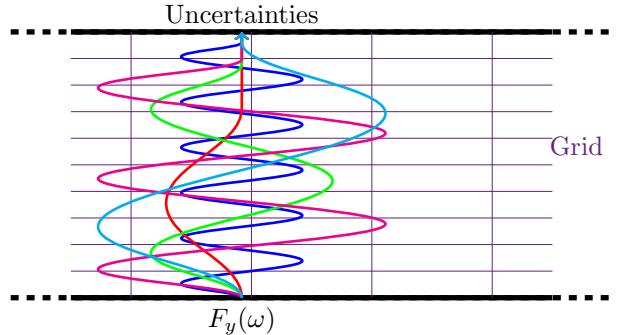


Figure 7.4: Description of the main statistical features of a stochastic process together with a resolution mesh on one realisation of the uncertain interface. Each cell crossed by the stochastic process is a random variable.

this leads to unaffordable stochastic dimension problems for a spectral method such as gPC. To give an idea, the later computations have been performed with 1000 cells in the y -direction and at least 16 cells to capture the initial interface position along the x -axis leading to a stochastic dimension of 16000. Such high dimension problem can only be handled applying an MC method (see figure 5.2 and the discussion of section 5.2.5) and is out of range for gPC.

In order to reduce the stochastic dimension, we rely on a Karhunen-Loeve decomposition. Suppose we initially know the covariance kernel

$$K(y, \zeta) = \mathbb{E}[F_y F_\zeta] = \int F_y(\omega) F_\zeta(\omega) d\mathcal{P}_X(\omega), \quad (7.16)$$

⁵as in [243] and we now know i-gPC confirms the results obtained with gPC, see [30].

of the stochastic process modeling the initial uncertain interface position, see figure 7.4. This kernel partly describes the statistics of the uncertain interface. A Karhunen-Loëve expansion [269] of a stochastic process having mean $\mu(y)$ and covariance $K(y, \zeta)$ defined by

$$F_y(\omega) = \mu(y) + \sum_{n=1}^{\infty} X_n(\omega) \sqrt{\lambda_n} e_n(y), \quad (7.17)$$

is an approximation of the stochastic process $F_y(\omega)_{y \in I, \omega \in \Omega}$ on a basis of orthonormal functions defined by the eigenvectors⁶ $(e_i)_{i \in \mathbb{N}}$ and eigenvalues $(\lambda_i)_{i \in \mathbb{N}}$ of the covariance operator

$$T_K f(y) = \int K(\zeta, y) f(\zeta) d\zeta, \forall f \in L^2(I). \quad (7.18)$$

The coefficients $(X_n)_{n \in \mathbb{N}}$ are centered⁷ normalized⁸ orthogonal random variables defined by

$$\forall i \in \mathbb{N}, \quad X_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_I F_y(\omega) e_i(y) dy. \quad (7.19)$$

In practice, (7.17) is truncated up to an order Q which denotes the stochastic dimension of our uncertainty propagation problem. Consequently, we consider an initial condition approximated by $F_y^Q(X(\omega)) \approx F_y(\omega)$ where $X = (X_1, \dots, X_Q)^t$. If the stochastic process can be accurately represented with a Karhunen-Loeve development with a reasonable size Q then the dimensionality of our problem is reduced and spectral methods may be effective.

7.2.3 The Multimaterial 2D Euler system

In the next section, we describe the uncertain multimaterial 2D Euler system we solve non-intrusively with a gPC approximation. The vector of unknowns is given by

$$U(x, y, t, X) = \begin{pmatrix} \rho(x, y, t, X) \alpha(x, y, t, X) \\ \rho(x, y, t, X) \\ \rho(x, y, t, X) u(x, y, t, X) \\ \rho(x, y, t, X) v(x, y, t, X) \\ \rho(x, y, t, X) e(x, y, t, X) \end{pmatrix},$$

and explicitly depends on $(x, y) \in \mathcal{D} \subset \mathbb{R}^2$, $t \in [0, T] \subset \mathbb{R}^{+,*}$ and $X = (X_1, \dots, X_Q)^t$. In the following sections, the dependences with respect to x, y, t, X are not reminded for the sake conciseness. The quantity α denotes the volume fraction of the fluids (equals to 0 for fluid 0 and 1 for fluid 1). The quantity ρ is the mass density, u and v are the components of the velocity, e is the specific total energy such that $e = \varepsilon + \frac{u^2}{2} + \frac{v^2}{2}$ with ε the specific internal energy. The different quantities are solutions of the uncertain multimaterial Euler system closed with a multimaterial isobare perfect gas closure defined by

$$\begin{cases} \partial_t \rho \alpha + \partial_x \rho u \alpha + \partial_y \rho v \alpha = 0, \\ \partial_t \rho + \partial_x \rho u + \partial_y \rho v = 0, \\ \partial_t \rho u + \partial_x (\rho u^2 + p) + \partial_y (\rho u v) = 0, \\ \partial_t \rho v + \partial_x (\rho u v) + \partial_y (\rho v^2 + p) = 0, \\ \partial_t \rho e + \partial_x (\rho u e + pu) + \partial_y (\rho v e + pv) = 0. \end{cases} \quad (7.20)$$

The first equation corresponds to a closure equation for the mixture model. The second equation corresponds to the mass conservation, the third and fourth to the conservation of momentum and the last one to the total energy conservation. The last equations are given by (7.21)–(7.22): we consider a perfect gas closure (7.21)

$$p(\rho, \varepsilon, \alpha) = (\Gamma(\alpha) - 1)\rho\varepsilon, \quad (7.21)$$

⁶the existence of these eigenvectors and eigenvalues is ensured by Mercer's theorem, see [202].

⁷zero mean: $\mathbb{E}[X_n] = 0, \forall n \in \mathbb{N}$.

⁸standard deviation equals to 1: $\mathbb{E}[X_n^2] = 1, \forall n \in \mathbb{N}$.

relying on the additivity of the internal energies hypothesis at the interface together with an isobare hypothesis (7.22)

$$\Gamma(\alpha) = -\frac{\gamma_0\gamma_1 - \gamma_1 + \alpha\gamma_1 - \alpha\gamma_0}{-\alpha\gamma_1 - \gamma_0 + 1 + \alpha\gamma_0}. \quad (7.22)$$

System (7.20)–(7.21)–(7.22) is hyperbolic under conditions $\varepsilon > 0$ and $0 \leq \alpha \leq 1$. From a numerical point of view, the system for a fixed random variable X is solved thanks to a 3rd order GAIA finite volume scheme with TVD limiters. Every details concerning the deterministic numerical discretisation are given in [154].

7.2.4 Observable of interest, Simulations and Comparisons with Experimental Results

Regarding the (statistical and physical) observable, we are obviously interested in the interface and its instabilities. By essence, this observable is a discontinuity. On such quantity, gPC will provide a poor convergence rate. The idea here is to slightly change the physical observable and consider a smooth one, still related to the interface position and relatively relevant (i.e. without losing too much information), on which spectral convergence will be bound to occur. For this reason, we are here interested in the size of the mixing zone with respect to time. It has to be defined in term of probabilistic quantity. For this, let us first denote by $x_{\text{int}}(t, y, X)$ the uncertain interface position⁹. The size of the mixing zone $\Delta x_{\text{int}}(t)$ must be averaged over the y -axis to obtain a scalar observable consistent with the information available in the experimental papers [244, 286]. Finally it must be reinterpreted in term of statistical quantity: we suggest defining the size of the mixing zone as the interval in which the probability of having x_{int} is beyond a certain threshold α . The quantity $\Delta x_{\text{int}}(t)$ can then be expressed as

$$\Delta x_{\text{int}}(t) = \int_I (x_{\text{max}}^\alpha(t, y) - x_{\text{min}}^\alpha(t, y)) dy, \quad (7.23)$$

in which x_{min}^α and x_{max}^α are such that

$$\begin{aligned} \mathbb{P}(x_{\text{int}}(t, y, X) \leq x_{\text{max}}^\alpha(t, y)) &= \alpha, & \mathbb{P}(x_{\text{int}}(t, y, X) \geq x_{\text{min}}^\alpha(t, y)) &= \alpha, \\ F_{x_{\text{int}}(t, y, X)}(x_{\text{max}}^\alpha(t, y)) &= \alpha, & F_{x_{\text{int}}(t, y, X)}(x_{\text{min}}^\alpha(t, y)) &= 1 - \alpha, \end{aligned}$$

leading to

$$x_{\text{max}}^\alpha(t, y) = F_{x_{\text{int}}(t, y, X)}^{-1}(\alpha), \quad \text{and} \quad x_{\text{min}}^\alpha(t, y) = F_{x_{\text{int}}(t, y, X)}^{-1}(1 - \alpha).$$

In the above expression, x_{max}^α and x_{min}^α depends on $x_{\text{int}}(t, y, X)$, which will be approximated with a gPC development by applying the materials of chapters 5 and 6: $x_{\text{int}}(t, y, X) \approx x_{\text{int}}^P(t, y, X) = \sum_{k=0}^P x_{\text{int}, k}(t, y) \psi_k(X)$ where $(x_{\text{int}, k})_{k \in \{0, \dots, P\}}$ are the coefficients in the gPC basis obtained non-intrusively (numerical integration). But before introducing this approximation, let us focus on their definitions:

$$\begin{aligned} F_{x_{\text{int}}(t, y, X)}(x_{\text{max}}^\alpha(t, y)) &= \alpha, \\ \int \mathbf{1}_{]-\infty, x_{\text{max}}^\alpha(t, y)]}(x) d\mathcal{P}_{x_{\text{int}}(t, y, X)}(x) &= \alpha, \\ \int \mathbf{1}_{]-\infty, x_{\text{max}}^\alpha(t, y)]}(x_{\text{int}}(t, y, X)) d\mathcal{P}_X(X) &= \alpha. \end{aligned} \quad (7.24)$$

In fact, x_{max}^α and x_{min}^α are both α -quantiles of the interface positions. A proportion of the interface position will be below x_{max}^α and another will be above x_{min}^α . This is emphasized in figure 7.2.4: it first shows one realisation of the interface (in fact it shows the volume fraction $\alpha(x, y, t = 12ms)$). The combination of (7.23) and (7.24) has to be compared to (3.1) and an attempt to apply Birkhoff's ergodic theorem: to evaluate certain quantities evolving with respect to time, we rely on averaging over a multidimensional measure (integration with respect to $d\mathcal{P}_X$ in (7.24)).

This was for the intention. Now, if we go back to more technical considerations, in (7.24) the interface position stochastic process $x_{\text{int}}(t, y, X)$ is approximated by a gPC development, denoted by $x_{\text{int}}^P(t, y, X)$.

⁹Note that $x_{\text{int}}(0, y, X(\omega)) = F_y^Q(X(\omega)) \approx F_y(\omega)$.

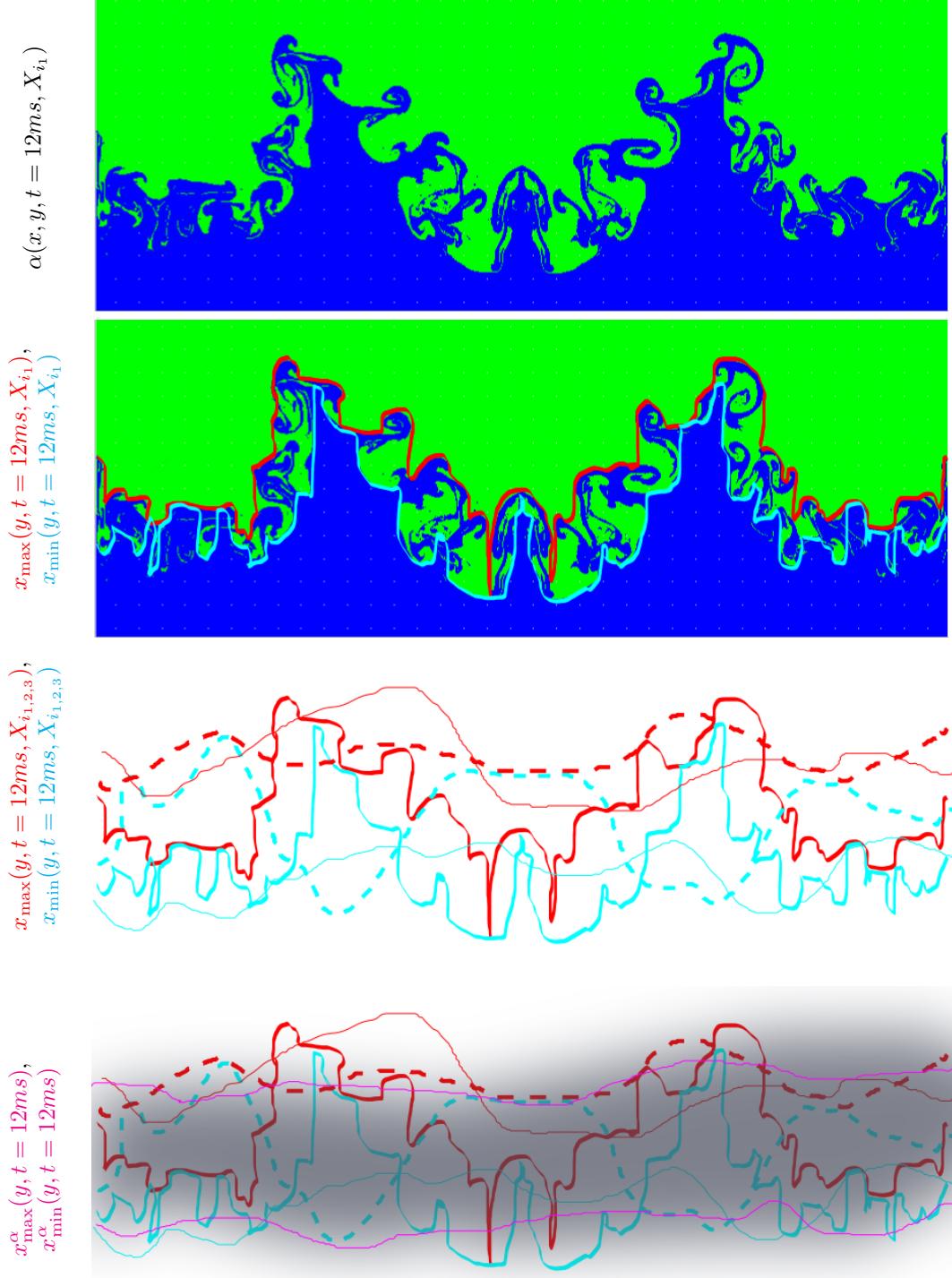


Figure 7.5: The top picture shows a zoom on the interface of one realisation X_{i_1} of a perturbed interface (volume fraction $\alpha(x, y, t = 12ms, X_{i_1})$). The second one superposes the x_{\max}^{α} and x_{\min}^{α} of the interface for this realisation. They are extracted and the same is done for other realisation (third picture with three other realisations corresponding to $X_{i_1}, X_{i_2}, X_{i_3}$). On the bottom picture, the α -quantiles are represented and are obtained from (7.24).

Second point, in (7.24), the unknown is x_{\max}^{α} for a given α . In other words, we have to numerically inverse relation (7.24) to obtain an estimation of $x_{\max}^{\alpha}(t, y)$ which stands for the upper extremity of the

mixing zone for at time t and ordinate y . The same applies to x_{\min}^α , the lower one. To perform the inversion, we rely on a Monte-Carlo (MC) sampling of X in the gPC metamodel/stochastic surrogate model/approximation $x_{\text{int}}^P(t, y, X)$ and an evaluation of $x_{\max}^\alpha, x_{\min}^\alpha$ from the last line of (7.24).

Remark 7.1 Note that this study has been carried out before the introduction of *i-gPC*. Consequently, care has been taken to approximate a relatively smooth observable with a gPC representation, in order to take advantage of an efficient convergence rate. In [30], *i-gPC* has been applied in this same context and recovered similar results. Due to the smoothness of the observable, *i-gPC* and gPC have the same performances.

Now remains to apply the above material – i.e. modelization of an uncertain interface through a stochastic process approximated by a KL expansion, resolution of the stochastic Euler (7.20) system at quadrature points with GAIA scheme and approximation of the stochastic process for $t > 0$ thanks to non-intrusive gPC – in order to approximate the uncertain position of the interface with respect to time.

Figure 7.6 shows the evolution of the size of mixing zones with respect to time: it presents a comparison between the experimental results of [244] (top) and [286] (bottom) and the numerical results obtained by solving our stochastic model and computing (7.23). The circles correspond to the results deduced from the experiments of [244, 286]. The size of the circles corresponds to an evaluation of the confidence in the experimental results: some uncertainties are remaining, for example concerning the temperature during the experiments¹⁰. Besides, the experimental mixing zone sizes are estimated from pictures implying possible chronometry difficulties and measurement uncertainties plus our retranscription errors¹¹. Still, we suppose these remaining uncertainties can be neglected in comparison to the one beared by the initial uncertain interface position.

The dotted points in figure 7.6 correspond to the numerical results – approximation of $\Delta x_{\text{int}}(t)$ of (7.23) – obtained by solving the stochastic Euler system with uncertain initial interface position with non-intrusive gPC. The held covariance kernels used in the experiments are given by

$$K_{V \text{aujours}}(x, y) = 0.002 \exp\left(-\frac{|x - y|}{2.5}\right), \text{ for [244]}, \quad (7.25)$$

and by

$$K_{VS}(x, y) = 0.002 \exp\left(-\frac{|x - y|}{0.8}\right), \text{ for [286]}. \quad (7.26)$$

We truncated the KL representations after the three main modes ($Q = 3$) and chose X as a vector of three independent uniform random variables on $[-1, 1]$: this last choice is arbitrary as we can not extract more information on the initial conditions from papers [244, 286]. The choice of a bounded support is consistent with the fact that the uncertainty is very localized initially and the independency hypothesis is convenient for computations. We used an order $P = 3$ in each stochastic directions for the gPC development, leading to $(P + 1)^3 = (3 + 1)^3 = 64$ gPC coefficients (full tensor of the gPC basis) estimated thanks to a full tensorization of the Gauss-Legendre quadrature rule, 3 points in each stochastic directions leading to 27 deterministic independent runs of the code for one curve.

On figure 7.6 (left), two numerical curves are displayed obtained with a spatial discretisation of 8 and 16 cells per amplitude¹² of the initial perturbation¹³. Both numerical results fit the experimental ones in the linear regime together with the nonlinear one and can be considered converged: the difference between the two curves can be explained by the use of an MC method in order to estimate the α -quantiles of $x_{\text{int}}^P(t, y, X)$. In other words, we experimentally observe a convergence behaviour for the observable of interest once our methodology applied. Theoretical convergence have not been investigated in our context but in a recent paper [109], the authors showed convergence results in some particular measured spaces very similar to our application context. It certainly represents a good kickstart for eventual theoretical analysis of our methodology.

¹⁰We fixed it to the ambient temperature (286 K).

¹¹We obtained the plots by directly reading values on the figures of the considered publication [244, 286].

¹²respectively 500×500 and 1000×1000 cells.

¹³On these problems, the modal discretisation can be considered fully resolved.

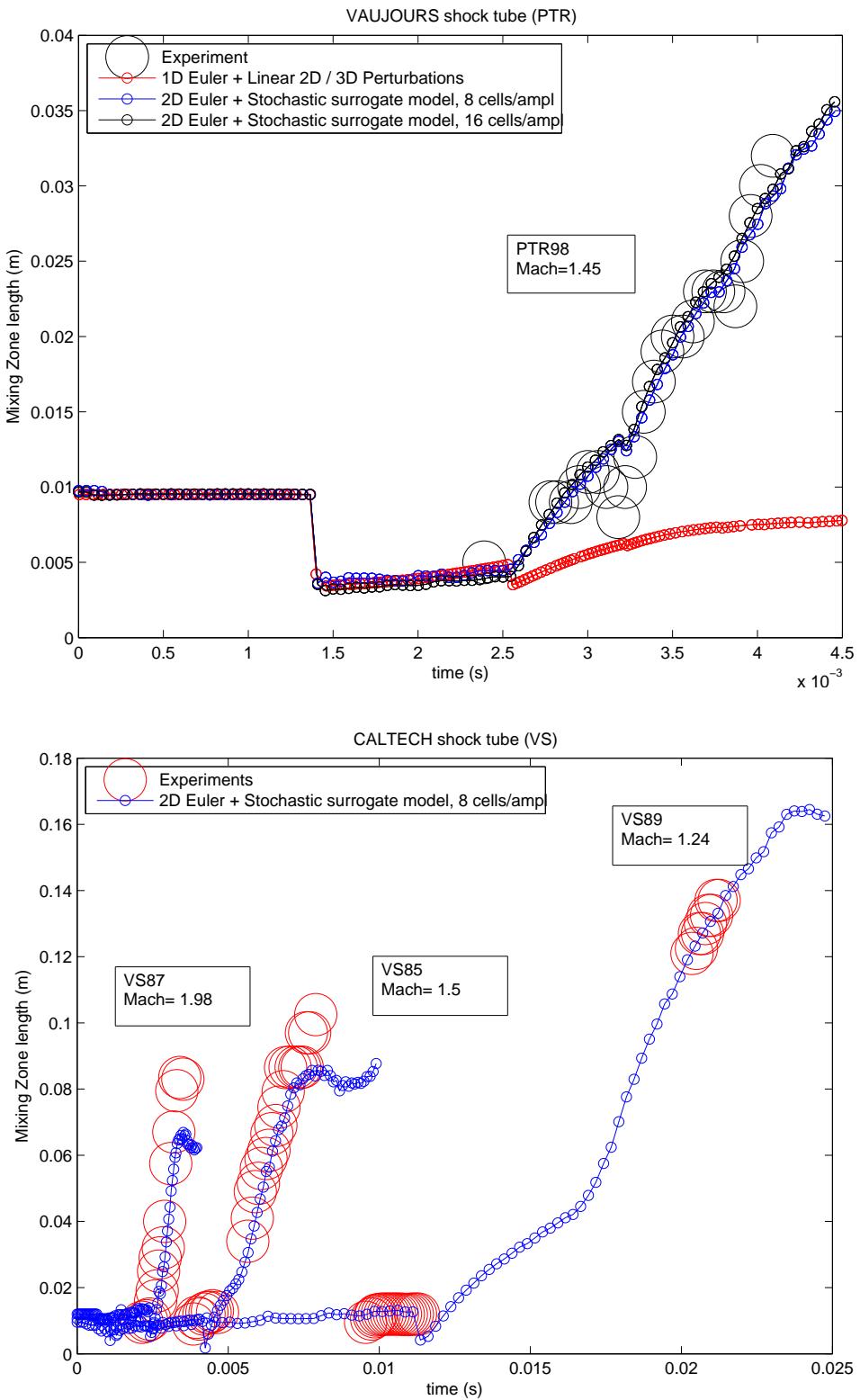


Figure 7.6: Comparisons between experimental (circles) and numerical (dots) results obtained with our stochastic model. Top: experiment from [244] with initial conditions of figure 7.3 top-left. Bottom: experiments from [286] with the initial conditions of figure 7.3 top-right, bottom-left and bottom-right.

Figure 7.6 (right) shows the same kind of results obtained on the three shock tubes of paper [286]. We calibrated the model – looked for the kernel allowing to recover the experimental results – on the second RM shock tube (VS85) and fed the two other studies (VS87 and VS89) with this same covariance (7.26) as an initial uncertain interface position: these results show the model is predictive on experiments obtained from the same experimental device.

For VS85 and VS87, the last experimental circles are not reached by our model. Several hypothesis can be made in order to explain this: first, it is known that gPC suffers long term integration problems [121, 292]. Besides, they can also be explained by 3D effects/flows. In this document, we only performed 2D simulations. We also considered initial uncertain interface positions with constant means which can be restrictive¹⁴. Finally, we can not forget an eventual inaccuracy or more important uncertainty in the experiments for high times. These hypothesis will be investigated in further publications.

7.3 Conclusion for the gPC application to chaotic flows

We have suggested a new process in order to predict the growth rates of mixing zones between two fluids in RM shock tubes. The modelization relies on a reinterpretation of shock tube experiments as uncertainty driven ones. This implies modeling the uncertain initial interface position by a stochastic process and solving the stochastic Euler system rather than relying on a turbulence model (as in DNS). The stochastic model and the statistics of the uncertain interface position with respect to time are recovered by building their gPC approximations.

Numerical results are compared to experimental ones and show a good agreement provided the parameters of the covariance kernel representing the initially uncertain interface position are known. The use of Karhunen-Loeve development allows for an important dimensionality reduction permitting the application of spectral methods. The new stochastic methodology seems predictive in the sense calibrating one simulation with respect to one experiment allows the prediction of the other experiments obtained with the same experimental device. The numerical results also presented a very interesting convergence behaviour: it has been observed in the numerical experiments but not tackled theoretically. In [109], the authors surely bring a theoretical background to the above numerical results/experiments and the paper represents a good starting point to new theoretical studies for our context. This can be counted amongst the possible perspectives of this work as the previous study can even be interpreted as an acceleration of the studies performed in [109] (together with comparisons to experiments). In this context, the capability for gPC to accelerate Monte-Carlo computations will be demonstrated in section 9.12 of chapter 9 for the resolution of the linear Boltzmann equation.

Several important points have been tackled since the publication of the above material in [243]: first, the new non-intrusive gPC based uncertainty propagation method, i-gPC, presented in chapter 6 (see [238]) has been applied to the same problems in [30]. The i-gPC approximations gave the same results as gPC, even for the later simulation times. This is mainly due to the smoothness of the observable. Secondly, in [243], due to the lack of information on the features of the initial interface position, we applied a brute force method in order to calibrate the covariance kernel describing the initial condition. In [30], the same kinds of stochastic inverse problem have been successfully solved applying Bayesian Inference. The methodology is described in [31]. It allows automatizing the calibration step. This last one can even be accelerated using perturbative models for a fast calibration by fitting first the early times before relying on the finer gPC one.

¹⁴as the pressure on the interface can be different at the center of the domain than at the boundaries for example.

Chapter 8

Toward an application of Cameron-Martin's theorem (not only its special case)

An attempt to apply the full version of Cameron-Martins's theorem

Contents

8.1	An attempt to apply theorem 3.3: an i-gPC decomposition of the residue	139
8.1.1	Analysis of theorem 3.3 and comparison to theorem 3.4	139
8.1.2	i-gPC decomposition of the residue in an infinite integration accuracy context .	140
8.1.3	i-gPC decomposition of the residue in a finite integration accuracy context . .	142
8.2	Numerical Applications of the i-gPC decomposition of residue method . .	144
8.2.1	Some (hydrodynamically motivated) 1D test-problems	144
8.2.2	Some (well-known in the literature) multidimensional test-cases	155
8.3	Summary for the i-gPC decomposition of the residue algorithm	159

The material of this chapter may appear singular. In the previous ones, especially in those dealing with the non-intrusive application of generalized Polynomial Chaos, the convergence theorem invoked to justify the mathematical legitimacy of the built gPC approximation was in fact *the special case* of Cameron-Martin's theorem (i.e. theorem 3.4). This has been briefly tackled in chapter 3¹ in which we identified theorem 3.4 as the spectral counterpart of the Central Limit Theorem for Monte-Carlo methods. In this chapter, we analyse the *complete* Cameron-Martin theorem 3.3² and build a new gPC based approximation method from it. Of course, invoking theorem 3.3 instead of theorem 3.4 is of poor interest if the gPC approximation obtained from the first one is accurate enough. But all along the previous chapters, we identified *regimes* for which the convergence of the gPC approximation is relatively slow. With i-gPC in chapter 6, we increased the accuracy of a gPC approximation³ but we identified regimes (stagnation) for which the improved approximation still strongly relies on theorem 3.4⁴. The i-gPC approximation can face two configurations, see inequalities (6.7)–(6.13), in a finite numerical integration accuracy context:

1. Numerical integration accuracy is reached after several iterations (i.e. we have a negligible truncation error): it corresponds to the ideal case as the number N of runs of the simulation code is

¹and more precisely in section 3.1.3.

²instead of its special case 3.4.

³especially for discontinuous solutions.

⁴i.e. the polynomial order has to be increased in order to improve the accuracy.

in general the limiting factor (costly codes) and they are consequently fully used/harnessed. An illustration of this behaviour is presented in section 6.2.1.

2. The accuracy of the approximation stagnates after a certain iteration j_0 and the truncation error remains preponderant. Consequently some information is still available from the N simulations points. An illustration is given in section 6.2.2.

We suggest here to revisit theorem 3.3 in order to improve the accuracy of the i-gPC approximation when the second case is encountered (see the numerical example of section 6.2.2). This new method is made possible by the use of i-gPC, this will be emphasized in the following discussion. This chapter is the only one in part II in which systems of conservation laws are only mentionned⁵. Every test-problems presented here are very simple ones. They are nonetheless motivated and built from difficulties encountered in hydrodynamical configurations (some are presented in [31]). Note that the material of this section has not been published in any paper.

8.1 An attempt to apply theorem 3.3: an i-gPC decomposition of the residue

In this chapter, we are motivated by revisiting theorem 3.3 in order to build a new method, a new algorithm, in order to improve the accuracy in case 2.) detailed above. For this, we first compare more precisely theorem 3.3 and theorem 3.4 and identify a new degree of freedom which is yet waiting to be exploited.

8.1.1 Analysis of theorem 3.3 and comparison to theorem 3.4

As explained before, the aim of this first section is to understand what has been neglected in the special case theorem 3.4 with respect to the complete theorem 3.3. Let us first recall the notations of section 3.1.3: remember that for any functional $u(f)$ satisfying the conditions of theorem 3.3, we have

$$\|u(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2 = \int_{C^0([a,b])}^w \left| u(f) - \sum_{m_1, \dots, m_P}^P u_{m_1, \dots, m_P} \Psi_{m_1, \dots, m_P}(f) \right|^2 d_w f < \infty \xrightarrow[P \rightarrow \infty]{} 0. \quad (8.1)$$

The coefficients u_{m_1, \dots, m_P} are the Fourier-Hermite coefficients defined by

$$u_{m_1, \dots, m_P} = \int_{C^0([a,b])}^w u(f) \Psi_{m_1, \dots, m_P}(f) d_w f,$$

and we introduced a more concise notation for both the norm on the space of $u(f)$ and the P^{th} order approximation $u^P(f)$ of $u(f)$. Now, we suggest plugging a new Q -dimensional functional $g(f)$ satisfying the condition of the special case (theorem 3.4) into (8.1). At this stage of the discussion, g is arbitrary. We can write

$$\|u(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2 = \|u(f) - g(f) + g(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2. \quad (8.2)$$

Now set $g(f) = u^{P_0}(f)$ with $P_0 \in \mathbb{P}$ such that g is a Q -dimensional polynomial of given maximal orders P_i in each directions $i \in \{1, \dots, Q\}$. With this obvious choice, we have

$$\begin{aligned} \|u(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2 &= \|u(f) - g(f) + g(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2, \\ &= \|u(f) - u^{P_0}(f) + u^{P_0}(f) - u^P(f)\|_{L^2_{C^0([a,b])}}^2, \\ &= \left\| u(f) - u^{P_0}(f) - \sum_{\substack{m_1, \dots, m_P \\ P > P_0}}^P u_{m_1, \dots, m_P} \Psi_{m_1, \dots, m_P}(f) \right\|_{L^2_{C^0([a,b])}}^2. \end{aligned} \quad (8.3)$$

⁵and not solved.

Introduce $R^{P_0}(f) = u(f) - u^{P_0}(f)$, the residue in the above norm of the functional $u(f)$ with respect to the P_0^{th} order approximation. Introduce also its P order polynomial approximation defined by

$$R^P(f) = \sum_{\substack{m_1, \dots, m_P \\ P > P_0}}^P u_{m_1, \dots, m_P} \Psi_{m_1, \dots, m_P}(f).$$

Then $R^P(f)$ is a polynomial approximation of order P with increasing dimensions. By denoting by $\psi_{q_1, \dots, q_P}(f)$ a simple renumerotation of the previous basis $\Psi_{m_1, \dots, m_P}(f)$ which drops the Q first components of $\Psi_{m_1, \dots, m_P}(f)$ at every polynomial orders, we obtain that $R^P(f)$ can be rewritten as

$$R^P(f) = \sum_{q_1, \dots, q_P}^P r_{q_1, \dots, q_P} \psi_{m_1, \dots, m_P}(f).$$

As such, $R^P(f)$ is a converging approximation of the residue $R^{P_0}(f)$ in a new basis, orthogonal to the initial one: we have

$$\|u(f) - u^P(f)\|_{L^2_{c^0([a,b])}}^2 = \|R^{P_0}(f) - R^P(f)\|_{L^2_{c^0([a,b])}}^2 \xrightarrow[P \rightarrow \infty]{} 0, \quad (8.4)$$

as the left hand side tends to zero with P .

The idea behind the comparison of both theorems/approximations is quite simple. The construction of a new algorithm taking advantage of this idea is more complex. To be efficient, the iterative process would have to *reuse* the initial experimental design and improve (or at least preserve) the accuracy thanks to an approximation of the residue of the solution. The idea is also very close to the kriging-gPC principle (see section 5.3.3) in which a gaussian process is introduced to approximate the residue.

8.1.2 i-gPC decomposition of the residue in an infinite integration accuracy context

We aim at incorporating the remarks of the previous section into a new algorithm in order to improve the approximation of an output with respect to an i-gPC development. Let X denote an input random variable of known probability measure $d\mathcal{P}_X$ and $X \rightarrow u(X)$ denote a transformation of X which we want to estimate.

For the first step of our new method, we suggest applying first i-gPC⁶: let us denote by $u_Z^P(Z)$ the random variable obtained from the i-gPC approximation of $u(X)$ where Z denotes the last random variable of iteration $Z = Z^{k_{\text{last iteration}}}$ of the i-gPC process. The second step consists in building the residue R of the i-gPC approximation $u_Z^P(Z)$ with respect to $u(X)$, i.e.

$$R(X) = u(X) - u_Z^P(Z(X)). \quad (8.5)$$

The idea now is to approximate it by its i-gPC development with initial gPC basis $(\phi_k^X(X))_{k \in \{0, \dots, P\}}$. This is made possible due to the fact that the initial gPC basis $(\phi_k^X(X))_{k \in \{0, \dots, P\}}$ is not necessarily orthonormal to the final gPC basis obtained after several iterations of i-gPC. We denote by $R_U^P(U) \approx R$ this i-gPC development where U denotes the last random variable of iteration $U = U^{k_{\text{last iteration}}}$ of the i-gPC process. If we now consider the sum of random variables $u_Z^P(Z) + R_U^P(U)$ then we have the following result.

Property 8.1 *With the previous notations, we have*

$$\|u(X) - (u_Z^P(Z) + R_U^P(U))\|_{L^2(\Omega)} \leq \|u(X) - u_Z^P(Z)\|_{L^2(\Omega)}. \quad (8.6)$$

Consequently, a new iterative approximation method is built. It enriches, for fixed polynomial order P , the number of random components of the gPC basis (it now depends on Z and U and not only on Z) as suggested by the *complete* Cameron-Martin's theorem 3.3. Note that the enrichment is only additive for the moment (we do not consider high order cross products of u_Z^P with R_U^P for example).

⁶which suppose applying first gPC.

Proof In order to prove the previous property, let us expand (8.6) into

$$\begin{aligned} \|u(X) - (u_Z^P(Z) + R_U^P(U))\|_{L^2(\Omega)}^2 &= \|u(X) - u_Z^P(Z)\|_{L^2(\Omega)}^2 + \underbrace{\|R_U^P(U)\|_{L^2(\Omega)}^2}_{-\underbrace{2\mathbb{E}[(u(X) - u_Z^P(Z))R_U^P(U)]}_{\Gamma_2}} \\ &\quad - 2\mathbb{E}[(u(X) - u_Z^P(Z))R_U^P(U)]. \end{aligned} \quad (8.7)$$

By definition of the i-gPC development of the residue R , we have

$$R \approx R_U^P(U) = \sum_{k=0}^P r_k^U \phi_k^U(U), \quad (8.8)$$

with $(\phi_k^U(U))_{k \in \mathbb{N}}$ the orthonormal basis with respect to $d\mathcal{P}_U$ and $\forall k \in \{0, \dots, P\}$

$$r_k^U = \mathbb{E}[R\phi_k^U(U)] = \int R(F_X^{-1}(F_U(u)))\phi_k^U(u)d\mathcal{P}_U(u). \quad (8.9)$$

In the previous expression, F_X and F_U are the cdfs of the random variables X and U . Consequently, we have

$$\Gamma_1 = \int (R_U^P(u))^2 d\mathcal{P}_U(u) = \sum_{k=0}^P (r_k^U)^2 \geq 0. \quad (8.10)$$

Besides, using the definition of $R = u(X) - u_Z^P(Z)$ and (8.8) one has

$$\begin{aligned} \Gamma_2 &= \mathbb{E}[R \times R_U^P(U)], \\ &= \sum_{k=0}^P \int R(F_X^{-1}(F_U(u)))r_k^U \phi_k^U(u)d\mathcal{P}_U(u), \\ &= \sum_{k=0}^P r_k^U \mathbb{E}[R\phi_k^U(U)] = \sum_{k=0}^P (r_k^U)^2, \\ &= \Gamma_1. \end{aligned} \quad (8.11)$$

Finally we get

$$\|u(X) - (u_Z^P(Z) + R_U^P(U))\|_{L^2(\Omega)}^2 = \|u(X) - u_Z^P(Z)\|_{L^2(\Omega)}^2 - \Gamma_1, \quad (8.12)$$

with $\Gamma_1 \geq 0$ so that

$$\|u(X) - (u_Z^P(Z) + R_U^P(U))\|_{L^2(\Omega)}^2 \leq \|u(X) - u_Z^P(Z)\|_{L^2(\Omega)}^2. \quad (8.13)$$

This ends the proof. ■

The new iterative approximation is not canonical in the sense the inequality in (8.6) is not strict and the approximation may stagnate. Some additional (regularity) hypothesis may allow identifying regimes for which the inequality is strict but as we aim at applying the approach to general applications. We consequently prefer keeping this general statement. Before studying any stagnation regime, the numerical analysis of i-gPC performed in the previous chapter showed it is important considering finite numerical integration accuracy *prior* to other considerations⁷. This is done in the next section. Nonetheless, inequality (8.6) shows what can be expected asymptotically with accurate integration techniques and seems interesting enough to go on in this direction.

⁷due to the possible appearance of numerical instabilities.

8.1.3 i-gPC decomposition of the residue in a finite integration accuracy context

This section is the *finite integration accuracy* counterpart of the previous one. It may appear redundant and the reader not interested in the proof could directly skip this part and consider directly property 8.2. For the interested reader, the proof gives hints at how intertwined integration and truncation errors may perturb the new algorithm.

As in the previous parts, suppose X denotes an input random variable of known probability measure $d\mathcal{P}_X$. Let $u(X)$ denote a transformation of X which we want to approximate. To do so, we suggest applying first i-gPC: let us denote by $u_{Z^N}^P(Z^N)$ the random variable obtained from the i-gPC approximation of $u(X)$ where Z^N denotes the last random variable of iteration $Z^N = Z^{N,k_{\text{last iteration}}}$ of the i-gPC process. The second step of our decomposition consists in *approximating* the residue R^N of the i-gPC approximation $u_{Z^N}^P(Z^N)$ with respect to $u(X)$,

$$R^N = u(X) - u_{Z^N}^P(Z^N). \quad (8.14)$$

To do so, we apply i-gPC approximation with respect to the initial gPC basis $(\phi_k^X(X))_{k \in \{0, \dots, P\}}$. For numerical integration, we rely *only on the initial experimental design* $(X_l, w_l)_{l \in \{1, \dots, N\}}$, just as for i-gPC in the previous chapter. We build the residue at these points

$$(R^N(X_l), w_l)_{l \in \{1, \dots, N\}} = (u(X_l) - u_{Z^N}^P(Z^N(X_l)), w_l)_{l \in \{1, \dots, N\}}.$$

Now, we denote by $R_{U^N}^P(U^N) \approx R^N$ the i-gPC development of the residue R^N where U^N denotes the last random variable of iteration $U^N = U^{N,k_{\text{last iteration}}}$ of the i-gPC process. By considering the sum of random variables $u_{Z^N}^P(Z^N) + R_{U^N}^P(U^N)$ then we have the following results.

Property 8.2 *With the previous notations and considering finite numerical integration accuracy, we have*

$$\begin{aligned} \|u(X) - (u_{Z^N}^P(Z^N) + R_{U^N}^P(U^N))\|_{L^2(\Omega)} - \|u(X) - u_{Z^N}^P(Z^N)\|_{L^2(\Omega)} &= e_p^{U^N, N} \\ &\quad - \sum_{k=0}^P (r_k^{U^N})^2. \end{aligned} \quad (8.15)$$

The term $e_p^{U^N, N}$ is the projection error of the i-gPC approximation of the residue.

Of course, asymptotically (infinite integration accuracy), we recover the inequality (8.6) and the approximation $u_Z^P(Z) + R_U^P(U)$ ensures a gain with respect to the i-gPC approximation $u_Z^P(Z)$.

Proof Let us expand the first term in (8.15) into

$$\begin{aligned} \|u(X) - (u_{Z^N}^P(Z^N) + R_{U^N}^P(U^N))\|_{L^2(\Omega)}^2 &= \|u(X) - u_{Z^N}^P(Z^N)\|_{L^2(\Omega)}^2 + \underbrace{\|R_{U^N}^P(U^N)\|_{L^2(\Omega)}^2}_{\Gamma_1^N} \\ &\quad - 2 \underbrace{\mathbb{E} [(u(X) - u_{Z^N}^P(Z^N)) R_{U^N}^P(U^N)]}_{\Gamma_2^N}. \end{aligned} \quad (8.16)$$

By definition of the i-gPC development of the residue, we have

$$R^N \approx R_{U^N}^P(U^N) = \sum_{k=0}^P r_k^{U^N, N} \phi_k^{U^N}(U^N), \quad (8.17)$$

with $(\phi_k^{U^N}(U^N))_{k \in \mathbb{N}}$ the orthonormal basis with respect to $d\mathcal{P}_{U^N}$. Besides, $\forall k \in \{0, \dots, P\}$ we have

$$r_k^{U^N} = \mathbb{E}[R\phi_k^{U^N}(U^N)] = \int R(F_X^{-1}(F_{U^N}(u))) \phi_k^{U^N}(u) d\mathcal{P}_{U^N}(u). \quad (8.18)$$

It holds also for its approximation

$$\begin{aligned} r_k^{U^N, N} &= \sum_{k=1}^N R^N(F_X^{-1}(F_{U^N}(X_i)))\phi_k^{U^N}(U^N(X_i))w_i, \\ r_k^{U^N, N} &= \sum_{k=1}^N (u(X_i) - u_{Z^N}^P(Z^N(X_i)))\phi_k^{U^N}(U^N(X_i))w_i. \end{aligned} \quad (8.19)$$

Once again, F_X and F_{U^N} are the cdfs of the random variables X and U^N . Consequently, we have

$$\Gamma_1^N = \int (R_{U^N}^P(u))^2 d\mathcal{P}_{U^N}(u) = \sum_{k=0}^P (r_k^{U^N, N})^2 \geq 0. \quad (8.20)$$

Using the definition of $R = u(X) - u_{Z^N}^P(Z^N)$ and (8.17) one has

$$\begin{aligned} \Gamma_2^N &= \mathbb{E}[R^N \times R_{U^N}^P(U^N)], \\ &= \sum_{k=0}^P \int R^N(F_X^{-1}(F_{U^N}(u)))r_k^{U^N, N}\phi_k^{U^N}(u)d\mathcal{P}_{U^N}(u), \\ &= \sum_{k=0}^P r_k^{U^N, N} \mathbb{E}[R\phi_k^{U^N}(U^N)] = \sum_{k=0}^P r_k^{U^N} r_k^{U^N, N}. \end{aligned} \quad (8.21)$$

Let us introduce the projection error on the residue

$$(e_{\text{int}, P}^{U^N, N})^2 = \sum_{k=0}^P (r_k^{U^N} - r_k^{U^N, N})^2 = \Gamma_1^N + \sum_{k=0}^P (r_k^{U^N})^2 - 2\Gamma_2^N. \quad (8.22)$$

Finally we get

$$\begin{aligned} \|u(X) - (u_{Z^N}^P(Z^N) + R_{U^N}^P(U^N))\|_{L^2(\Omega)}^2 &= \|u(X) - u_{Z^N}^P(Z^N)\|_{L^2(\Omega)}^2 \\ &\quad + (e_{\text{int}, P}^{U^N, N})^2 - \sum_{k=0}^P (r_k^{U^N})^2, \end{aligned} \quad (8.23)$$

which ends the proof. ■

Inequality (8.15) is the basis of a new iterative algorithm which *adds* random variables in the approximation basis just as the *complete* Cameron-Martin's 3.3 may suggest. Suppose we perform K iterations for the i-gPC decomposition of the residue, and suppose for the transformation of interest the inequality in (8.6) is strict, then one has

$$u_{Z^N}^P(Z^N) + \sum_{k=1}^K R_{U^{N,k}, k}^P(U^{N,k}) \xrightarrow[N \rightarrow \infty]{K \rightarrow \infty} u(X).$$

If the inequality is not strict, one can still rely on convergence with respect to P .

Regarding finite numerical integration, according to property 8.2, the accuracy of the approximation is increased under some condition (see the proof) on the terms Γ_1^N and Γ_2^N , intertwining integration and truncation errors. Of course, this condition is not always ensured and the projection error $(e_{\text{int}, P}^{U^N, N})^2$ may become preponderant in comparison to $\sum_{k=2}^P (r_k^{U^N})^2$. This is the case when numerical integration accuracy is reached. Consequently, if we do not want the approximation error to potentially increase after some iteration, we once again have to *stop it before the projection error becomes preponderant* in (8.6). The main reason this part of the document does not appear amongst my list of publications comes from the fact that I did not have the time to study the intrications of the aliasing errors between integration and truncation in the previous inequality. The second reason comes from the rate of convergence with respect to the new parameter (number of variables for the approximation): (8.6) tells an improvement

may be obtained but we have no information concerning the convergence rate of the approximation. As will be presented (experimentally) in the following examples, the gain is slow with respect to K . For i-gPC, the convergence rate was also not available but at least the experimental tests on benchmarks gave very satisfying results in some cases (discontinuity). Nonetheless, I suggest presenting results on some applications in section 8.2. I had resort to arbitrary considerations for the stopping criterion. They are probably not optimal: the iterative process is stopped when the L^2 -norm at the sampled points between two iterations is greater than the L^2 of the approximation at the experimental design points. This criterion is applied when the overall polynomial order becomes greater than the number of points (see chapter 5 and section 5.2). This corresponds to a very simple heuristic criterion which gives the results obtained in the following sections and allows degenerating toward i-gPC and even gPC when the integration accuracy is not good enough.

8.2 Numerical Applications of the i-gPC decomposition of residue method

In this section, we suggest revisiting the problem of section 6.2.2 applying the new iterative approach before considering two other difficult test-cases inspired by compressible gas dynamics problems [31].

8.2.1 Some (hydrodynamically motivated) 1D test-problems

Let us begin with the problem of section 6.2.2.

Smooth Solution: Legendre Polynomial

We consider the test-case of section 6.2.2 which motivated the introduction of the new approximation due to a stagnation of the i-gPC process. We recall $X \sim \mathcal{U}([-1, 1])$ and $u(X)$ is defined by

$$u(x) = P_0^X(x) + P_3^X(x) + P_{10}^X(x), \quad (8.24)$$

where P_n^X is a one-dimensional Legendre polynomial of order n associated to the random variable X . If $P = 10$ then classical gPC, i-gPC and the new i-gPC residue based approaches are exact (up to numerical integration accuracy). In general, one does not have *a priori* estimations of the truncation order needed. In this section, we deliberately choose to underestimate the truncation order (we systematically take $P < 10$) and apply the new method.

Figure 8.1 presents the results obtained with the new i-gPC residue based approach for different truncation order, a level $k = 8$ of Clenshaw Curtis rule, i.e. $N = 257$ points, and a total of 40 successive approximations of the residue. We choose, first, to take a quite important number of points in order to have an idea of what can be asymptotically reached and remain in the conditions of property 8.1. Figure 8.1(b) presents a convergence study with respect to the number of residue iterations for several polynomial orders P . The first iteration corresponds to the gPC results, the second one to the i-gPC results and the followings to the proper iterations of the i-gPC/residue method. As expected, see theorem 8.1, the accuracy of the approximation increases with the iterations on the residue when the numerical integration accuracy is controlled. The new method allows a gain of about a decade on this test-case. The fact that numerical integration accuracy might be reached for some polynomial orders is tricky to verify experimentally. Note that the convergence curves are not anymore monotonic with respect to P : for example, $P = 6$ gives better results than $P = 8$ or $P = 9$ (parity/imparity or quality of the quadrature reached?). Figure 8.1(a)–8.1(c)–8.1(d) presents the approximations obtained from gPC, i-gPC and the i-gPC/residue approach for polynomial orders $P = 4$, $P = 5$ and $P = 6$. For the three orders P , the gPC approximations consists in the same polynomial of order $P = 3$. The i-gPC approximations are slightly different for the different orders: the iterative approach, for the different orders, does not explore exactly the same approximation spaces. Nevertheless, even with i-gPC, the improvement is small due to the fact that after few iterations of i-gPC, the basis built at each iterations are orthogonal to the residue of the solution. The i-gPC/residue approximation, on another hand, present very different results. First, note that for each polynomial orders, the i-gPC/residue approximations allows recovering the three modes of the solution. Depending on the truncation order P (at this fixed quadrature level), spurious modes of

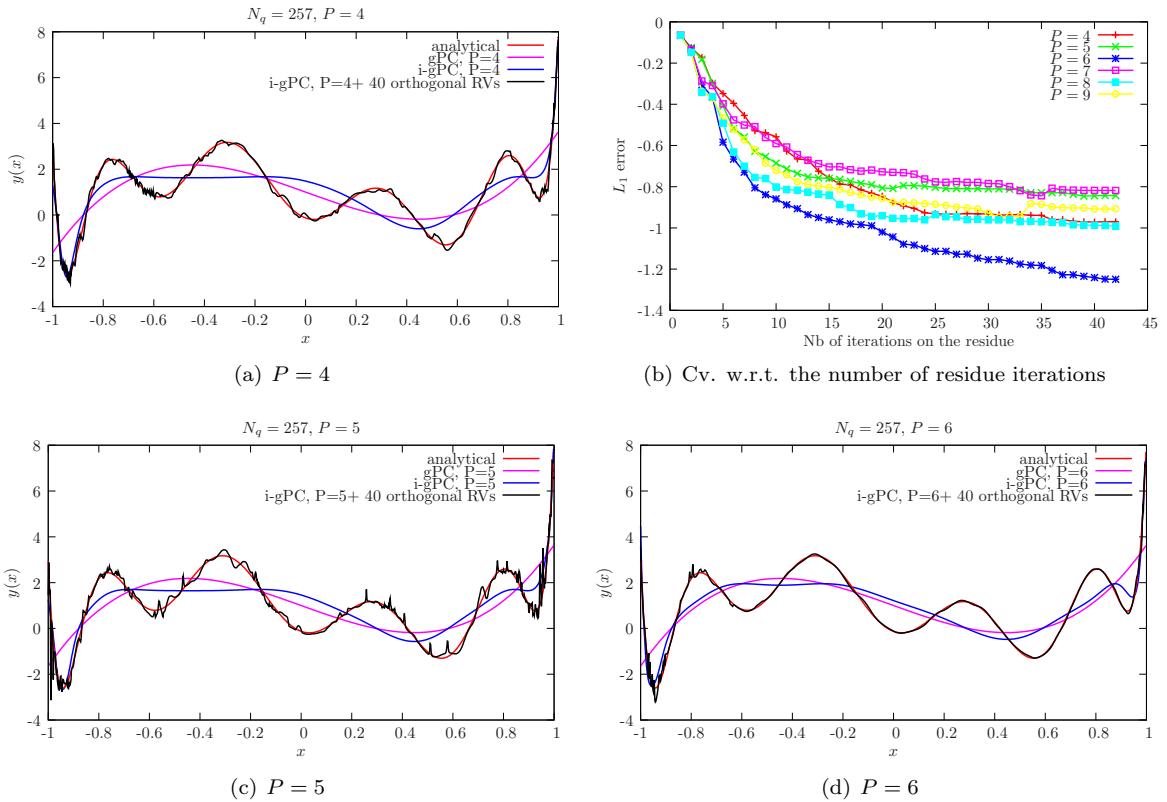
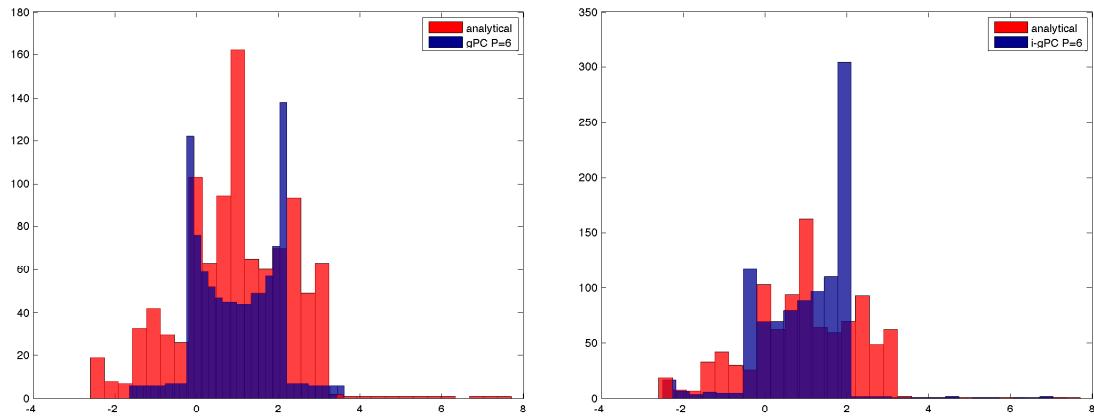
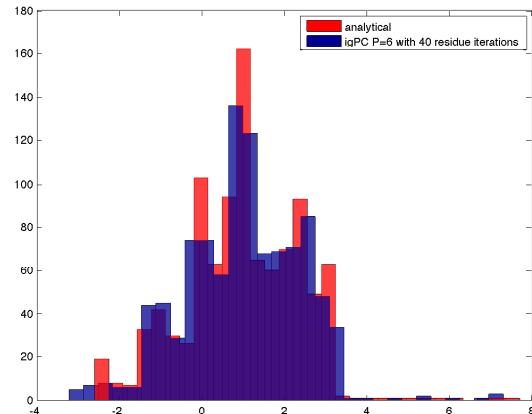


Figure 8.1: Comparisons between the results obtained from gPC, i-gPC and i-gPC/residue for $P = 4$, $P = 5$, and $P = 6$ on problem (8.24). The top right picture shows convergence curves with respect to the number of residue iterations.



(a) Analytical and gPC approximation

(b) Analytical and i-gPC approximation



(c) Analytical and i-gPC/residue approximation

Figure 8.2: Histograms of the pdfs of the gPC, i-gPC and i-gPC/residue approximations with $P = 6$ for problem (8.24).

different amplitudes are introduced, see 8.1(a) and 8.1(c). Figure 8.2 presents histograms of the pdfs of the gPC, i-gPC and i-gPC/residue approximations for $P = 6$. On this test-case, gPC and i-gPC both leads to approximations of the histograms which could lead to bad interpretations of the results: for both representations, the statistic is not recovered. The new method, on the contrary, even if generating spurious modes, allows recovering the statistical properties of the solution with an interesting agreement.

Same example with a smaller experimental design: $N_q = 65$

Figure 8.3 presents the same results as previously but in a more realistic context in the sense we take only 65 integration points. Figure 8.3(a) and 8.3(b) presents the comparison between the gPC, i-gPC and i-gPC/residue approximations for $P = 5$ and $P = 6$. The gPC and i-gPC approximations are only slightly affected by the loss in the numerical integration accuracy: this is due to the fact that for both representations, the truncation error remains preponderant with respect to the projection error. On the contrary, the i-gPC/residue approach is affected by this same loss of accuracy: for example, for $P = 6$, more spurious modes appear in the low resolution approximation of picture 8.3(b) than on picture 8.1(d). The i-gPC/residue approximation on this test-problem may allow tackling the projection error. Figure 8.3(c)–8.3(d)–8.3(e) present the histograms of the pdfs of the three approximations for $P = 6$. Note that even if the i-gPC/residue approximation with $N_q = 65$ points introduces more spurious modes than the one with $N_q = 257$, the approximation allows a good statistical interpretation.

First test-case inspired by compressible gas dynamics

Let us consider a new test-case inspired by what can be expected solving uncertainty propagation problems in compressible gas dynamics flows⁸, see [31]. The considered test-function consists in a constant state and an affine state both separated by a discontinuity:

$$u(x) = \begin{cases} 1 & \text{if } x \leq 0, \\ ax & \text{if } x > 0. \end{cases} \quad (8.25)$$

In this section, we take $a = -0.3$. The test-case is difficult in the sense the solution is not anymore piecewise constant but piecewise polynomial. In fact, the solution exhibits a mixed behaviour with a discrete part and a uniform one. One simple way to represent the solution consists in writing $u(X) = p_0\mathcal{U}_{[0,1]} + p_1\delta_{u=1}$. In the previous expression, $p_0 + p_1 = 1$, and $p_0 = p_1 = \frac{1}{2}$. They are the probability for the solution to be represented by a uniform random variable or a Dirac in 1. Figure 8.4 presents comparisons between the analytical solution and the gPC, i-gPC and i-gPC/residue approximations for different truncation orders P . Figure 8.4(b) shows a convergence study of the i-gPC/residue method with respect to the number of residue iterations. The first iteration corresponds to the gPC results, the second one to the i-gPC results and the following ones to the proper iterations of the i-gPC/residue method. Once again, the different iterations/applications of i-gPC on the residue of the solution ensure a gain in accuracy up to a stagnation occurring at the 12th iterations for almost every polynomial orders P . Almost one decade is gained with respect to gPC and i-gPC applying this i-gPC/residue approach. The convergence with respect to P of the new approach is not monotonic. Figure 8.4(a)–8.4(c)–8.4(d) compares the different methods for truncation orders $P = 4$, $P = 5$ and $P = 6$. For the three orders, gPC leads to very oscillating approximations and poorly represents the solution, the discontinuity is not captured neither the affine part. The i-gPC approximations allows capturing the discontinuous counterpart of the solution but miss the affine part. See figure 8.4(c) for example, the i-gPC approximations behaves as a piecewise constant approximation rather than a piecewise polynomial one. The i-gPC/residue approach now allows capturing both counterparts of the solution, the discontinuity, mainly captured by application of i-gPC in the early steps, and the affine one during the other iterations. The new approach introduces once again spurious modes but we insist on the fact that these spurious modes do not prevent from having a good statistical interpretation of the solution. This is emphasized in figure 8.5 where we consider the histograms of the pdfs of the approximations for $P = 4$. Figure 8.5(a) compares the histograms of the pdfs for the analytical solution together with the one obtained from the gPC approximation. The statistics of the solution is missed by the approximation method. Figure 8.5(b) compares the histograms

⁸In this section, we do not solve Euler system, we rely on built function mimicing complex behaviours in compressible gas dynamics flows, see [31].

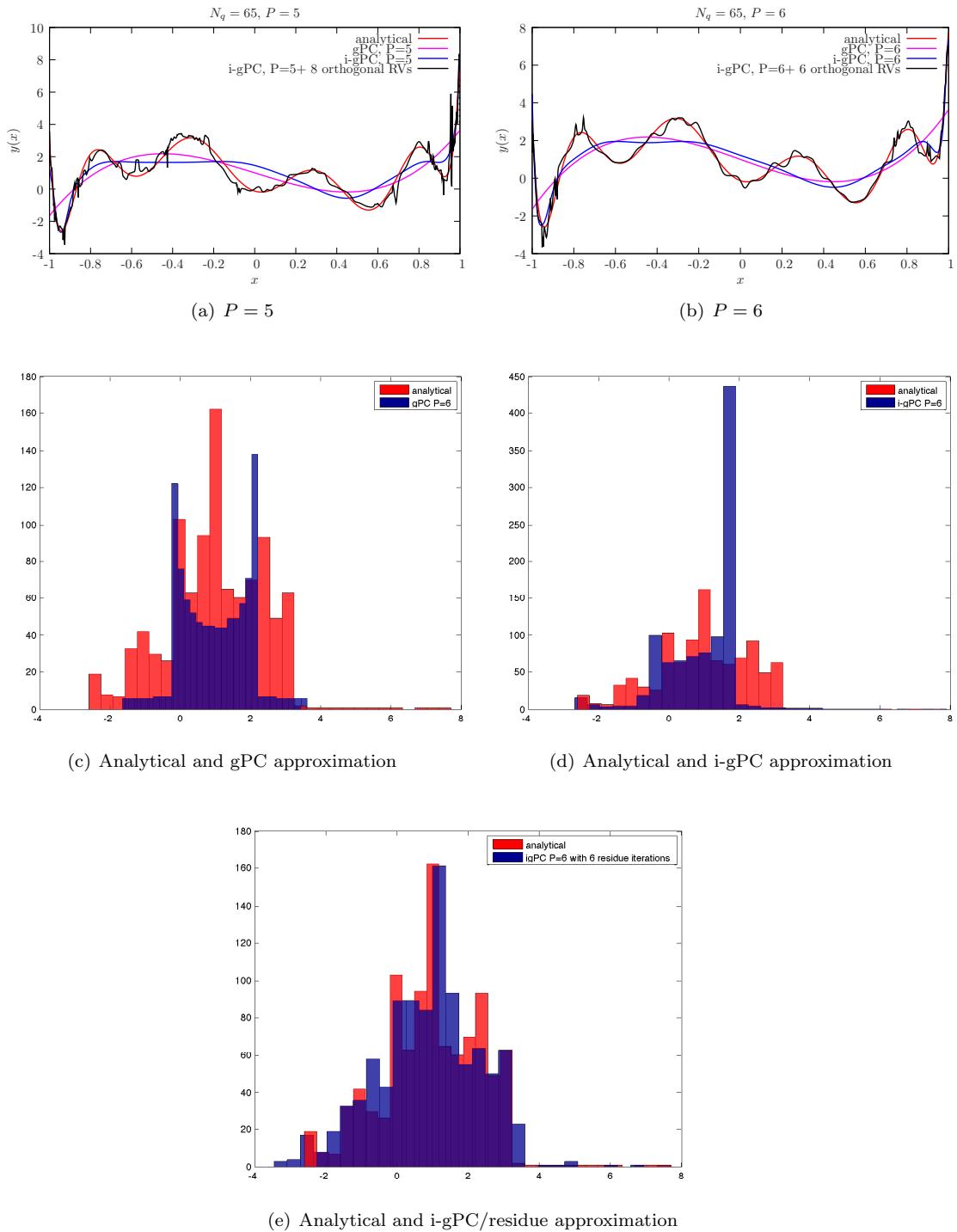


Figure 8.3: Stochastic representations and histograms of the pdfs of the solution in low resolution context ($N_q = 65$) for problem (8.24).

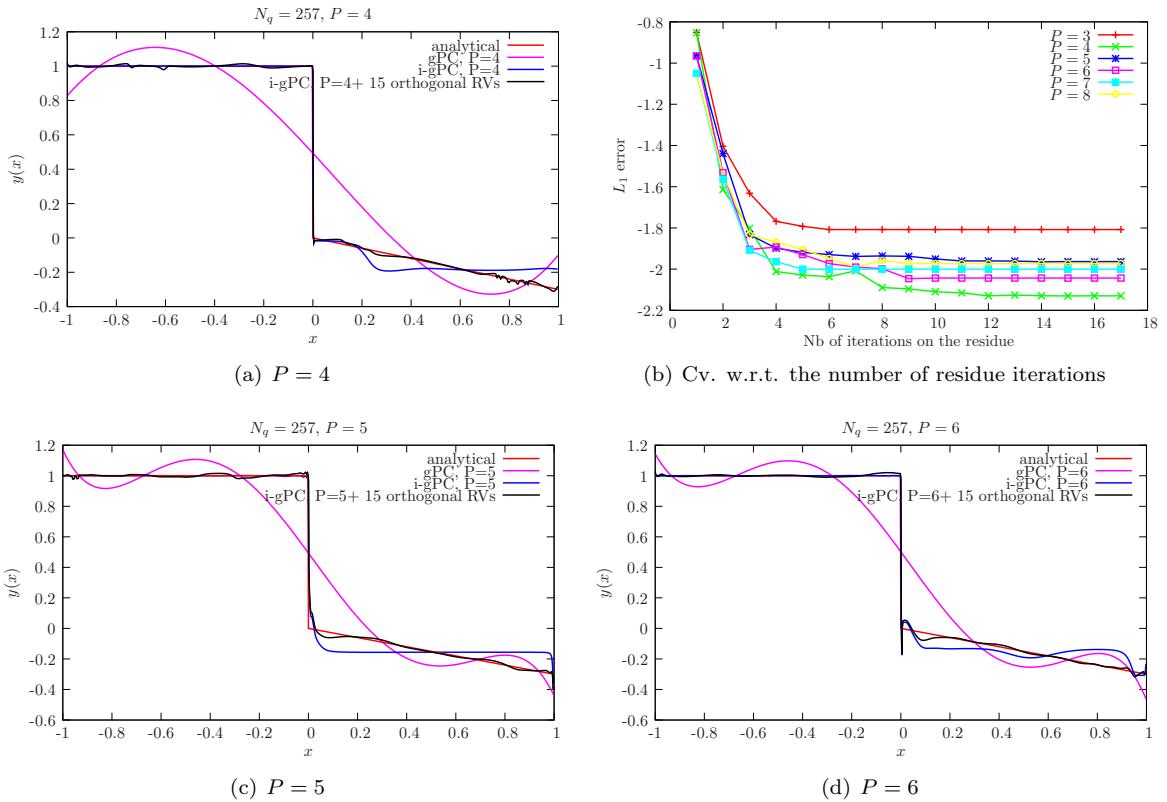


Figure 8.4: Comparisons between the results obtained from gPC, i-gPC and i-gPC/residue for $P = 4$, $P = 5$, and $P = 6$ on problem (8.25). The top right picture shows convergence curves with respect to the number of residue iterations.

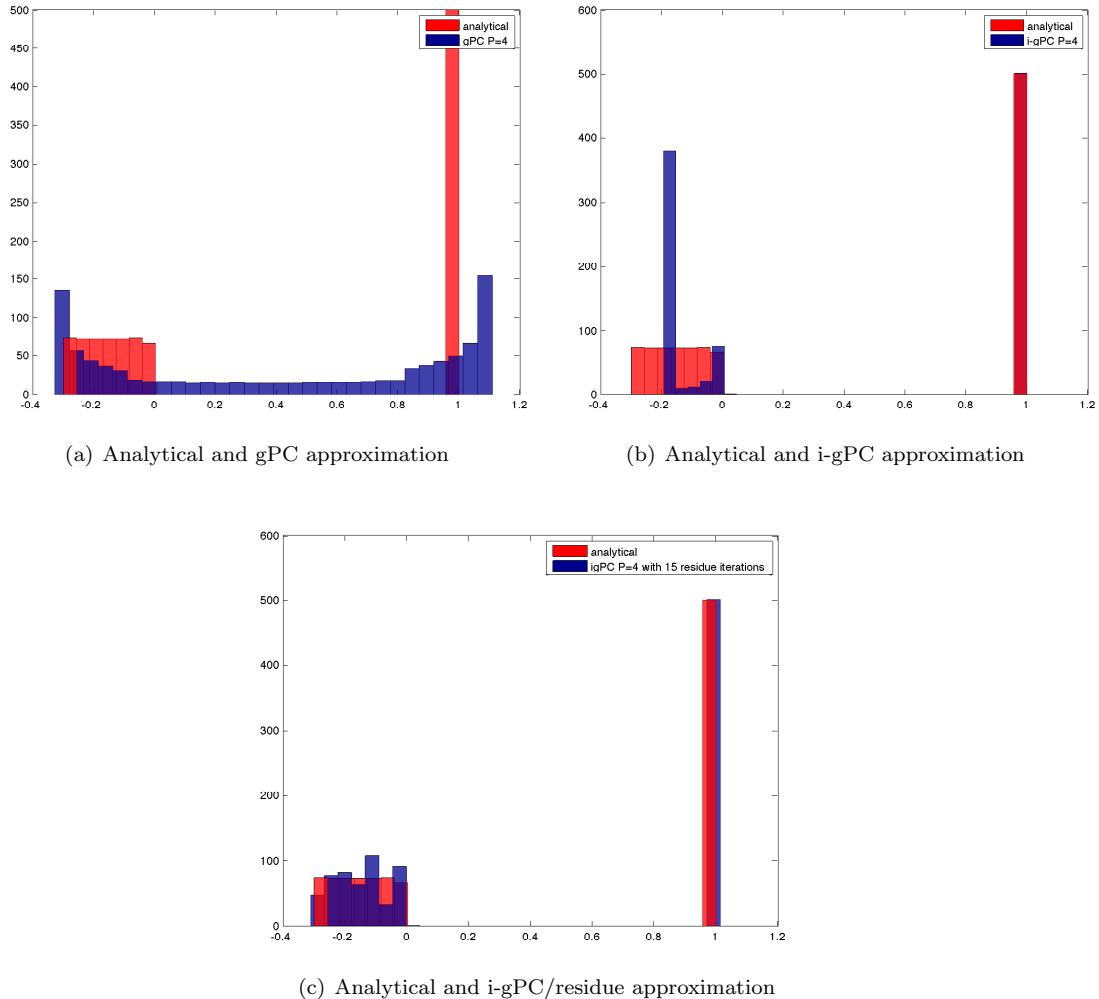


Figure 8.5: Histograms of the pdfs of the gPC, i-gPC and i-gPC/residue approximations with $P = 5$ for problem (8.25).

of the pdfs for the analytical solution together with the one obtained from the i-gPC approximation. The method allows capturing the discontinuity and the probability $p_1 = \frac{1}{2}$ of the event $u = 1$ with a good agreement (Dirac mass at $u = 1$). On the other hand, the uniform part of the random variable is not well represented. Consider the i-gPC/residue approximation, figure 8.5(c) shows that both the uniform part and the discrete one are well estimated with the good probabilities despite the introduction of spurious modes.

Same example with a smaller experimental design: $N_q = 65$

Figure 8.6 presents the same results as previously but in a more realistic context in the sense we take only 65 integration points. Figure 8.6(a) and 8.6(b) presents the comparison between the gPC, i-gPC and i-gPC/residue approximations for $P = 5$ and $P = 6$. The gPC approximations are only slightly affected by the loss in the numerical integration accuracy. This is due to the fact that for these representations, the truncation error remains preponderant with respect to the projection error. On the contrary, the i-gPC and i-gPC/residue approaches are affected by this same loss of accuracy. Both representations for both polynomial orders have more spurious modes on pictures 8.6(a)–8.6(b) in comparison to pictures 8.4(a)–8.4(d). Once again, by considering figures 8.6(c)–8.6(d)–8.6(e) we aim at emphasizing that even with the appearance of spurious modes for the new i-gPC/residue approach, the histograms of the pdfs of the approximations allow a good statistical interpretation.

Second test-case inspired by compressible gas dynamics

We once again consider a test-case inspired by what can be expected solving uncertainty propagation problems in compressible gas dynamics flows (Euler system), see [31]. The expression of the test-function is as follows:

$$u(x) = \begin{cases} u_0 & \text{if } x \leq x_0, \\ u_1 & \text{if } x_0 < x \leq x_1, \\ u_1 + (x - x_1)^2 & \text{if } x_1 < x. \end{cases} \quad (8.26)$$

In this section, we choose $u_0 = 0$, $u_1 = 1$, $x_0 = -0.6$ and $x_1 = -0.3$. In term of Euler system, this test-case can be interpreted as a shock interacting with a rarefaction fan, see [31]. In term of random variable, $u(X)$ can be viewed as the sum of a discrete random variable $\delta_{u=u_0=0}$ with probability p_0 and $\delta_{u=u_1=1}$ with probability p_1 and another continuous random variable which we denote by \mathcal{C} with probability p_2 so that

$$u(X) = p_0\delta_{u=u_0} + p_1\delta_{u=u_1} + p_2\mathcal{C}.$$

The data of the problem are such that $p_0 = 0.2$, $p_1 = 0.15$ and $p_2 = 0.65$.

Figure 8.7 presents the results obtained with the new i-gPC residue based approach for different truncation order, a level $k = 8$ of Clenshaw Curtis rule, i.e. $N = 257$ points, and a total of 50 successive approximations of the residue. Figure 8.7 presents comparisons between the analytical solution and the gPC, i-gPC and i-gPC/residue approximations for different truncation orders P . Figure 8.7(b) shows a convergence study of the i-gPC/residue method with respect to the number of residue iterations. The first iteration corresponds to the gPC results, the second one to the i-gPC results and the followings to the proper iterations of the i-gPC/residue method. Once again, the different iterations/applications of i-gPC on the residue of the solution ensure a gain in accuracy up to a stagnation occurring at different iterations depending on the polynomial order P . Once again, almost one decade is gained with respect to gPC and i-gPC applying this i-gPC/residue approach and the convergence with respect to P of the new approach is not monotonic. Figure 8.7(a)–8.7(c)–8.7(d) compares the different methods for truncation orders $P = 4$, $P = 7$ and $P = 8$. For the three orders, gPC leads to very oscillating approximations and poorly represents the solution, the discontinuity is not captured neither the continuous part. The i-gPC approximations allows capturing the discontinuous counterpart up to a certain polynomial order. For example, for $P = 4$, on figure 8.7(a), i-gPC is not much more accurate than gPC whereas for $P = 7$ and $P = 8$ on figure 8.7(c)–8.7(d), i-gPC allows capturing the discontinuous state. On the other hand, i-gPC does capture accurately the continuous part of the random variable for $P = 4$ whereas it does less accurately for high orders $P = 7$ and $P = 8$ (appearance of small oscillations in the continuous part of the curves on figures 8.7(c)–8.7(d)). The i-gPC/residue approach now allows capturing both

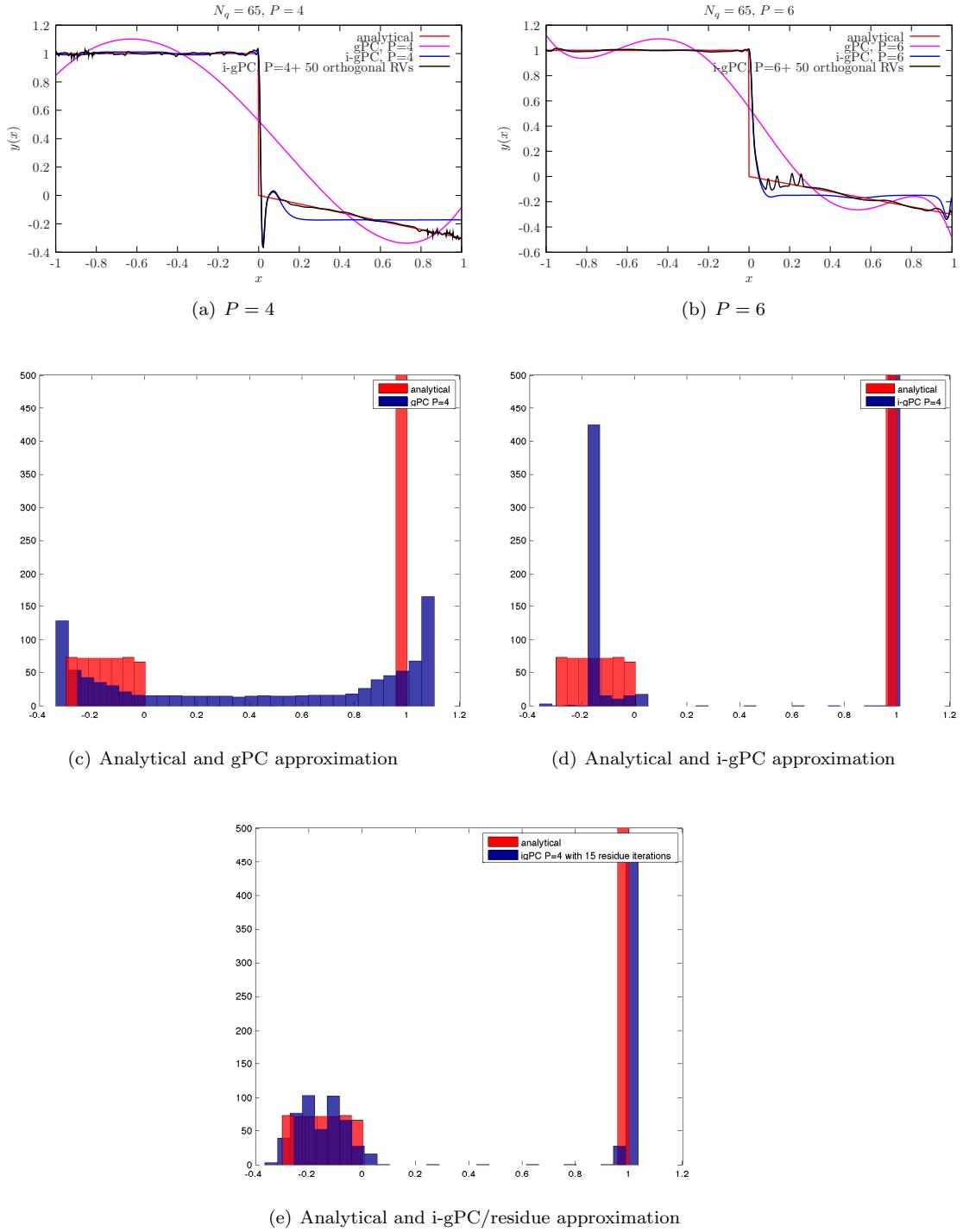


Figure 8.6: Stochastic representations and histograms of the pdfs of the solution in low resolution context ($N_q = 65$) for problem (8.25).

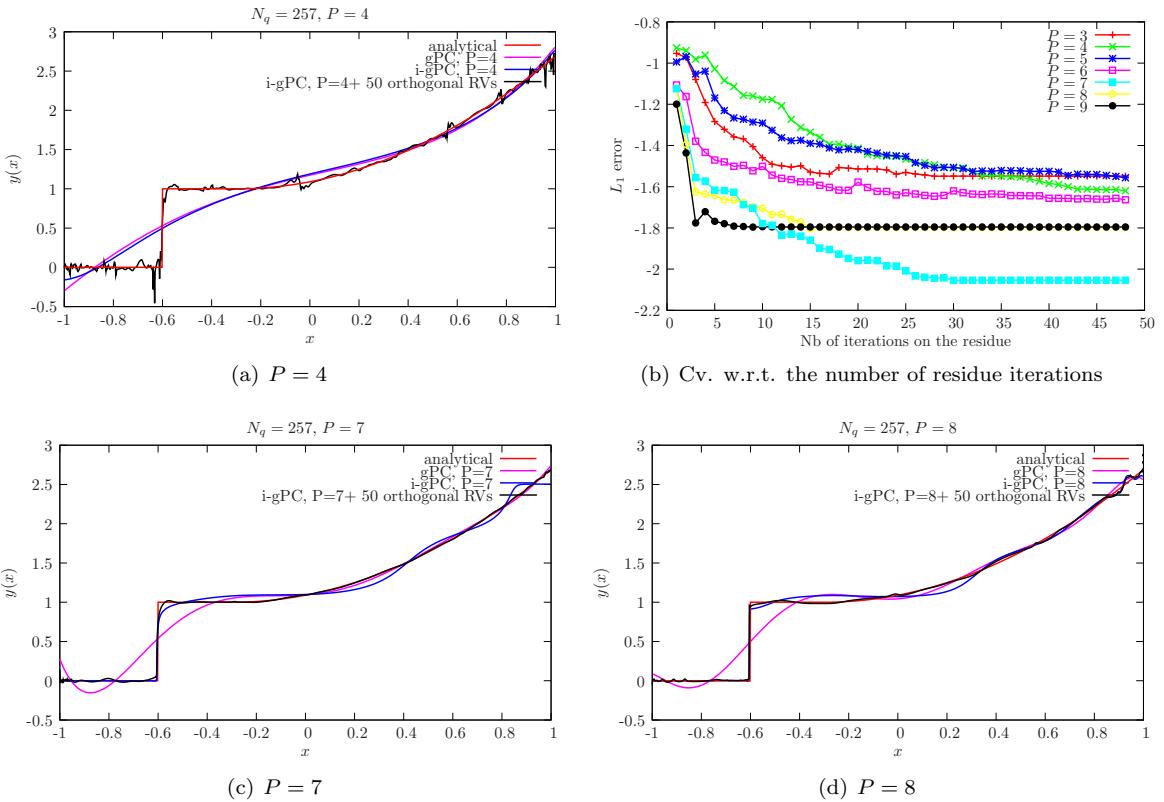
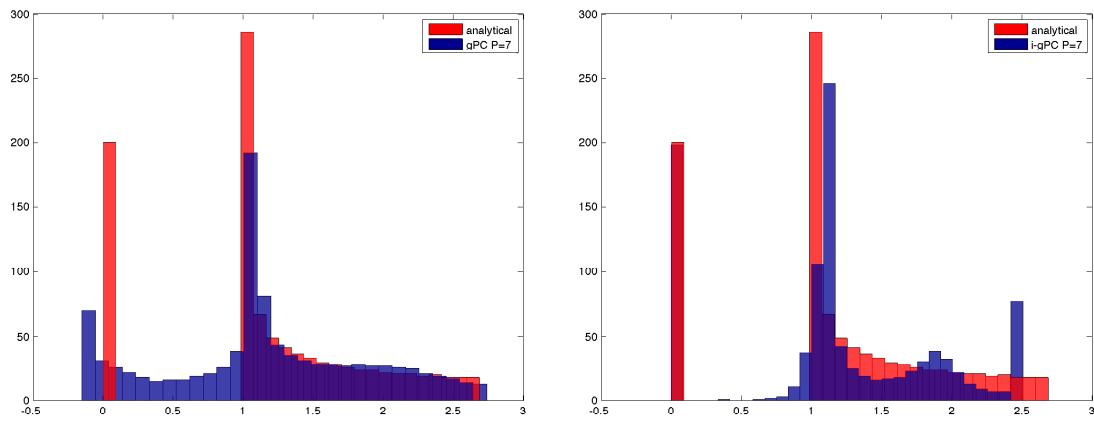
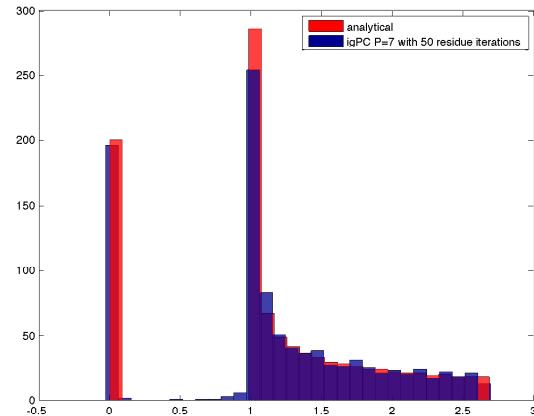


Figure 8.7: Comparisons between the results obtained from gPC, i-gPC and i-gPC/residue for $P = 4$, $P = 5$, and $P = 6$ for problem (8.26). The top right picture shows convergence curves with respect to the number of residue iterations.



(a) Analytical and gPC approximation

(b) Analytical and i-gPC approximation



(c) Analytical and i-gPC/residue approximation

Figure 8.8: Histograms of the pdfs of the gPC, i-gPC and i-gPC/residue approximations with $P = 7$ for problem (8.26).

counterparts of the solution for every orders $P = 4$, $P = 7$ and $P = 8$, see figures 8.7(a)–8.7(c)–8.7(d), despite the appearance of spurious modes for low order $P = 4$ on figure 8.7(a). Figure 8.8 compares the histograms of the pdfs for the analytical solution together with the ones obtained from the different approximation. Figure 8.8(a) compares the histograms of the pdfs for the analytical solution together with the one obtained from the gPC approximation. The method does not allow capturing the discontinuity and the probability $p_0 = 0.2$ of the event $u = u_0 = 0$. On the other hand, the continuous part of the random variable is quite well represented. The i-gPC representation for this test-case ensures the discrete part of the random variable is recovered accurately ($p_0 = 0.2$ for $u = u_0 = 0$). On the other hand, the continuous part of the random variable is not well resolved, see figure 8.8(b). Consider the i-gPC/residue approximation, figure 8.8(c) shows that both the continuous part and the discrete one are well estimated with the good probabilities despite the introduction of spurious modes.

Same example with a smaller experimental design: $N_q = 65$

Figure 8.9 presents the same results as previously but in a more realistic context in the sense we take only 65 integration points. Figure 8.9(a) and 8.9(b) presents the comparison between the gPC, i-gPC and i-gPC/residue approximations for $P = 7$ and $P = 8$. Once again, the gPC approximations are only slightly affected by the loss in the numerical integration accuracy. This is due to the fact that for these representations, the truncation error remains preponderant with respect to the projection error. On the contrary, the i-gPC and i-gPC/residue approaches are affected by this same loss of accuracy. Both representations for both polynomial orders slightly miss the discontinuity position. For this test case, even if the number of points is less important than previously, the representations do not exhibit more important spurious modes pictures 8.9(a)–8.9(b) than on pictures 8.7(c)–8.7(d). Figures 8.9(c)–8.9(d)–8.9(e) present the histograms of the pdfs of the three approximations for $P = 7$. On figure 8.9(c), the gPC approximation allows a good agreement for the continuous part of the random variable. On the other hand, the discrete counterpart with the two Dirac masses $\delta_{u=u_0}$ and $\delta_{u=u_1}$ is missed. On the contrary, in the same conditions, the i-gPC approximation allows a good representation of the discrete part of the random variable but misses the continuous one. The i-gPC/residue approach allows taking advantage of both representations with an interesting agreement on both parts of the random variable.

8.2.2 Some (well-known in the literature) multidimensional test-cases

In this section, we apply the i-gPC/residue approach to multidimensional test-cases, the g-function of Sobol which is a well-known benchmark in sensitivity analysis, see [36, 254].

This function is in fact very similar to the C^0 CONTINUOUS integrand of the Genz package of testing functions [120]. The model is

$$Y = \prod_{i=1}^Q \frac{|X_i| + a_i}{1 + a_i}, \quad \text{with } \forall i \in \{1, \dots, D\}, a_i \geq 0, \quad (8.27)$$

where the $(X_i)_{i \in \{1, \dots, D\}}$ are independent identically distributed uniform random variables on $[-2, 2]$ (as in [254]). For this model, the mean and variance can be determined exactly

$$\begin{cases} \bar{Y} &= 1, \\ \text{Var}(Y) &= \prod_{i=1}^Q \left(\frac{1}{3(1+a_i)^2} \right) - 1. \end{cases} \quad (8.28)$$

Parameter a_i controls the stiffness of the model. Lower values of a_i have the tendency to increase the discontinuity jump of the absolute value derivative profile, while higher values tend to decrease it.

Sobol in 2 D

We choose first a $Q = 2$ dimension Sobol' function and we test a stiff isotropic problem by taking $\vec{a} = (a_1, a_2) = (0, 0)$. Figure 8.10 presents the results on the 2-D Sobol g-funtion. Figure 8.10(b) shows the convergence curves with respect to the number of iterations on the residue. The first point

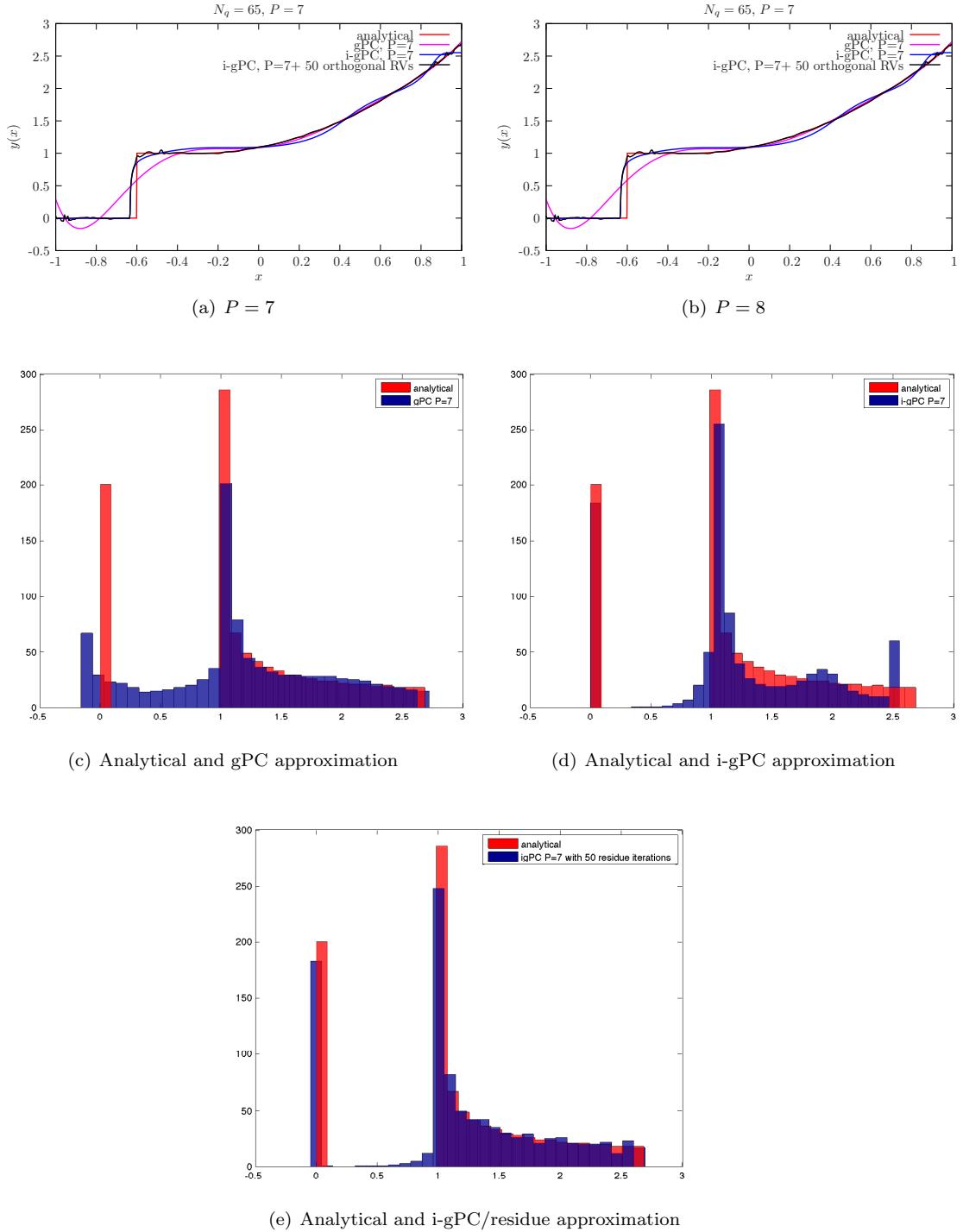


Figure 8.9: Stochastic representations and histograms of the pdfs of the solution in low resolution context ($N_q = 65$) for problem (8.26).

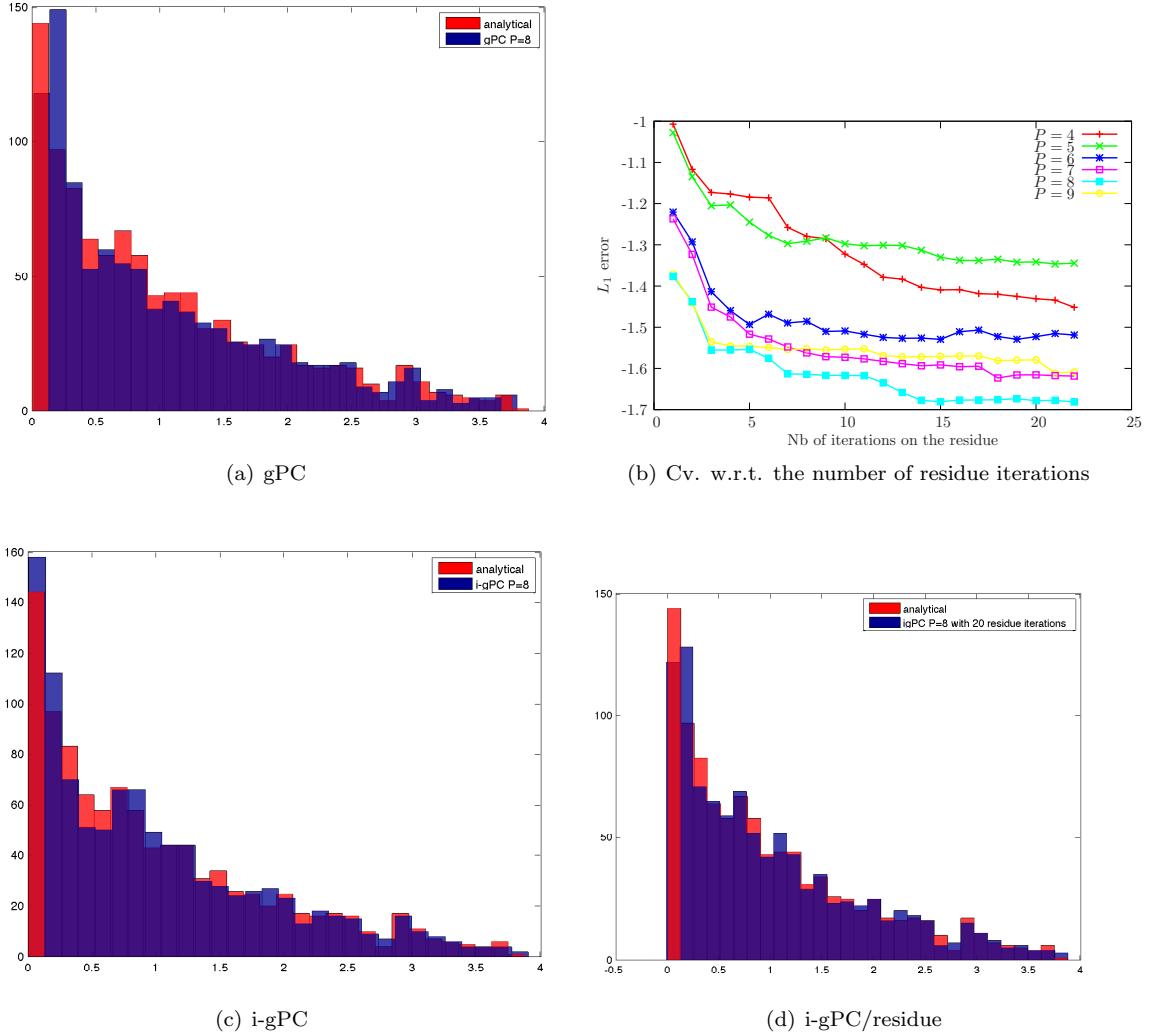


Figure 8.10: Comparisons between the results obtained from gPC, i-gPC and i-gPC/residue for $P = 8$ for problem (8.28) in 2-D. The top right picture shows convergence curves with respect to the number of residue iterations.

corresponds to the application of gPC, the second one to the application of i-gPC and the following ones to the different iterations on the residue. The method ensures a gain with respect to both gPC and i-gPC, even if this gain is less important than on the other test problems. Figure 8.10(a)–8.10(c)–8.10(d) compares the histograms of the pdfs of the approximations to the analytical one. The gain is visible from applications of gPC, i-gPC and the i-gPC/residue methods.

Sobol in 5 D

We choose a $Q = 5$ dimension Sobol' function and we test an isotropic problem with $\vec{a} = (a_1, a_2, a_3, a_4, a_5) = (0.5, 0.5, 0.5, 0.5, 0.5)$ so that every stochastic dimensions have the same importance. Figure 8.11 presents

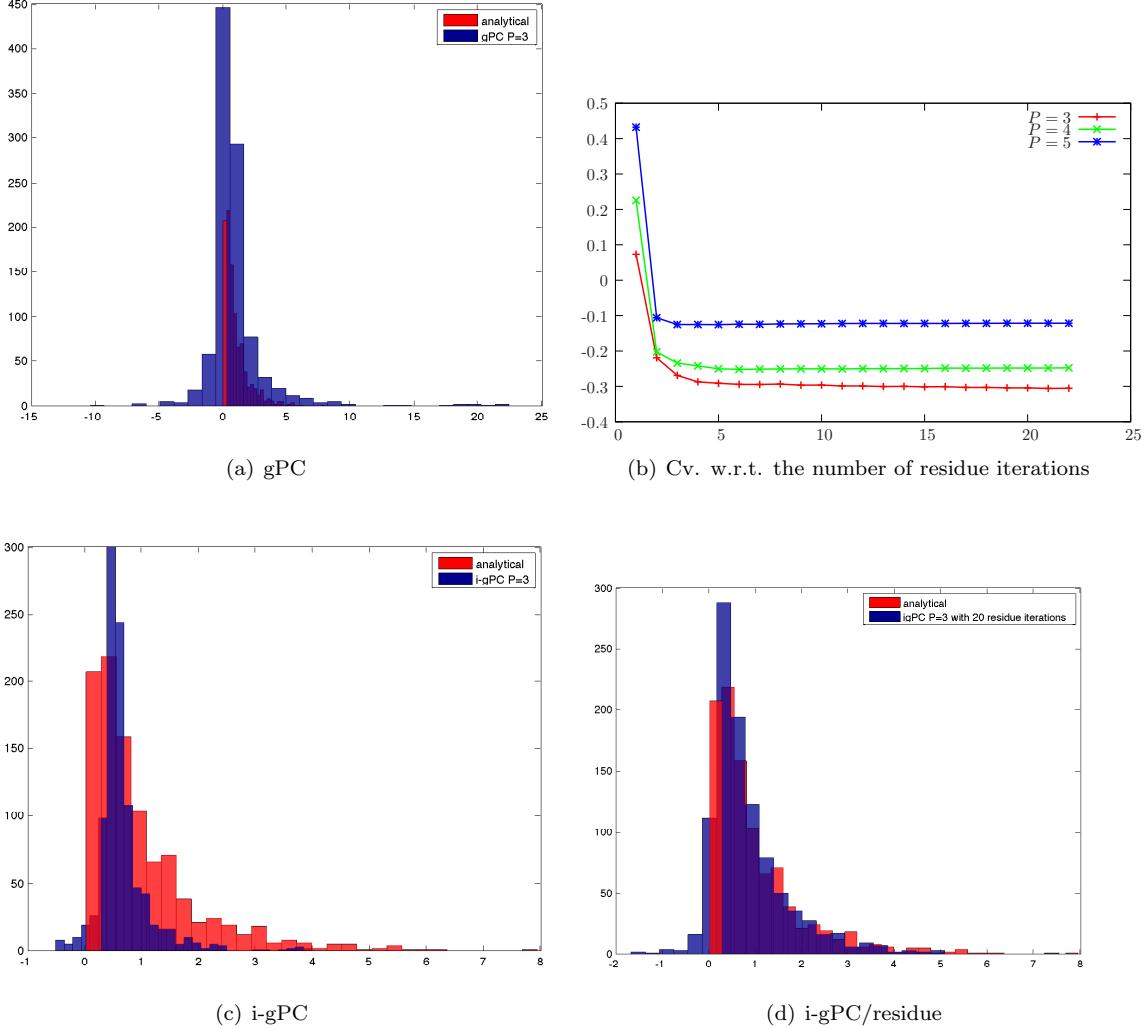


Figure 8.11: Comparisons between the results obtained from gPC, i-gPC and i-gPC/residue for $P = 3$ for problem (8.28) in 5-D. The top right picture shows convergence curves with respect to the number of residue iterations.

the results on the 5-D Sobol g-funtion. Figure 8.11(b) shows the convergence curves with respect to the number of iterations on the residue. The first point corresponds to the application of gPC, the second one to the application of i-gPC and the followings to the different iterations on the residue. The method ensures a gain with respect to both gPC and i-gPC. Figure 8.11(a)–8.11(c)–8.11(d) compares the histograms of the pdfs of the approximations to the analytical one. The gain is visible from applications of gPC, i-gPC and the i-gPC/residue methods. On this test-case, both gPC and i-gPC do not allow

recovering the statistics of the solution. On the other hand, the i-gPC/residue approach does.

8.3 Summary for the i-gPC decomposition of the residue algorithm

In this chapter, we wanted to highlight the fact that for uncertainty quantification applications, Cameron-Martin's theorem has probably not been fully taken advantage of in term of approximation algorithm. We wanted to insist on the fact that some parts of it are not invoked through the gPC algorithm presented in the literature.

The i-gPC decomposition of the residue algorithm presented in this chapter is one way to put forward the above fact but it may not represent a new *viable* alternative to gPC or i-gPC. As can be seen experimentally in the previous applications, the convergence rate of the i-gPC/residue approximations remains relatively slow with respect to the increase of the number of degree of freedom (number of residue approximations). For small but realistic experimental designs N , the stopping criterion is activated immediately after few iterations and the i-gPC/residue method degenerates toward i-gPC. The quality of the approximation is not really improved with respect to i-gPC. This *testifies that the regime of interest is probably not yet well identified hence not captured*. This is mainly why this chapter has not been submitted for publication and why the term *attempt* is emphasized in its title. Nonetheless, we wanted to document this attempt as its elaboration (which is probably a dead end) made us realize and understand some aspects of both publications [295] and [55]. For example, the relation between Cameron-Martin's theorem and Wiener's homogeneous Chaos is not straightforward as the ergodicity property omnipresent in Wiener's paper is not explicit in Cameron-Martin's one. This ergodicity notion may be recovered in [55] through the summation over a set of increasing number of random variables. It may echo the possibility to average in the direction of increasing random variables Q rather than in the direction of increasing polynomial orders P . I insist here that regarding ergodicity, I may have made wrong statements along the document. I am not expert with this notion but I am more and more interested by this property and aware of what can be cast behind it.

Part III

Monte-Carlo schemes for the (non)linear Boltzmann equation

Chapter 9

Monte-Carlo methods for the linear Boltzmann equation

An old topic (once again...) revisited

Contents

9.1	General Methodology for the construction of an MC scheme	164
9.2	The analog (Adjoint) MC scheme (mimics physics)	165
9.2.1	Expectation form over the analog set of random variables	165
9.2.2	Construction of the analog MC scheme	167
9.3	The semi-analog (Adjoint) MC scheme (implicit capture)	170
9.3.1	Expectation form over the semi-analog set of random variables	170
9.3.2	Construction of the semi-analog MC scheme	171
9.4	The non-analog (Adjoint) MC scheme	173
9.4.1	Expectation form over the non-analog set of random variables	173
9.4.2	Construction of the Adjoint non-analog MC scheme	174
9.5	Direct formulation and direct set of random variables	176
9.5.1	Adjoint and direct formulations of the same transport equation	176
9.5.2	Direct Integral formulation for the non-analog scheme	177
9.5.3	Construction of the direct non-analog MC scheme	177
9.6	Common approximations to simplify the samplings and resolutions	180
9.6.1	The interaction time τ of probability measure $f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds$	180
9.6.2	The energy and angle correlated samplings $\mathbf{V}' = V'W'$	184
9.6.3	The modification of the weight of the particle $w_p(t)$	186
9.7	Variance and moments of the MC schemes	187
9.7.1	Asymptotic variance of the analog scheme (full_analog and multiplicity)	187
9.7.2	Asymptotic variance of the semi-analog scheme	190
9.7.3	Asymptotic variance of the non-analog scheme	191
9.7.4	Comparisons of the standard deviations of the MC schemes (homogeneous)	192
9.8	A general canvas for developing MC schemes	195
9.8.1	Sampling the initial MC particle population	195
9.8.2	A general skeleton in order to develop each scheme in the same platform	203
9.9	Taking into account a source term	206
9.9.1	Application of Duhammel's principle: source sampling (direct)	207
9.9.2	Quasi-Static method for the transport equation with source term	209
9.10	Taking into account an acceleration term in MC resolution schemes	212
9.10.1	An MC resolution with curved trajectories in the comobile frame	213

9.10.2	An MC resolution with straight trajectories in a new frame	214
9.11	The Uncertain Linear Boltzmann equation	221
9.11.1	Non-intrusive resolution of the uncertain linear Boltzmann equation	222
9.11.2	A gPC-intrusive Monte-Carlo scheme for the uncertain linear Boltzmann equation	224
9.11.3	Summary	229
9.12	Application of gPC for MC accelerations for the linear Boltzmann equation	230
9.12.1	Variance reduction, AP scheme, same problems, different denominations	233
9.12.2	Application of gPC to accelerate MC integration	233
9.12.3	Acceleration by gPC of the MC resolution of the linear Boltzmann equation	246
9.12.4	Summary	252

The linear Boltzmann equation corresponds to limit (1.30) of the Boltzmann equation (1.1). It models the behaviour of particles in a background collisional media. Its linear aspect neglects particle-particle interactions and assumes only particle-matter collisions, matter being unaffected by the particle flow. It is particularly relevant for physical particles of weights way smaller than the constituents of the matter (i.e. neutrons versus (big) atoms, photons versus electrons etc.), see chapter 1. This class of PDE is very efficient in order to describe the mean behaviour of an important number of particles¹. The unknown of the equation is the density of particles $u(\mathbf{x}, t, \mathbf{v}) \geq 0$ having position $\mathbf{x} \in \mathcal{D} \subset \mathbb{R}^3$, at time $t \in [0, T]$ and velocity $\mathbf{v} \in \mathbb{R}^3$. Such models can be used for neutronics applications [268, 173], photonics [203, 59], plasma physics [29, 24, 25], population dynamics [223] and so on. At some stage of the discussion, we may consider $u(\mathbf{x}, t, \mathbf{v}) = u(\mathbf{x}, t, v, \omega)$ depending on energy² $|\mathbf{v}| = v \in \mathbb{R}^+$ and with angle $\frac{\mathbf{v}}{v} = \omega \in \mathbb{S}^2$.

Many *deterministic* resolution schemes are available for this equation, such as Kinetic ones [27], P_n [281, 196, 136, 141, 116], M_n [93, 226, 47, 137, 94, 133, 228, 279, 214, 227, 8], S_n [16, 92, 61, 116, 56]. In this part of the document, the unknown u depending on $3(\mathbf{x}) + 1(t) + 3(\mathbf{v}) = 7$ variables, we consider the resolution of the linear Boltzmann equation to be a high dimensional problem. We consequently focus on Monte-Carlo resolution schemes. The direct consequence of such a choice is (see [165]) mainly having a numerical method whose convergence rate is independent of the regularity of the solution and of the number of variables but relatively slow, $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}}}\right)$ where N_{MC} denotes the number of simulated MC particles (the term *MC particle* will be defined later in the document).

The aim of this chapter is to present the general construction of an MC scheme for the resolution of the linear Boltzmann equation. We explain how the description of an MC scheme resumes to identifying several samplings. **The way the MC schemes are built ensures their convergence according to theorem 3.2.1 of [165] for any linear Boltzmann equations** (inductive reasoning and verification). In this chapter we present the most common MC schemes and even suggest some original ones (see section 9.6). We detail the construction of

- the *non-analog* scheme, used mainly in photonic applications,
- the *semi-analog*³ one, intensively used for neutronics applications,
- and the *analog*⁴ scheme, applied mainly in order to estimate probabilities of extinction of a population of particles (small number of physical particles).

Every schemes aim at (at least) approximating the density of particle $u(\mathbf{x}, t, \mathbf{v})$, $\forall \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^3, t \in [0, T], \mathbf{v} \in \mathbb{R}^3$, solution of the linear Boltzmann equation:

$$\partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \cdot \nabla_x u(\mathbf{x}, t, \mathbf{v}) + v \sigma_t(\mathbf{x}, t, \mathbf{v}) u(\mathbf{x}, t, \mathbf{v}) = \int v \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}' + S(\mathbf{x}, t, \mathbf{v}). \quad (9.1)$$

¹The model becomes irrelevant for modeling a small number of physical particles, see [285, 200, 18, 52] for examples in neutronics.

²Depending on the physics of interest, v may not exactly refer to energy but all along the document, v will be related to the energy of the particles *via* $\frac{1}{2}mv^2$ where v is the velocity for example for neutrons or *via* hv where v is the frequency for photons.

³also referred as *implicit capture scheme*.

⁴known to *mimic* the physical behaviour of the particles.

The quantities $(\sigma_\alpha)_{\alpha \in \{s,t\}}$ are called the cross-sections⁵ in this document. The total cross-section σ_t describes the collision rate of the particles in the media whereas the scattering cross-sections describes its relative absorption, diffusion or multiplicative rate⁶. When there is no ambiguity, the more concise notation $\sigma_t(\mathbf{x}, t, \cdot) = \sigma_t(\mathbf{x}, t, \mathbf{v})$ and $\sigma_s(\mathbf{x}, t, \cdot) = \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}')$ may be used in the following chapters. At some point in the document, the scattering cross-section may be expanded into a sum over a certain number of reactions N_R :

$$\begin{aligned}\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') &= \sum_{r=0}^{N_R} \nu_r(v) \sigma_r(\mathbf{x}, t, \mathbf{v}, \mathbf{v}'), \\ \sigma_t(\mathbf{x}, t, \mathbf{v}) &= \sum_{r=0}^{N_R} \int \sigma_r(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') d\mathbf{v}' = \sum_{r=0}^{N_R} \sigma_r(\mathbf{x}, t, \mathbf{v}).\end{aligned}\tag{9.2}$$

In (9.2), each $(\sigma_r)_{r \in \{0, \dots, N_R\}}$ describes the rate of occurrence of reaction $r \in \{0, \dots, N_R\}$. The coefficients $(\nu_r(v))_{r \in \{0, \dots, N_R\}}$ are the multiplicities of the corresponding reactions: $\forall r \in \{0, \dots, N_R\}$, $\nu_r(v) \in \mathbb{R}^+$ but in general $\nu_r(v) \in \mathbb{N}$. We may also introduce the absorption cross-section σ_0 , defined by the reaction of multiplicity $\nu_0 = 0$. Note that "reaction" can be understood in a very broad sense. It can describe physical reactions as implied in the previous description: for example in neutronics the $(n, 2n)$ has multiplicity $\nu_{(n, 2n)} = 2$ etc. see [52]. But it can also describe more "artificial" reactions, introduced only for practical reasons: for example, one may need "reaction" 0 to gather isotropic contributions (and in this case $\nu_0 \neq 0$) and reactions of higher number to treat progressively the anisotropy. Many other decompositions can be applied, more or less computationally efficient, depending on their probabilities of occurrence. The term S is a source term, it will be dealt with separately in the following sections (see mainly section 9.9) due to the slightly different structure it confers to the equation. We finally emphasize a notation trick which will be used all along the following chapters and which has already been encountered in (9.2):

$$\begin{aligned}\sigma_r(\mathbf{x}, t, \mathbf{v}) &= \iint \sigma_r(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') d\mathbf{v}' = \iint \sigma_r(\mathbf{x}, t, v\omega, v'\omega') d\mathbf{v}' d\omega', \\ &= \iint \sigma_r(\mathbf{x}, t, v, v', \omega \cdot \omega') d\mathbf{v}' d\omega' = \sigma_r(\mathbf{x}, t, \mathbf{v}).\end{aligned}$$

In the above expression, the dependence with respect to ω seemed to be omitted on the right hand side (and not in the left hand side) and the notation may appear abusive. We insist it is not, as the angular distribution of ω' with respect to ω can be described through the scalar product $\omega \cdot \omega'$, hence with only one variable (naming $\omega \cdot \omega'$) instead of two (naming ω and ω'). The notation is not conventional but is more adapted for the material of some of the following chapters⁷.

Equation (9.1) must come with proper initial and boundary conditions for wellposedness. We denote by n_x the outward unitary vector normal at point $\mathbf{x} \in \partial\mathcal{D}$. Furthermore, we introduce

$$\Gamma^- = \{(\mathbf{x}, \mathbf{v}) \in \partial\mathcal{D} \times \mathbb{R}^3 | \mathbf{v} \cdot n_x < 0\}.$$

We also generally denote by $\mathcal{C}_b(\mathbb{X})$ the space of continuous functions from \mathbb{X} to \mathbb{R} bounded on \mathbb{X} .

Let us consider

- the initial condition $u_0(\mathbf{x}, \mathbf{v}) \in \mathcal{C}_b(\overline{\mathcal{D}} \times \mathbb{R}^3)$,
- the boundary condition $u^- \in \mathcal{C}_b([0, T] \times \Gamma^-)$ must be in agreement with the initial condition at the boundary of the domain, i.e. such that

$$u_0(\mathbf{x}, \mathbf{v}) = u^-(0, \mathbf{x}, \mathbf{v}), \forall (\mathbf{x}, \mathbf{v}) \in \Gamma^-.$$

- and the term source $S(\mathbf{x}, t, \mathbf{v}) \in \mathcal{C}_b([0, T] \times \overline{\mathcal{D}} \times \mathbb{R}^3)$.

⁵The term *cross-sections* is commonly used in neutronics. In photonics, authors usually use the term *opacities*.

⁶Depending on the sign of the quantity $\sigma_t - \iint \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') d\mathbf{v}'$.

⁷The same notation trick is used in [245].

Then there exists $u \in \mathcal{C}_b([0, T] \times \mathcal{D} \times \mathbb{R}^3)$ unique solution of (9.1), see [127, 58, 132, 10]. In the following sections, we suppose every considered problem comes with proper initial and boundary conditions even if they are not (abusively) reminded. Consequently the solution of our problem always exists and it is legit to look for it. In this document, this is done with a MC resolution scheme.

When there are no ambiguities, the dependences $(\mathbf{x}, t, \mathbf{v})$ may not be recalled. In the first following sections, without loss of generalities, we drop the source term and detail how it can be taken into account in section 9.9.

9.1 General Methodology for the construction of an MC scheme

In this section, we describe the general methodology applied in order to build an MC scheme for solving (9.1). The first step consists in rewriting (9.1) in an integral form. The computations may appear tedious due to the fact that all dependences with respect to $(\mathbf{x}, t, \mathbf{v})$ *must* be recalled: it is important in order to identify and apply the less constraining hypothesis during the MC resolution. In order to rewrite (9.1) in integral form, we perform several successive *exact* changes of variable. By *exact* we mean *no approximations on the shapes of the cross-sections are made before section 9.6*. We detail every of them in the following sections. The described methodology leads to the adjoint MC resolution of the equation (or backward Kolmogorov equation, see [219]). The direct counterpart (forward Kolmogorov equation) will be studied in section 9.5.

As explained before, the methodology resumes to a succession of changes of variable. The first one consists in rewriting the transport equation (9.1) on a characteristic $\mathbf{x} + \mathbf{v}t$. Equation (9.1) without source term ($S = 0$) rewritten along a characteristic $(\mathbf{x} + \mathbf{v}s, s, \mathbf{v})$ becomes

$$\partial_s u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}) = -v\sigma_t(\mathbf{x} + \mathbf{v}s, s, v)u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}) + \int v\sigma_s(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}')d\mathbf{v}'. \quad (9.3)$$

Let us multiply each side of the equality by

$$\exp \left[\int_0^s v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha \right].$$

We then get

$$\partial_s \left[u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}) e^{\int_0^s v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \right] = e^{\int_0^s v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \int v\sigma_s(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}')d\mathbf{v}'. \quad (9.4)$$

Integrating (9.4) in the time interval $[0, t]$ leads to

$$u(\mathbf{x} + \mathbf{v}t, t, \mathbf{v}) = u_0(\mathbf{x}, \mathbf{v})e^{-\int_0^t v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} + \int_0^t e^{-\int_s^t v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \int v\sigma_s(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}')d\mathbf{v}'ds. \quad (9.5)$$

We have then

$$u(\mathbf{x}, t, \mathbf{v}) = u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v})e^{-\int_0^t v\sigma_t(\mathbf{x} - \mathbf{v}(t-\alpha), \alpha, v)d\alpha} + \int_0^t e^{-\int_s^t v\sigma_t(\mathbf{x} - \mathbf{v}(t-\alpha), \alpha, v)d\alpha} \iint v\sigma_s(\mathbf{x} - \mathbf{v}(t-s), s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} - \mathbf{v}(t-s), s, \mathbf{v}')d\mathbf{v}'ds. \quad (9.6)$$

Equation (9.6) is an integral equation but still needs to be worked on: first, notice that

$$\begin{aligned} e^{-\int_0^t v\sigma_t(\mathbf{x} - \mathbf{v}(t-\alpha), \alpha, v)d\alpha} &= e^{-\int_0^t v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t-\alpha, v)d\alpha}, \\ &= \int_t^\infty v\sigma_t(\mathbf{x} - \mathbf{v}s, t-s, v)e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t-\alpha, v)d\alpha}ds. \end{aligned}$$

Then, the integral counterpart of (9.1) is given by

$$\begin{aligned} u(\mathbf{x}, t, \mathbf{v}) &= \\ &+ \int_t^\infty u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) v\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v) d\alpha} ds \\ &+ \int_0^t e^{-\int_s^t v\sigma_t(\mathbf{x} - \mathbf{v}(t - \alpha), \alpha, v) d\alpha} \iint v\sigma_s(\mathbf{x} - \mathbf{v}(t - s), s, \mathbf{v}, \mathbf{v}') u(\mathbf{x} - \mathbf{v}(t - s), s, \mathbf{v}') d\mathbf{v}' ds. \end{aligned} \quad (9.7)$$

Building an MC scheme now implies introducing a set of random variables together with their probability measure in order to rewrite (9.7) as an expectation. The choice of the set of random variables is not unique and consequently leads to different MC schemes having different properties. In the following sections, we detail the construction of three MC schemes

- the analog one (section 9.2),
- the semi-analog one (section 9.3),
- and the non-analog one (section 9.4).

Their asymptotic properties will be investigated later on, in section 9.7.

9.2 The analog (Adjoint) MC scheme (mimics physics)

In this section, we describe the analog MC scheme. This scheme is usually hinted at as the scheme mimicing the physics of the particles. This will be clarified in the following sections. For the moment we focus on its construction. Every dependences with respect to $(\mathbf{x}, t, \mathbf{v})$ of the cross-sections are explicit. The notations and equations are consequently heavy but it helps identifying the treatments to perform on a *MC particle*⁸ in order to solve (9.1) with this MC method.

9.2.1 Expectation form over the analog set of random variables

The first step in order to rewrite (9.7) as an expectation consists in identifying a probability measure relative to the time integration in equation (9.7). Let us perform a change of variable ($\beta = t - s$ and β is immediately replaced by s) in the time integrations in the scattering part. We obtain

$$\begin{aligned} u(\mathbf{x}, t, \mathbf{v}) &= \\ &+ \int_t^\infty u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) v\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v) d\alpha} ds \\ &+ \int_0^t e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v) d\alpha} \int v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') d\mathbf{v}' ds. \end{aligned} \quad (9.8)$$

It is then possible to factorize by

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds = \mathbf{1}_{[0, \infty]}(s) v\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v) d\alpha} ds.$$

The above expression is a probability measure $\forall (\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$: indeed, it is positive and sums up to 1 $\forall (\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$. Using its expression in (9.8) leads to

$$u(\mathbf{x}, t, \mathbf{v}) = \iint \begin{bmatrix} +\mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) & \delta_{\mathbf{v}}(\mathbf{v}') \\ +\mathbf{1}_{[0, t]}(s) & u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') & \frac{\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}')}{\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)} \end{bmatrix} f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds d\mathbf{v}'. \quad (9.9)$$

Let us work on the components of the scattering cross-section σ_s : without loss of generality, one can decompose each reaction cross-sections $\forall r \in \{0, \dots, N_R\}$ as

$$v\sigma_r(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') = v\sigma_r(\mathbf{x} - \mathbf{v}s, t - s, v) P_r(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}').$$

⁸The term will be defined very soon.

In the above expression, $\forall(\mathbf{y}, \beta, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$ we have

$$\begin{aligned}\sigma_r(\mathbf{y}, \beta, v) &= \int \sigma_r(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') d\mathbf{v}', \forall r \in \{0, \dots, N_R\}, \\ P_r(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') &= \frac{\sigma_r(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}')}{\sigma_r(\mathbf{y}, \beta, v)}, \forall r \in \{0, \dots, N_R\}, \\ \sigma_s(\mathbf{y}, \beta, v) &= \int \sigma_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') d\mathbf{v}', \\ \sigma_s(\mathbf{y}, \beta, v) &= \sum_{r=0}^{N_R} \nu_r(v) \sigma_r(\mathbf{y}, \beta, v).\end{aligned}\tag{9.10}$$

Consequently $\forall r \in \{0, \dots, N_R\}$ and $\forall(\mathbf{y}, \beta, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$, we can identify a three dimensional (as $\mathbf{v}' \in \mathbb{R}^3$) positive measure, summing up to 1 (i.e. a probability measure)

$$P_{\mathbf{V}'}^r(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}') d\mathbf{v}' = P_r(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') d\mathbf{v}', \forall(\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3.\tag{9.11}$$

Equation (9.9) can then be rewritten

$$\begin{aligned}u(\mathbf{x}, t, \mathbf{v}) &= \\ \iint &\left[\begin{array}{ll} +\mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) \\ +\mathbf{1}_{[0, t]}(s) \sum_{r=0}^{N_R} & \nu_r(v) u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') \quad \frac{\sigma_r(\mathbf{x} - \mathbf{v}s, t - s, v)}{\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)} P_{\mathbf{V}'}^r(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}') \\ f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds d\mathbf{v}' \end{array} \right] \end{aligned}\tag{9.12}$$

From the definition of the total cross-section with respect to the reaction ones, $\sigma_t = \sum_{r=0}^{N_R} \sigma_r$, we have $\forall r \in \{0, \dots, N_R\}$ and $\forall(\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$

$$0 \leq \frac{\sigma_r(\mathbf{x}, t, v)}{\sigma_t(\mathbf{x}, t, v)} \leq 1, \quad \text{and} \quad \sum_{r=0}^{N_R} \frac{\sigma_r(\mathbf{x}, t, v)}{\sigma_t(\mathbf{x}, t, v)} = 1.$$

Consequently, the quantity $f_{\mathcal{B}}(\mathbf{x}, t, s, \mathbf{v}, b) db$ defined as

$$f_{\mathcal{B}}(\mathbf{x}, t, s, \mathbf{v}, b) db = \sum_{r=0}^{N_R} \frac{\sigma_r(\mathbf{x} - \mathbf{v}s, t - s, v)}{\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)} \delta_r(b),$$

is a probability measure. In fact, it is the probability measure of a multinomial (i.e. discrete) random variable. It is denoted as

$$\mathcal{B} \sim \mathcal{M} \left(r \in \{0, \dots, N_R\}, \left(\frac{\sigma_r}{\sigma_t} \right)_{r \in \{0, \dots, N_R\}} \right),$$

where the $N_R + 1$ states $r \in \{0, \dots, N_R\}$ with respective probabilities $(\frac{\sigma_r}{\sigma_t})_{r \in \{0, \dots, N_R\}}$ $\forall(\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$ are explicated. Let us now introduce the following random variables associated to the previously identified probability measures:

$$\begin{cases} \tau & \text{with probability measure } f_\tau(\mathbf{x}, t, \mathbf{v}) ds, \\ \mathbf{V}' & \text{with probability measure } P_{\mathbf{V}'}^r(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}') d\mathbf{v}', \\ \mathcal{B} & \text{with probability measure } f_{\mathcal{B}}(\mathbf{x}, t, s, \mathbf{v}, b) db. \end{cases}\tag{9.13}$$

Then (9.12) can be rewritten *in an adjoint recursive way* as an expectation over the above set of random variables (9.13)

$$u(\mathbf{x}, t, \mathbf{v}) = \mathbb{E} \left[\mathbf{1}_{[t, \infty]}(\tau) u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) + \mathbf{1}_{[0, t]}(\tau) \sum_{r=0}^{N_R} u(\mathbf{x} - \mathbf{v}\tau, t - \tau, \mathbf{V}') \nu_r(v) \delta_r(\mathcal{B}) \right].\tag{9.14}$$

Such adjoint formulation allows for example solving the transport equation and obtain approximation of the solution at a chosen point $(\mathbf{x}, t, \mathbf{v})$ of the simulation domain. Typically, this formulation is intensively used when one wants to recover accurate approximations on shielded detectors [173, 268].

The next step in order to describe an MC resolution is to define a *MC particle* together with the treatments one must apply to it in order to solve (9.14).

9.2.2 Construction of the analog MC scheme

Once the linear Boltzmann equation (9.1) rewritten as an expectation (9.14) over a set of defined random variables, formally, the construction of an MC scheme relies on looking for solutions of (9.14) having the particular form

$$u_p(\mathbf{x}, t, \mathbf{v}) = w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)). \quad (9.15)$$

Such solution u_p is commonly called a *MC particle*. The MC scheme intensively uses the linearity of equation (9.1): if $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ are independent solutions of (9.1) then $\sum_{p=1}^{N_{MC}} u_p$ is also solution of (9.1). Now, remains to identify the operations one has to apply in order to make sure each $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ is effectively an MC solution of (9.14). To do so, we plug u_p into (9.14) and solve a system of (compatible) equations of unknowns $w_p(t), \mathbf{x}_p(t), \mathbf{v}_p(t)$. Plugging (9.15) in (9.14) leads to

$$w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)) = \begin{array}{ccccc} +\mathbf{1}_{[t, \infty[}(\tau) & w_p(0) & \delta_{\mathbf{x}-\mathbf{v}t}(\mathbf{x}_p(0)) & \delta_{\mathbf{v}}(\mathbf{v}_p(0)) \\ +\mathbf{1}_{[0, t]}(\tau) & \delta_r(\mathcal{B}) & \nu_r(v) w_p(t-\tau) & \delta_{\mathbf{x}-\mathbf{v}\tau}(\mathbf{x}_p(t-\tau)) & \delta_{\mathbf{v}'}(\mathbf{v}_p(t-\tau)). \end{array}$$

The above expression may be disturbing but allows identifying the conditional samplings (successions of samplings) and treatments they imply. The equations satisfied by the unknown fields of the solution u_p , naming $w_p(t), \mathbf{x}_p(t), \mathbf{v}_p(t)$, are

$$\begin{aligned} w_p(t) &= \mathbf{1}_{[t, \infty[}(\tau) w_p(0) & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) \nu_r(v) w_p(t-\tau), \\ \delta_{\mathbf{x}}(\mathbf{x}_p(t)) &= \mathbf{1}_{[t, \infty[}(\tau) \delta_{\mathbf{x}-\mathbf{v}t}(\mathbf{x}_p(0)) & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) \delta_{\mathbf{x}-\mathbf{v}\tau}(\mathbf{x}_p(t-\tau)), \\ \delta_{\mathbf{v}}(\mathbf{v}_p(t)) &= \mathbf{1}_{[t, \infty[}(\tau) \delta_{\mathbf{v}}(\mathbf{v}_p(0)) & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) \delta_{\mathbf{v}'}(\mathbf{v}_p(t-\tau)). \end{aligned}$$

Their resolution leads to

$$\left\{ \begin{array}{ll} w_p(t) &= \mathbf{1}_{[t, \infty[}(\tau) w_p(0) & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) \nu_r(v) w_p(t-\tau), \\ \mathbf{x}_p(t) &= \mathbf{1}_{[t, \infty[}(\tau) (\mathbf{x}_0 + \mathbf{v}t) & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) (\mathbf{x}_{t-\tau} + \mathbf{v}\tau), \\ \mathbf{v}_p(t) &= \mathbf{1}_{[t, \infty[}(\tau) \mathbf{v} & +\mathbf{1}_{[0, t]}(\tau) \delta_r(\mathcal{B}) \mathbf{V}'. \end{array} \right. \quad (9.16)$$

Practically, the above *recursive* system of equation in term of weight, position and velocity leads to the numerical treatment/algorithm (remember we here detailed the adjoint formulation) summed up in algorithm 1.

Remark 9.1 Note that within algorithm 1, we put forward the possibility to use two 'options'. Those options in fact correspond to two slightly different MC schemes. They are treated as options mainly because the practical differences in term of code developments are small:

- the first option is called 'full_analog'. It implies creating as many particles as the multiplicity ν_r of the sampled reaction r .
- The second option is called 'multiplicity'. It implies multiplying the weight of the MC particle enduring a reaction r by its multiplicity ν_r .

If the differences are very simple in practice, the practical effects can be very important. For example, in a fast multiplying media

- the 'full_analog' option may imply a fast growth of the number of MC particles to be treated. It can generate memory and/or computational issues.
- On the other hand, in the same conditions, the 'multiplicity' option ensures having a quite constant number of particles ($\approx N_{MC}$). The computation can be carried out.

However, these options do not bear the same asymptotical properties and do not capture the same physical regime. These asymptotical properties will be studied and analysed in section 9.7.

Note that algorithm 1 does not describe treatments for computing integrated values (called 'track length estimators' in the literature, see for example [165]) but only punctual one at time t (called 'indicated value' in the literature). Integrated observables will be dealt with in chapter 10. In this chapter, we will consider the linear Boltzmann equation coupled to different systems. In this context, a track length estimator may be mandatory for consistency.

Algorithm 1: The MC analog scheme described in term of algorithmic operations in order to compute (adjoint) $u(\mathbf{x}, t, \mathbf{v})$.

```

1 set  $u(\mathbf{x}, t, \mathbf{v}) = 0$ 
2 for  $p \in \{1, \dots, N_{MC}\}$  do
3   set  $s_p = t$  #this will be the life time of particle  $p$ 
4   set  $x_p = \mathbf{x}$ 
5   set  $\mathbf{v}_p = \mathbf{v}$ 
6   set  $w_p = \frac{1}{N_{MC}}$ 
7   while  $s_p > 0$  and  $w_p > 0$  do
8     if  $x_p \notin \mathcal{D}$  then
9       #here a general function for the application of arbitrary boundary conditions
10      apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
11    end
12    Sample  $\tau$  from the distribution having probability measure  $f_\tau(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p)ds$ .
13    if  $\tau > s_p$  then
14      #see the treatment in factor of  $\mathbf{1}_{[t, \infty]}(\tau)$  in (9.16)
15      #move the particle  $p$ 
16       $x_p \leftarrow \mathbf{x}_p + \mathbf{v}_p s_p$ ,
17      #set the life time of particle  $p$  to zero:
18       $s_p \leftarrow 0$ 
19      #do not change the velocity of particle  $p$ 
20      #do not change the weight of particle  $p$ 
21      #tally the contribution of particle  $p$ 
22       $u(\mathbf{x}, t, \mathbf{v}) += w_p \times u_0(\mathbf{x}_p, \mathbf{v}_p)$ 
23    end
24  else
25    #see the recursive treatment in factor of  $\mathbf{1}_{[0, t]}(\tau)$  in (9.16)
26    Sample  $\mathcal{B}$  from a multinomial law of probability measure  $f_{\mathcal{B}}(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, b)db$ 
27    if  $\mathcal{B} = r$  then
28      if full_analog then
29        #do not change its weight and split the particle
30        create  $\nu_r(\mathbf{v}_p)$  particles  $p' \in \{1, \dots, \nu_r(\mathbf{v}_p)\}$  with the same characteristics as  $p$ 
31        for  $p' \in \{1, \dots, \nu_r(\mathbf{v}_p)\}$  do
32          #Sample their velocities from  $P_{\mathbf{v}'}^r(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}')$ 
33           $\mathbf{v}_{p'} = V'$  for every created particle  $p' \in \{1, \dots, \nu_r(\mathbf{v}_p)\}$ 
34        end
35      end
36      if multiplicity then
37        #change the weight of the particle
38         $w_p \leftarrow \nu_r(\mathbf{v}_p) w_p$ 
39        #Sample the velocity of particle  $p$  from  $P_{\mathbf{v}'}^r(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}')$ 
40         $\mathbf{v}_p = V'$ 
41      end
42    end
43    #move the particle  $p$ 
44     $x_p \leftarrow \mathbf{x}_p + \mathbf{v}_p \tau$ ,
45    #set the life time of particle  $p$  to:
46     $s_p \leftarrow s_p - \tau > 0$ 
47  end
48 end
49 end

```

The description of algorithm 1 deduced from the recursive set of equations (9.16) shows that the

analog MC scheme is defined by a set of samplings depending on almost every variables $\mathbf{x}, t, \mathbf{v}, \mathbf{v}'$. In practice, some additional approximations are made in order to make the samplings easier to compute. Those are presented later in section 9.6. Note that during the analog treatment of an MC particle, the weight of a particle does not change. In this sense, the scheme mimics physics as one can set the initial weight to m^9 so that an MC particle represents a physical one. The asymptotic property of the scheme will be emphasized in section 9.7.

9.3 The semi-analog (Adjoint) MC scheme (implicit capture)

In the previous section, we presented the analog MC scheme built from the integral form (9.7) of the linear Boltzmann equation (9.1). We furthermore introduced the set of random variables to rewrite it as an expectation (9.14). The semi-analog MC scheme (also known as ‘implicit capture’ in the literature [173, 268]) starts from the same integral form (9.7)¹⁰. We can start the description of the semi-analog scheme from (9.9) introducing the probability measure of the interaction time

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds = \mathbf{1}_{[0, \infty]}(s)v\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)e^{-\int_0^s v\sigma_t(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v)d\alpha}ds, \\ \forall(\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3. \quad (9.17)$$

It is the same as in the previous section.

9.3.1 Expectation form over the semi-analog set of random variables

The description of the semi-analog scheme begins with (9.9), reminded here

$$u(\mathbf{x}, t, \mathbf{v}) = \iint \begin{bmatrix} +\mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) & \delta_{\mathbf{v}}(\mathbf{v}') \\ +\mathbf{1}_{[0, t]}(s) & u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') & \frac{\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}')}{\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)} \end{bmatrix} f_\tau(\mathbf{x}, t, \mathbf{v}, s)dsd\mathbf{v}'. \quad (9.18)$$

The scheme mainly differs from the analog one by the choice of the random variables introduced for the scattering cross-section. Without loss of generality, we can write

$$v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') = v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v)P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}').$$

In the above expression, $\forall(\mathbf{y}, \beta) \in \mathcal{D} \times [0, T]$ we have

$$\sigma_s(\mathbf{y}, \beta, v) = \int \sigma_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}')d\mathbf{v}', \\ P_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') = \frac{\sigma_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}')}{\sigma_s(\mathbf{y}, \beta, v)}. \quad (9.19)$$

The quantity $P_{\mathbf{V}'}^s(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}')d\mathbf{v}' = P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}')d\mathbf{v}'$ is positive and is summing up to 1. It is consequently a three-dimensional probability measure $\forall(\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$. The difference between the analog scheme of the previous section and the one presented here comes from the fact the probability measure for the samplings of the velocity \mathbf{V}' is here *averaged over the set of reactions* $r \in \{0, \dots, N_R\}$. Equation (9.18) can then be rewritten

$$u(\mathbf{x}, t, \mathbf{v}) = \iint \begin{bmatrix} +\mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) \\ +\mathbf{1}_{[0, t]}(s) & u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') \end{bmatrix} \frac{\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v)}{\sigma_t(\mathbf{x} - \mathbf{v}s, t - s, v)} P_{\mathbf{V}'}^s(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}') f_\tau(\mathbf{x}, t, \mathbf{v}, s)dsd\mathbf{v}'. \quad (9.20)$$

Introduce the following random variables associated to the previously identified probability measures

$$\begin{cases} \tau & \text{with probability measure } f_\tau(\mathbf{x}, t, \mathbf{v})ds, \\ \mathbf{V}' & \text{with probability measure } P_{\mathbf{V}'}^s(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}')d\mathbf{v}'. \end{cases} \quad (9.21)$$

⁹i.e. the physical weight of the particles of interest, see part I.

¹⁰this will not be the case for the non-analog MC scheme of section 9.4.

Then (9.20) can be rewritten *in an adjoint recursive way* as an expectation over the above set of random variables (9.21)

$$u(\mathbf{x}, t, \mathbf{v}) = \mathbb{E} \left[\mathbf{1}_{[t, \infty[}(\tau) u_0(\mathbf{x} - \mathbf{v}\tau, \mathbf{v}) + \mathbf{1}_{[0, t]}(\tau) \frac{\sigma_s(\mathbf{x} - \mathbf{v}\tau, t - \tau, v)}{\sigma_t(\mathbf{x} - \mathbf{v}\tau, t - \tau, v)} u(\mathbf{x} - \mathbf{v}\tau, t - \tau, \mathbf{V}') \right]. \quad (9.22)$$

For the semi-analog scheme, the reaction at the interaction time is not sampled. The MC particle endures *an averaged* reaction. The ratio $\frac{\sigma_s}{\sigma_t}$ corresponds to the probability for the particle of being scattered and verifies

$$\mathbb{E} \left[\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}) \right] = \sum_{r=0}^{N_R} \nu_r \frac{\sigma_r}{\sigma_t} = \frac{\sigma_s}{\sigma_t},$$

where \mathcal{B} is the multinomial random variable defined in the previous section.

9.3.2 Construction of the semi-analog MC scheme

Regarding the construction of the semi-analog MC scheme, the steps are the same as in section 9.2.2. We consider 'particle' solutions $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ of (9.22) having the particular form

$$u_p(\mathbf{x}, t, \mathbf{v}) = w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)). \quad (9.23)$$

Let us plug them into (9.22) to identify the operations to perform to make sure each $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ is solution of (9.22). This leads to

$$w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)) = \begin{cases} \mathbf{1}_{[t, \infty[}(\tau) & w_p(0) \\ \mathbf{1}_{[0, t]}(\tau) & \frac{\sigma_s}{\sigma_t}(\mathbf{x} - \mathbf{v}\tau, t - \tau, \mathbf{v}) w_p(t - \tau) \end{cases} \begin{matrix} \delta_{\mathbf{x}-\mathbf{v}t}(\mathbf{x}_p(0)) \\ \delta_{\mathbf{x}-\mathbf{v}\tau}(\mathbf{x}_p(t - \tau)) \end{matrix} \begin{matrix} \delta_{\mathbf{v}}(\mathbf{v}_p(0)) \\ \delta_{\mathbf{v}'}(\mathbf{v}_p(t - \tau)), \end{matrix}$$

so that the weight, the position and the velocity satisfy

$$\begin{cases} w_p(t) = \mathbf{1}_{[t, \infty[}(\tau) w_p(0) + \mathbf{1}_{[0, t]}(\tau) \frac{\sigma_s}{\sigma_t}(\mathbf{x}_p(t - \tau), t - \tau, \mathbf{v}_p(t - \tau)) w_p(t - \tau), \\ \mathbf{x}_p(t) = \mathbf{1}_{[t, \infty[}(\tau)(\mathbf{x}_0 + \mathbf{v}t) + \mathbf{1}_{[0, t]}(\tau)(\mathbf{x}_{t-\tau} + \mathbf{v}\tau), \\ \mathbf{v}_p(t) = \mathbf{1}_{[t, \infty[}(\tau)\mathbf{v} + \mathbf{1}_{[0, t]}(\tau)\mathbf{V}'. \end{cases} \quad (9.24)$$

Practically, the above system of equation in term of weight, position and velocity leads to the recursive numerical treatment/algorithm (remember we here detailed the adjoint formulation) summed up in

algorithm 2.

Algorithm 2: The MC semi-analog scheme described in term of algorithmic operations in order to compute (adjoint) $u(\mathbf{x}, t, \mathbf{v})$.

```

1 set  $u(\mathbf{x}, t, \mathbf{v}) = 0$ 
2 for  $p \in \{1, \dots, N_{MC}\}$  do
3   set  $s_p = t$  #this will be the remaining life time of particle p
4   set  $\mathbf{x}_p = \mathbf{x}$ 
5   set  $\mathbf{v}_p = \mathbf{v}$ 
6   set  $w_p(t) = \frac{1}{N_{MC}}$ 
7   while  $s_p > 0$  and  $w_p > 0$  do
8     if  $x_p \notin \mathcal{D}$  then
9       #here a general function for the application of arbitrary boundary conditions
10      apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
11    end
12    Sample  $\tau$  from the distribution having probability measure  $f_\tau(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p)ds$ .
13    if  $\tau > s_p$  then
14      #see the treatment in factor of  $\mathbf{1}_{[t, \infty]}(\tau)$  in (9.24)
15      #move the particle p
16       $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p s_p$ ,
17      #set the life time of particle p to zero:
18       $s_p \leftarrow 0$ 
19      #do not change the velocity of particle p
20      #do not change the weight of particle p
21      #tally the contribution of particle p
22       $u(\mathbf{x}, t, \mathbf{v}) += w_p \times u_0(\mathbf{x}_p, \mathbf{v}_p)$ 
23    end
24  else
25    #see the recursive treatment in factor of  $\mathbf{1}_{[0, t]}(\tau)$  in (9.24)
26    #change the particle weight
27     $w_p \leftarrow \frac{\sigma_s(\mathbf{x}_p, s_p - \tau, \mathbf{v}_p)}{\sigma_t(\mathbf{x}_p, s_p - \tau, \mathbf{v}_p)} w_p$ 
28    #Sample the velocity  $\mathbf{V}'$  of particle p from  $P_{\mathbf{V}'}^s(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}')d\mathbf{v}'$ 
29     $\mathbf{v}_p = \mathbf{V}'$ 
30    #move the particle p
31     $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p \tau$ ,
32    #set the life time of particle p to:
33     $s_p \leftarrow s_p - \tau > 0$ 
34  end
35 end
36 end

```

Algorithm 2 only differs from algorithm 1 in the recursive part of the treatment (if $\tau < s_p$). The sampling of the velocity \mathbf{V}' is averaged and the weight of the particle is multiplied by the ratio $\frac{\sigma_s}{\sigma_t}$ at the position and at the instant of the shock. The latter ratio corresponds to the probability for the particle of being scattered. The MC particle does not anymore represent the behaviour of a physical particle at $\mathbf{x}, t, \mathbf{v}$ but rather the behaviour of a population of physical particles at $\mathbf{x}, t, \mathbf{v}$. With this treatment, the weight of an MC particle never goes to zero if $\sigma_s \neq 0$. An MC particle is never explicitly captured hence the denomination 'implicit capture' for this scheme. The asymptotic property of the scheme will be emphasized in section 9.7.

Once again, we insist on the fact that the semi-analog MC scheme is defined by a set of samplings depending on almost every variables $\mathbf{x}, t, \mathbf{v}, \mathbf{v}'$. In practice, some additional approximations are made in order to make the samplings easier to compute. Those are presented later in section 9.6.

9.4 The non-analog (Adjoint) MC scheme

In the previous sections, we presented two MC schemes for solving the linear Boltzmann equation (9.1). Those are mostly used in neutronics applications. In this section, we describe the non-analog scheme, intensively applied in photonic ones. As in the previous sections, we first rewrite the linear Boltzmann equation (9.1) in an integral form. The two previous schemes were both built from (9.7). The non-analog one is obtained from different changes of variables which are detailed in the next section. We then present the set of random variables at the basis of the MC scheme. The scheme is also referred as 'capture along the flight path' in the literature and care will be taken to emphasize why.

9.4.1 Expectation form over the non-analog set of random variables

First, as in the previous section 9.3, we rewrite the scattering cross-section

$$v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') = v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v)P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}').$$

In the above expression, $\forall (\mathbf{y}, \beta) \in \mathcal{D} \times [0, T]$ we have

$$\begin{aligned}\sigma_s(\mathbf{y}, \beta, v) &= \int \sigma_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') d\mathbf{v}', \\ P_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}') &= \frac{\sigma_s(\mathbf{y}, \beta, \mathbf{v}, \mathbf{v}')}{\sigma_s(\mathbf{y}, \beta, v)}.\end{aligned}\tag{9.25}$$

The quantity $P_{\mathbf{V}'}^s(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}') d\mathbf{v}' = P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') d\mathbf{v}'$ is positive and is summing up to 1. It is consequently a three-dimensional probability measure $\forall (\mathbf{x}, t, \mathbf{v}) \in \mathcal{D} \times [0, T] \times \mathbb{R}^3$. It is the same as for the semi-analog scheme of section 9.3. The non-analog scheme now needs the introduction of

$$\sigma_a = \sigma_t - \sigma_s.$$

The quantity σ_a is not always equal to the absorption cross-section σ_0 (cross-section of multiplicity $\nu_0 = 0$). It is the case only for a particular set of reactions of the form $r \in \{0, 1\}$. Let us decompose σ_t into $\sigma_a + \sigma_s$ in (9.6). This allows keeping the term $e^{-\int_0^s v\sigma_a(\mathbf{x}-\mathbf{v}(t-\alpha), \alpha, v) d\alpha}$ in factor of u_0 and u . Now using the fact that

$$e^{-\int_0^t v\sigma_s(\mathbf{x}-\mathbf{v}(t-\alpha), \alpha, v) d\alpha} = e^{-\int_0^t v\sigma_s(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} = \int_t^\infty v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_s(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} ds,$$

equation (9.6) rewrites

$$\begin{aligned}u(\mathbf{x}, t, \mathbf{v}) &= \\ &+ \int_t^\infty u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) e^{-\int_0^s v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_s(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} ds \\ &+ \int_0^t \left[v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_s(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} e^{-\int_0^s v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} \right. \\ &\quad \left. \times \iint P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') d\mathbf{v}' \right] ds.\end{aligned}\tag{9.26}$$

It is then possible to factorize by

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds = \mathbf{1}_{[0, \infty]}(s) v\sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v) e^{-\int_0^s v\sigma_s(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} ds.\tag{9.27}$$

It is also a probability measure (with respect to σ_s rather than σ_t). We then rewrite the linear Boltzmann equation in another integral form

$$\begin{aligned}u(\mathbf{x}, t, \mathbf{v}) &= \\ &\iint \left[\begin{array}{cccc} \mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}) & e^{-\int_0^s v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} & \delta_{\mathbf{v}}(\mathbf{v}') \\ \mathbf{1}_{[0, t]}(s) & u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}') & e^{-\int_0^s v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v) d\alpha} & P_s(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}, \mathbf{v}') \end{array} \right] \\ &\quad \times f_\tau(\mathbf{x}, t, \mathbf{v}, s) d\mathbf{v}' ds.\end{aligned}\tag{9.28}$$

Integral form (9.28) obtained here is different from the one (9.7) used for the two previous schemes. It mainly differs due to the exponential term multiplying u_0 and u . Let us now introduce the random variables

$$\begin{cases} \tau & \text{with probability measure } f_\tau(\mathbf{x}, t, \mathbf{v})ds, \\ \mathbf{V}' & \text{with probability measure } P_{\mathbf{V}'}^s(\mathbf{x}, t, s, \mathbf{v}, \mathbf{v}')d\mathbf{v}'. \end{cases} \quad (9.29)$$

Equation (9.28) can then be rewritten *in an adjoint recursive way* as an expectation over the above set of non-analog random variables (9.29)

$$u(\mathbf{x}, t, \mathbf{v}) = \mathbb{E} \left[\begin{array}{ccc} +\mathbf{1}_{[t, \infty[}(\tau) & e^{-\int_0^t v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha} & u_0(\mathbf{x}-\mathbf{v}t, \mathbf{v}) \\ +\mathbf{1}_{[0, t]}(\tau) & e^{-\int_0^\tau v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha} & u(\mathbf{x}-\mathbf{v}\tau, t-\tau, \mathbf{V}') \end{array} \right]. \quad (9.30)$$

In the next section we deduce the MC treatments to apply to solve (9.30).

9.4.2 Construction of the Adjoint non-analog MC scheme

The steps for the construction of the non-analog MC scheme are similar to the previous ones. Let us consider 'particle' solutions $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ of (9.30) having the particular form

$$u_p(\mathbf{x}, t, \mathbf{v}) = w_p(t)\delta_{\mathbf{x}}(\mathbf{x}_p(t))\delta_{\mathbf{v}}(\mathbf{v}_p(t)). \quad (9.31)$$

Let us plug them into (9.30) in order to identify the operations to perform to make sure each $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ is solution of (9.30). This leads to

$$\begin{aligned} w_p(t)\delta_{\mathbf{x}}(\mathbf{x}_p(t))\delta_{\mathbf{v}}(\mathbf{v}_p(t)) = \\ +\mathbf{1}_{[t, \infty[}(\tau) w_p(0) \delta_{\mathbf{x}-\mathbf{v}t}(\mathbf{x}_p(0)) \delta_{\mathbf{v}}(\mathbf{v}_p(0)) e^{-\int_0^t v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha}, \\ +\mathbf{1}_{[0, t]}(\tau) w_p(t-\tau) \delta_{\mathbf{x}-\mathbf{v}\tau}(\mathbf{x}_p(t-\tau)) \delta_{\mathbf{v}'}(\mathbf{v}_p(t-\tau)) e^{-\int_0^\tau v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha}, \end{aligned}$$

so that the weight, the position and the velocity satisfy

$$\begin{cases} w_p(t) = \mathbf{1}_{[t, \infty[}(\tau)w_p(0)e^{-\int_0^t v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha} + \mathbf{1}_{[0, t]}(\tau)e^{-\int_0^\tau v\sigma_a(\mathbf{x}-\mathbf{v}\alpha, t-\alpha, v)d\alpha}w_p(t-\tau), \\ \mathbf{x}_p(t) = \mathbf{1}_{[t, \infty[}(\tau)(\mathbf{x}_0 + \mathbf{v}t) + \mathbf{1}_{[0, t]}(\tau)(\mathbf{x}_{t-\tau} + \mathbf{v}\tau), \\ \mathbf{v}_p(t) = \mathbf{1}_{[t, \infty[}(\tau)\mathbf{v} + \mathbf{1}_{[0, t]}(\tau)\mathbf{V}'. \end{cases} \quad (9.32)$$

Practically, the above system of equation in term of weight, position and velocity leads to the recursive numerical treatment/algorithm (remember we here detailed the adjoint formulation) summed up in

algorithm 3.

Algorithm 3: The MC non-analog scheme described in term of algorithmic operations in order to compute (adjoint) $u(\mathbf{x}, t, \mathbf{v})$.

```

1 set  $u(\mathbf{x}, t, \mathbf{v}) = 0$ 
2 for  $p = 1 \in \{1, \dots, N_{MC}\}$  do
3   set  $s_p = t$  #this will be the life time of particle p
4   set  $\mathbf{x}_p = \mathbf{x}$ 
5   set  $\mathbf{v}_p = \mathbf{v}$ 
6   set  $w_p(t) = \frac{1}{N_{MC}}$ 
7   while  $s_p > 0$  and  $w_p > 0$  do
8     if  $\mathbf{x}_p \notin \mathcal{D}$  then
9       #here a general function for the application of arbitrary boundary conditions
10      apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
11    end
12    Sample  $\tau$  from the distribution having probability measure  $f_\tau(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p)ds$ .
13    if  $\tau > s_p$  then
14      #see the treatment in factor of  $\mathbf{1}_{[t, \infty]}(\tau)$  in (9.32)
15      #change its weight
16       $w_p \leftarrow e^{-\int_0^{s_p} v_p \sigma_a(\mathbf{x}_p - \mathbf{v}_p \alpha, s_p - \alpha, \mathbf{v}_p) d\alpha} w_p$ 
17      #move the particle p
18       $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p \tau,$ 
19      #set the life time of particle p to zero:
20       $s_p \leftarrow 0$ 
21      #do not change the angle or velocity of particle p
22      #tally the contribution of particle p
23       $u(\mathbf{x}, t, \mathbf{v}) += w_p \times u_0(\mathbf{x}_p, \mathbf{v}_p)$ 
24    end
25  else
26    #see the recursive treatment in factor of  $\mathbf{1}_{[0, t]}(\tau)$  in (9.32)
27    #change the particle weight
28     $w_p \leftarrow e^{-\int_0^\tau v_p \sigma_a(\mathbf{x}_p - \mathbf{v}_p \alpha, s_p - \alpha, \mathbf{v}_p) d\alpha} w_p$ 
29    #Sample the velocity  $\mathbf{V}'$  of particle p from  $P_{\mathbf{V}'}^s(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}')d\mathbf{v}'$ 
30     $\mathbf{v}_p = \mathbf{V}'$ 
31    #move the particle p
32     $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p \tau,$ 
33    #set the life time of particle p to:
34     $s_p \leftarrow s_p - \tau > 0$ 
35  end
36 end
37 end

```

Algorithm 3 mainly differs from the two previous ones (algorithms 1 and 2) by the fact that

- the interaction time is sampled from σ_s rather than from σ_t ,
- the weight of the particle is modified along its flight path.

The sampling of the velocity \mathbf{V}' is averaged over the set of reactions at the position and at the instant of the interaction (as in section 9.3). Once again, the MC particle does not anymore represent the behaviour of a physical particle at $\mathbf{x}, t, \mathbf{v}$ but rather the behaviour of a population of physical particles at $\mathbf{x}, t, \mathbf{v}$ averaged in a homogeneous media. Indeed, the weight modification along the flight path of a particle corresponds to the solution of a punctual/homogeneous problem given by

$$\partial_s U_{\mathbf{x}, t, \mathbf{v}}(s) = -v \sigma_a(\mathbf{x} - \mathbf{v}s, t - s, v) U_{\mathbf{x}, t, \mathbf{v}}(s). \quad (9.33)$$

It is equivalent to having

$$\frac{U_{\mathbf{x},t,\mathbf{v}}(t)}{U_{\mathbf{x},t,\mathbf{v}}(0)} = e^{-\int_0^t v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v) d\alpha}, \quad (9.34)$$

and exactly corresponds to the weight modification along the flight path of each MC particles. The asymptotic property of the scheme will be emphasized in section 9.7.

We have already insisted on the fact that the non-analog MC scheme is defined by a set of samplings depending on almost every variables $\mathbf{x}, t, \mathbf{v}, \mathbf{v}'$. In practice, some additional approximations are often made in order to make the samplings easier to compute. Those are presented later in section 9.6.

9.5 Direct formulation and direct set of random variables

In the previous sections, we described the different sets of samplings/random variables allowing an MC resolution for the adjoint formulation. The built MC schemes allowed computing the solution at *a priori* prescribed points $(\mathbf{x}, t, \mathbf{v})$ of the simulation domain. In many applications, one is interested in the solution on the whole domain at a given time, i.e. $\forall \mathbf{x} \in \mathcal{D}$ and at time $t \in [0, T]$. In order to be able to compute the contribution of MC particles on the whole domain, it is convenient adopting a direct formulation.

9.5.1 Adjoint and direct formulations of the same transport equation

The linear Boltzmann equation of unknown \tilde{u} from which we built the three previous adjoint MC schemes is recalled here

$$\partial_t \tilde{u}(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_x \tilde{u}(\mathbf{x}, t, \mathbf{v}) = -v \sigma_t(\mathbf{x}, t, \mathbf{v}) \tilde{u}(\mathbf{x}, t, \mathbf{v}) + v \sigma_s(\mathbf{x}, t, \mathbf{v}) \int P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') \tilde{u}(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}'. \quad (9.35)$$

The notations of (9.25) for the scattering term are applied. The direct counterpart of (9.35) is given (see [219]) by

$$-\partial_t u(\mathbf{x}, t, \mathbf{v}) - \mathbf{v} \partial_x u(\mathbf{x}, t, \mathbf{v}) = -v \sigma_t(\mathbf{x}, t, \mathbf{v}) u(\mathbf{x}, t, \mathbf{v}) + \int v' \sigma_s(\mathbf{x}, t, \mathbf{v}') P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}'. \quad (9.36)$$

In this section, \tilde{u} is solution of the Kolmogorov Backward equation and u is solution of the Kolmogorov Forward equation, see [219]. Let us introduce $v \sigma_S(\mathbf{x}, t, \mathbf{v}) P_S(\mathbf{x}, t, \mathbf{v}, \mathbf{v}')$ such that

$$v \sigma_S(\mathbf{x}, t, \mathbf{v}) P_S(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') = v' \sigma_s(\mathbf{x}, t, \mathbf{v}') P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}').$$

By definition we have

$$v \sigma_S(\mathbf{x}, t, \mathbf{v}) = \int v' \sigma_s(\mathbf{x}, t, \mathbf{v}') P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') d\mathbf{v}',$$

and

$$P_S(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') = \frac{v' \sigma_s(\mathbf{x}, t, \mathbf{v}') P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}')}{v \sigma_S(\mathbf{x}, t, \mathbf{v})}. \quad (9.37)$$

The above expressions are general and independent of the shapes of the cross-sections. The previous definitions allow rewriting (9.36) as

$$-\partial_t u(\mathbf{x}, t, \mathbf{v}) - \mathbf{v} \partial_x u(\mathbf{x}, t, \mathbf{v}) = -v \sigma_t(\mathbf{x}, t, \mathbf{v}) u(\mathbf{x}, t, \mathbf{v}) + v \sigma_S(\mathbf{x}, t, \mathbf{v}) \int P_S(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}'. \quad (9.38)$$

In the following section, we present the non-analog direct MC scheme in order to solve (9.38). The constructions of the direct analog and semi-analog ones are not presented, they would be redundant and do not present particular additional difficulties.

9.5.2 Direct Integral formulation for the non-analog scheme

Exactly as in section 9.2, we suggest rewriting equation (9.36) in an integral form, as an expectation over an identified set of random variables and obtain the direct set of samplings for the non-analog scheme of section 9.4. Equation (9.38) rewritten on a characteristic becomes

$$-\partial_s u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}) = -v\sigma_t(\mathbf{x} + \mathbf{v}s, s, v)u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}) \\ + v\sigma_S(\mathbf{x} + \mathbf{v}s, s, v) \int P_S(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}')d\mathbf{v}'. \quad (9.39)$$

Integrating the total cross-section term in the time derivative, we obtain

$$-\partial_s u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v})e^{-\int_0^s v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} = \\ v\sigma_S(\mathbf{x} + \mathbf{v}s, s, v)e^{-\int_0^s v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \int P_S(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}')d\mathbf{v}'. \quad (9.40)$$

Integrating the latter equation with respect to time on the time step $[0, t]$ we get

$$u_0(\mathbf{x}, \mathbf{v}) = u(\mathbf{x} + \mathbf{v}t, t, \mathbf{v})e^{-\int_0^t v\sigma_t(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \\ + \int_0^t \left[\int P_S(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}') \right] d\mathbf{v}' ds. \quad (9.41)$$

Let us introduce the quantity

$$\sigma_A = \sigma_t - \sigma_S.$$

Cross-section σ_A does not generally denote the absorption cross-section σ_0 . It is, under certain assumptions (only two reactions σ_0, σ_1 with σ_1 of multiplicity $\nu_1 = 1$ for example). Applying the change of variable $\beta = \alpha - t$ we obtain

$$u_0(\mathbf{x}, \mathbf{v}) = u(\mathbf{x} + \mathbf{v}t, t, \mathbf{v})e^{-\int_0^t v\sigma_A(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \int_t^\infty v\sigma_S(\mathbf{x} + \mathbf{v}s, s, v)e^{-\int_0^s v\sigma_S(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} ds \\ + \int_0^t \left[\int P_S(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')u(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}') \right] d\mathbf{v}' ds. \quad (9.42)$$

Let us introduce the probability measure of the interaction time

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds = \mathbf{1}_{[0, \infty]}(s)v\sigma_S(\mathbf{x} + \mathbf{v}s, s, v)e^{-\int_0^s v\sigma_S(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} ds, \quad (9.43)$$

and the probability measure of the outer velocity

$$P_S(\mathbf{x} + \mathbf{v}s, s, \mathbf{v}, \mathbf{v}')d\mathbf{v}'.$$

Then the expectation form of (9.42) is given by

$$u_0(\mathbf{x}, \mathbf{v}) = \mathbb{E} \left[\begin{array}{ccc} +\mathbf{1}_{[t, \infty]}(\tau) & u(\mathbf{x} + \mathbf{v}t, t, \mathbf{v}) & e^{-\int_0^t v\sigma_A(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \\ +\mathbf{1}_{[0, t]}(\tau) & u(\mathbf{x} + \mathbf{v}\tau, \tau, \mathbf{V}') & e^{-\int_0^\tau v\sigma_A(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha} \end{array} \right]. \quad (9.44)$$

The above expression (9.44) relates recursively the initial condition u_0 to the solution at different times $t > 0$ and *a priori* unknown positions and velocities $\mathbf{x} \in \mathcal{D}, \mathbf{v} \in \mathbb{R}^3$ at time t .

9.5.3 Construction of the direct non-analog MC scheme

The steps for the construction of the direct non-analog MC scheme are once again pretty similar to the previous ones. Let us consider 'particle' solutions $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ of (9.44) having the particular form

$$u_p(\mathbf{x}, t, \mathbf{v}) = w_p(t)\delta_{\mathbf{x}}(\mathbf{x}_p(t))\delta_{\mathbf{v}}(\mathbf{v}_p(t)). \quad (9.45)$$

Let us plug them into (9.44) in order to identify the operations to perform to make sure each $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ is solution of (9.44). This leads to

$$w_p(0)\delta_{\mathbf{x}}(\mathbf{x}_p(0))\delta_{\mathbf{v}}(\mathbf{v}_p(0)) = \begin{matrix} +\mathbf{1}_{[t, \infty[}(\tau) & w_p(t) & \delta_{\mathbf{x}+\mathbf{v}t}(\mathbf{x}_p(t)) & \delta_{\mathbf{v}}(\mathbf{v}_p(t)) & e^{-\int_0^t v\sigma_A(\mathbf{x}+\mathbf{v}\alpha, \alpha, v)d\alpha} \\ +\mathbf{1}_{[0, t]}(\tau) & w_p(\tau) & \delta_{\mathbf{x}+\mathbf{v}\tau}(\mathbf{x}_p(\tau)) & \delta_{\mathbf{v}'}(\mathbf{v}_p(\tau)) & e^{-\int_0^\tau v\sigma_A(\mathbf{x}+\mathbf{v}\alpha, \alpha, v)d\alpha} \end{matrix}.$$

The weight, the position and the velocity satisfy

$$\left\{ \begin{array}{ll} w_p(0) = & \mathbf{1}_{[t, \infty[}(\tau)w_p(t)e^{-\int_0^t v\sigma_A(\mathbf{x}+\mathbf{v}\alpha, \alpha, v)d\alpha} + \mathbf{1}_{[0, t]}(\tau)e^{-\int_0^\tau v\sigma_A(\mathbf{x}+\mathbf{v}\alpha, \alpha, v)d\alpha}w_p(\tau), \\ \mathbf{x}_p(0) = & \mathbf{1}_{[t, \infty[}(\tau)(\mathbf{x}_t - \mathbf{v}t) + \mathbf{1}_{[0, t]}(\tau)(\mathbf{x}_\tau - \mathbf{v}\tau), \\ \mathbf{v}_p(0) = & \mathbf{1}_{[t, \infty[}(\tau)\mathbf{v} + \mathbf{1}_{[0, t]}(\tau)\mathbf{V}' \end{array} \right. \quad (9.46)$$

Practically, the above system of equation in term of weight, position and velocity leads to the recursive numerical treatment/algorithm (remember we here detailed the *direct* formulation) summed up in algorithm 4. Algorithm 4 mainly differs from the previous *backward* ones due to the *initial sampling* step which consists in making sure the MC particle discretisation of the initial condition is representative of u_0 . The sampling part of the algorithm is briefly and generally described in algorithm 5. But it deserves a more detailed attention and will be dealt with in section 9.8 in a more practical way (once a grid introduced for example). Once again, the MC particle does not anymore represent the behaviour of a physical particle at $\mathbf{x}, t, \mathbf{v}$ but rather the behaviour of a population of physical particles at $\mathbf{x}, t, \mathbf{v}$ averaged in a homogeneous media. The weight modification along the flight path of a particle corresponds to the solution of a punctual/homogeneous problem given by

$$\partial_s U_{\mathbf{x}, t, \mathbf{v}}(s) = -v\sigma_A(\mathbf{x} + \mathbf{v}s, s, v)U_{\mathbf{x}, t, \mathbf{v}}(s). \quad (9.47)$$

It is equivalent to having

$$\frac{U_{\mathbf{x}, t, \mathbf{v}}(t)}{U_{\mathbf{x}, t, \mathbf{v}}(0)} = e^{-\int_0^t v\sigma_A(\mathbf{x}+\mathbf{v}\alpha, \alpha, v)d\alpha}, \quad (9.48)$$

which exactly corresponds to the weight modification along the flight path of a particle.

Algorithm 4: The MC non-analog scheme described in term of algorithmic operations in order to compute (direct) $U(x, t)$.

```

1 #Initialize to zero the quantity of interest on the whole simulation domain  $\mathcal{D}$ 
2 set  $U(\mathbf{x}, t) = 0 \forall \mathbf{x} \in \mathcal{D}$ 
3 #SAMPLING: call the sampling algorithm 5
4 Sampling( $N_{MC}$ )
5 #TRACKING: make sure each  $u_p$  is an MC particles
6 for  $p \in \{1, \dots, N_{MC}\}$  do
7   set  $s_p = 0$  #this will be the current time of particle  $p$ 
8   while  $s_p < t$  and  $w_p > 0$  do
9     if  $x_p \notin \mathcal{D}$  then
10      #here a general function for the application of arbitrary boundary conditions
11      apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
12    end
13    Sample  $\tau$  from the distribution having probability measure  $f_\tau(\mathbf{x}_p, s_p, s, \mathbf{v}_p)ds$ .
14    if  $\tau > t$  then
15      #see the treatment in factor of  $\mathbf{1}_{[t, \infty]}(\tau)$  in (9.32)
16      #change the particle weight
17       $w_p \leftarrow e^{-\int_0^{t-\tau} v_p \sigma_A(\mathbf{x}_p + \mathbf{v}_p \alpha, s_p + \alpha, \mathbf{v}_p) d\alpha} w_p$ 
18      #move the particle  $p$ 
19       $\mathbf{x}_p = \mathbf{x}_p - \mathbf{v}_p \times (t - \tau),$ 
20      #set the life time of particle  $p$  to zero:
21       $s_p \leftarrow t$ 
22      #do not change the angle or velocity of particle  $p$ 
23      #tally the contribution of particle  $p$ 
24       $U(\mathbf{x}_p, t) += w_p$ 
25    end
26  else
27    #see the recursive treatment in factor of  $\mathbf{1}_{[0, t]}(\tau)$  in (9.32)
28    #change the particle weight
29     $w_p \leftarrow e^{-\int_0^\tau v_p \sigma_A(\mathbf{x}_p + \mathbf{v}_p \alpha, s_p + \alpha, \mathbf{v}_p) d\alpha} w_p$ 
30    #Sample the velocity  $\mathbf{V}'$  of particle  $p$  from  $P_{\mathbf{V}'}^s(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}')d\mathbf{v}'$ 
31     $\mathbf{v}_p = \mathbf{V}'$ 
32    #move the particle  $p$ 
33     $\mathbf{x}_p \leftarrow \mathbf{x}_p - \mathbf{v}_p \tau,$ 
34    #set the life time of particle  $p$  to:
35     $s_p \leftarrow s_p + \tau < t$ 
36  end
37 end
38 end

```

We here derived the direct counterpart of the non-analog MC scheme. The derivation of the analog and the non-analog ones are almost identical. We just want to hint at the fact that the semi-analog scheme may be an interesting solver if one needs to solve both the direct and the adjoint problem in the same simulation code. Indeed, it limits the differences between the direct and the adjoint solvers (the only difference coming from the 'inversion' of the scattering laws) with a relatively good asymptotic variance (see section 9.7) ensuring a good compromise between accuracy/efficiency of the solver and development/verification times (V&V, see [13]).

Algorithm 5: Sampling step in order to represent the initial condition $u_0(\mathbf{x}, \mathbf{v})$ with N_{MC} MC particles. More details are given in section 9.8.1.

```

1 Function sampling( $N_{MC}$ )
2   compute  $U_0 = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \int u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}$ 
3   #SAMPLING: Sample the MC particles in order to represent the initial condition  $u_0(\mathbf{x}, \mathbf{v})$ 
4   for  $p \in \{1, \dots, N_{MC}\}$  do
5     #Introduce the probability measure  $du_0(\mathbf{x}, \mathbf{v}) = \frac{u_0(\mathbf{x}, \mathbf{v})}{U_0} d\mathbf{x} d\mathbf{v}$ 
6     #Sample the correlated realisation  $X_p, V_p$ 
7      $(X_p, V_p) = \text{sample\_from\_} du_0()$  #This function is detailed in section 9.8.1
8     set  $\mathbf{x}_p = X_p$ 
9     set  $\mathbf{v}_p = V_p$ 
10    set  $w_p = \frac{U_0}{N_{MC}}$ 
11  end
12  return The population of MC particles is an  $N_{MC}$  approximation of the initial condition

```

The three most common MC schemes in order to solve the linear Boltzmann equation have been presented all along the previous sections. We identified the samplings they imply and the algorithmic treatments induced. The developer familiar with MC code/resolution may not exactly recognize the samplings detailed above, mainly as they were presented in their most general form. We did not make any assumptions on the 'shapes' of the cross-sections with respect to $(\mathbf{x}, t, \mathbf{v})$ because we wanted to highlight the fact that an MC scheme does not need them. In the following section, we present some classical approximations. We insist on the fact that they are only introduced in order to simplify the computations for the samplings and are not dictated by the MC resolution. Furthermore, we presented three MC schemes but we did not compare their relative performances. In section 9.7, some theoretical considerations on the convergence of the MC schemes (their asymptotic variance, their moments of high orders) are detailed. We first tackle the most common approximations encountered in MC simulation codes in order to simplify the samplings.

9.6 Common approximations to simplify the samplings and resolutions

In the above descriptions of the three most common MC schemes (adjoint or direct), we identified four sets of probability measures. The samplings from the latters can be very complex in practice mainly due to the dependence with respect to $(\mathbf{x}, t, \mathbf{v})$ of the cross-sections. Our aim in this section is to present the most classical approximations used in order to simplify those samplings. Developers of MC codes will be able to recognize the operations they apply together with the set of assumptions made. We insist that none of these choices are induced by the MC discretisation but rather by practical considerations. To emphasize this, we review the previous different samplings (time interaction, energy/angle) and state the most common assumptions made. We also highlight some (scarcely applied in practice) possibilities in order to sample the interaction time and compute the weight modification of the MC particle along its flight path. These new possibilities paves the path toward the new MC schemes [3] we present in the next chapter 10.

9.6.1 The interaction time τ of probability measure $f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds$

We first consider the sampling of the interaction time denoted by τ in the previous sections, having probability measure $f_\tau(\mathbf{x}, t, \mathbf{v}, s)ds$. Depending on the MC scheme, f_τ may imply

- the use of the total cross-section σ_t (see sections 9.2 and 9.3),
- or of the averaged scattering cross-section σ_s (see section 9.4) or its direct counterpart σ_S (see section 9.5).

In order to treat the more generally possible every cases, we here introduce the (more concise and lighter) notation

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds = \mathbf{1}_{[0, \infty]}(s) v \sigma(\mathbf{x} + \mathbf{v}s, s, v) e^{-\int_0^s v \sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v) d\alpha} ds. \quad (9.49)$$

Expression (9.49) corresponds to the direct counterpart of the probability measure of the interaction time, see (9.43), rather than the adjoint one (9.27). But the material of this section can be directly applied in the adjoint samplings once noticing that

$$\mathbb{P}^{adjoint}(\tau < t) = \mathbb{P}^{direct}(\tau < -t).$$

The above relation can be deduced comparing (9.27) and (9.43) for example.

In order to sample from an arbitrary probability measure, a common method consists in inverting its cumulative density function (cdf) defined by

$$F_\tau(\mathbf{x}, t, \mathbf{v}, s) = \int_{-\infty}^s f_\tau(\mathbf{x}, t, \mathbf{v}, \alpha) d\alpha. \quad (9.50)$$

It is a classical probability result [256] that if $\mathcal{U} \sim \mathcal{U}([0, 1])$ is sampled from a uniform random variable on $[0, 1]$, then τ , defined by

$$\mathcal{U} = F_\tau(\mathbf{x}, t, \mathbf{v}, \tau), \quad (9.51)$$

follows the desired distribution. If we now use the expression of the probability measure (9.49) in (9.51), we get

$$\begin{aligned} \mathcal{U} = F_\tau(\mathbf{x}, t, \mathbf{v}, \tau) &= \int_{-\infty}^\tau f_\tau(\mathbf{x}, t, \mathbf{v}, \alpha) d\alpha = \int_0^\tau v \sigma(\mathbf{x} + \mathbf{v}s, s, v) e^{-\int_0^s v \sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v) d\alpha} ds, \\ &= 1 - \int_\tau^\infty v \sigma(\mathbf{x} + \mathbf{v}s, s, v) e^{-\int_0^s v \sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v) d\alpha} ds, \\ &= 1 - e^{-\int_0^\tau v \sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v) d\alpha}. \end{aligned} \quad (9.52)$$

Noticing that if $\mathcal{U} \sim \mathcal{U}([0, 1])$ then $\tilde{\mathcal{U}} = 1 - \mathcal{U} \sim \mathcal{U}([0, 1])$, we can write without loss of generality

$$-\ln(\mathcal{U}) = \int_0^\tau v \sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v) d\alpha. \quad (9.53)$$

The above expression is analytical but even if we have access to the function $\mathbf{x}, t, \mathbf{v} \rightarrow \sigma(\mathbf{x}, t, \mathbf{v})$, this does not imply the integral in (9.53) is easily invertible along the flight path of each particle (i.e. for every direction $\omega = \frac{\mathbf{v}}{v}$ from position \mathbf{x}). In practice, some approximations are made in order to make the sampling easier. The most common choice is presented in the next section.

The classical approximations for the interaction time

If the integration along the flight path is difficult, some additional assumptions can be made in order to simplify the samplings. For example, it is usually assumed the cross-sections are constant

- with respect to time for a given set of non-overlapping time steps such that $[0, t] = \bigcup_{i=1}^{N_t} [t^i, t^{i+1}]$,
- and with respect to space for a given set of non-overlapping cells such that $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$.

During one time step $[t^n, t^{n+1}]$ of size $t^{n+1} - t^n = \Delta t_n$, the cross-section may be approximated by $\sigma(\mathbf{x}, t \in [t^n, t^{n+1}], v) \approx \sum_{i=1}^{N_x} \sigma_i^n(v) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x})$, where

$$\sigma_i^n(v) = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} \frac{1}{\Delta t_n} \int_{t^n}^{t^{n+1}} \sigma(\mathbf{x}, s, v) dx ds.$$

The sampling of the interaction time is now easier as (9.53) becomes

$$\begin{aligned} -\ln(\mathcal{U}) &= \int_0^\tau v\sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha, \\ &\approx \sum_{i=1}^{N_x} \int_0^\tau v\sigma_i^n(v) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x} + \mathbf{v}\alpha)d\alpha. \end{aligned} \quad (9.54)$$

Suppose the particle stays within the same cell \mathcal{D}_i between two events, then

$$\tau = -\frac{\ln(\mathcal{U})}{v\sigma_i^n(v)}. \quad (9.55)$$

It corresponds to the classical sampling of an exponential law of parameter $v\sigma_i^n(v)$. It is denoted by $\mathcal{E}(v\sigma_i^n(v))$ in the following lines. Expression (9.55) is independent of the energy/velocity discretisation of $\sigma_i^n(v)$ and additional approximation for this dimension may be performed.

Depending on how the cross-sections are averaged with respect to time and space, such hypothesis may be more or less constraining. For example, in k_{eff} computations in neutronics, the cross-sections do not depend on time and the spatial change in cross-sections depends on the choice of the materials. In other words, for such physical applications, it is enough spatially describing the skin of the materials to be analytical/exact. In this case, the approximation is not restrictive. When dealing with irradiation in neutronics, see [97, 98, 148, 95, 96], [3], the approximation may not hold as cross-sections may bear fast evolution during one time step (see section 10.1). In photonic applications, see chapter 10, one may need either a very fine mesh either an MC scheme having particular properties (see section 10.2) with respect to the cell size $|\mathcal{D}_i|$.

Suppose now the particle crosses several cells along its flight path. Without loss of generality we can assume the particle only crosses two cells and apply the result in a recursive way. Let us denote by \mathcal{D}_1 and \mathcal{D}_2 the crossed cells and introduce $t_c \in [0, \tau]$ the time in interval $[0, \tau]$ at which the particle reaches the interface between the two cells. In this case, (9.53) becomes

$$\begin{aligned} -\ln(\mathcal{U}) &= \int_0^{t_c} v\sigma_1^n(v)d\alpha + \int_{t_c}^\tau v\sigma_2^n(v)d\alpha, \\ -\ln(\mathcal{U}) &= t_c v\sigma_1^n(v) + (\tau - t_c)v\sigma_2^n(v). \end{aligned} \quad (9.56)$$

In this case, the interaction time can be rewritten

$$\tau = -\frac{\ln(\mathcal{U})}{v\sigma_2^n(v)} + t_c \left(1 - \frac{\sigma_1^n(v)}{\sigma_2^n(v)}\right). \quad (9.57)$$

Expression (9.57) recovers the classical exponential sampling (9.55) if

- the particle stays within the cell,
- or even if the cross-sections are the same on each side of the interface. If $\sigma_1^n = \sigma_2^n$, then the corrective term in (9.57) is zero the interaction time degenerates toward (9.55).

If $t_c \in [0, \tau[$ and $\sigma_1^n \neq \sigma_2^n$, a correction to the sampling $-\frac{\ln(\mathcal{U})}{v\sigma_2^n(v)}$ is applied. It depends on the time t_c at which the particle crosses the interface between the two cells but also on the ratio of cross-sections on each sides of the interface. Due to the *memorylessness* of exponential laws (see [219]), we can rely on another solution in order to sample the interaction time when the particle crosses the interface between two cells. Statistically, (9.56) is also equivalent to

$$\tau_1 = -\frac{\ln(\mathcal{U}_1)}{v\sigma_1^n(v)}, \text{ and } \begin{cases} \text{if } \tau_1 < t_c \text{ then } \tau = \tau_1, \\ \text{else } \tau_2 = -\frac{\ln(\mathcal{U}_2)}{v\sigma_2^n(v)} \text{ and } \tau = \tau_2, \end{cases} \quad (9.58)$$

where $(\mathcal{U}_1, \mathcal{U}_2)$ are two independent uniform random variables on $[0, 1]$. If the MC particle stays in the same cell, (9.58) degenerates toward (9.55). If not, even if $\sigma_1^n = \sigma_2^n$, a new uniform sampling \mathcal{U}_2 is introduced. This sampling is intensively applied in MC codes. It practically means that

- we can stop any MC particle at the interface between two cells,

- decrease its life time by t_c ,
- and sample a new uniform random variable in order to obtain a new sampled interaction time.

Equations (9.58) and (9.57) are statistically equivalent, the first one only needs one sampling whereas the second one needs as many samplings as crossed cells. With nowadays computers and *random number generators (RNGs)*, sampling one random variable or a set of several is not constraining. But we can imagine the corrected sampling (9.55), needing only one sampling, may have been interesting on architectures which were not able to ensure a large enough periodicity [171] of the RNGs or for computationally intensive ones [171, 220].

The reader interested in both adjoint and direct resolutions may notice that such approximation with respect to time and space has the following property: direct and adjoint samplings are the same. Performing the same computations as in (9.52) but to the adjoint formulation of the interaction time probability measure leads to the same expression of τ , given by (9.55)¹¹. In this particular case, for the analog and the semi-analog schemes for which the interaction time is sampled from σ_t , the samplings are the same in adjoint and direct formulation.

An original approximation in order to sample the interaction time

The above approximation is commonly used in many MC codes. It presents the drawback of imposing the same time and space discretisation to every simulated MC particles. For example, suppose there exists two main areas in a simulation domain

- \mathcal{A}_1 where σ needs small time steps to be accurate,
- and \mathcal{A}_2 where σ remains constant with respect to time.

Then an MC particle in area \mathcal{A}_2 will be imposed the same time step as the one in area \mathcal{A}_1 whereas it does not need it. In practice, this can lead to possible loss of computational time. On another hand, expression (9.49) shows that the time discretisation for sampling the interaction time only depends on the flight path of the MC particles. Now, assume we have access to the function $\forall(\mathbf{x}, t, v) \in \mathcal{D} \times [0, T] \times \mathbb{R}^+ \rightarrow \sigma(\mathbf{x}, t, v)$. The question is *what prevents us from analytically sampling from (9.49) via (9.53)?* The main difficulty may come from the fact that an analytical integration in (9.53) $\forall \omega = \frac{\mathbf{v}}{v} \in \mathbb{S}^2$ is difficult. In this section, we introduce a new approximation, motivated by the previous observation, which is intensively used in chapter 10 for the new unsplit MC scheme presented in [3].

Let us introduce a sequence of time steps $(\Delta t_i^p = t_p^{i+1} - t_p^i)_{i \in \{1, \dots, N_t^p\}}$ depending on the MC particle p such that $\forall p \in \{1, \dots, N_{MC}\}$, $\sum_{i=1}^{N_t^p} \Delta t_i^p = t$, and such that we can approximate (9.53) along the flight path of each particle by a (second order here) numerical integration strategy

$$\begin{aligned} & \int_0^t v\sigma(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha \\ & \approx \sum_{i=1}^{N_t^p} v \frac{\Delta t_i^p}{2} (\mathbf{1}_{[0, \tau]}(t_p^i)\sigma(\mathbf{x} + \mathbf{v}t_p^i, t_p^i, v) + \mathbf{1}_{[0, \tau]}(t_p^{i+1})\sigma(\mathbf{x} + \mathbf{v}t_p^{i+1}, t_p^{i+1}, v)). \end{aligned} \quad (9.59)$$

The number of integration point N_t^p depends on the MC particle p . In fact, it depends on the time and spatial area it evolves in. Where steep gradients of $\mathbf{x}, s, \mathbf{v} \rightarrow \sigma(\mathbf{x} + \mathbf{v}s, s, v)$ are encountered, N_t^p may be important whereas it may be smaller where σ behaves as a constant. Note that in (9.59), the integration is a second order one but higher order ones (quadrature rules, Simpson, Gauss-Lobatto, or others as chapter 5 of part II etc.) could have been used. The integration method here is not really the purpose. The idea is that the integral in (9.59) can be approximated along the flight path of a particle of direction $\omega = \frac{\mathbf{v}}{v}$ summing the contribution in each sub-interval $[t^i, t^{i+1}]$. From (9.59), we can obtain an approximation of the sampling τ : for a realisation \mathcal{U} of a uniform random variable on $[0, 1]$, let us

¹¹The only difference comes from the fact particle crosses first cell \mathcal{D}_2 and then eventually cell \mathcal{D}_1 .

denote by n_τ^p the first $i \in \{1, \dots, N_t^p\}$ such that

$$-\ln(\mathcal{U}) < \sum_{i=1}^{n_\tau^p} v \frac{\Delta t_i^p}{2} (\mathbf{1}_{[0,\tau]}(t_p^i) \sigma(\mathbf{x} + \mathbf{v} t_p^i, t_p^i, v) + \mathbf{1}_{[0,\tau]}(t_p^{i+1}) \sigma(\mathbf{x} + \mathbf{v} t_p^{i+1}, t_p^{i+1}, v)). \quad (9.60)$$

Then $t^{n_\tau^p} = \tau^p + \mathcal{O}\left(\max_{i \in \{1, \dots, N_t^p\}} (\Delta t_i^p)^\gamma\right)$ is an approximation of order γ^{12} of the exact sampling τ^p and is built along the flight path of the particle. It needs the evaluation of function $\mathbf{x}, s, \mathbf{v} \rightarrow \sigma(\mathbf{x} + \mathbf{v}s, s, v)$ at successive positions $\mathbf{x} + \mathbf{v}t_p^i$ and times t_p^i for $i \in \{1, \dots, n_\tau^p + 1\}$. The previous process for sampling τ^p is licit even for discontinuous cross-sections (jump at a cell interface for example), it does not need any regularity assumptions. In fact, it resumes to approximating the interaction time sampling thanks to a discrete law of (not equiprobable) states $i \in \{1, \dots, N_t^p\}$. This strategy is intensively applied in [3] in a coupling context (see also chapter 10).

In the next section, we recall the most common approximations made on the cross-sections with respect to energy v and angle ω (such that $\mathbf{v} = v\omega$).

9.6.2 The energy and angle correlated samplings $\mathbf{V}' = V'W'$

Depending on the scheme of interest, the scattering velocity can be sampled

- from $P_{\mathbf{V}'}^r$ as in (9.11) for the analog scheme of section 9.2,
- or from $P_{\mathbf{V}'}^s$ as in (9.25) for the semi-analog and non-analog ones of sections 9.3 and 9.4,
- or even from $P_{\mathbf{V}'}^S$ as in (9.37) for its direct counterpart in section 9.5.

In order to treat every MC schemes, we simplify the notations and consider the probability measure of the random variables $\mathbf{V}' = V'W'$ where $V' = |\mathbf{V}'|$ and $W' = \frac{\mathbf{V}'}{V'}$ is denoted by

$$P_{\mathbf{V}'}(\mathbf{v}, \mathbf{v}') d\mathbf{v}' = P_{V', W'}(v' \omega', v \omega) dv' d\omega' = P_{V', W'}(v', v, \omega', \omega) dv' d\omega'.$$

At first glance, one may argue the last expression neglects spatial and time discretisations. It is in fact general enough: velocity changes being punctual, i.e. only at the interaction times and not integrated along the flight path of the MC particles, space and time are only parameters of the sampling law. They can be dropped in the following descriptions.

Treating the correlated samplings of V', W'

Depending on the physics of interest, the energy $v = |\mathbf{v}|$ and the angle $\omega = \frac{\mathbf{v}}{v}$ can be independent or correlated. In neutronics, they are correlated. Sampling from a multidimensional law can be complex and time consuming (see Gibbs algorithm, Metropolis-hasting etc. [252, 46]). Such strategies may not be affordable as our MC resolution needs a correlated sampling at each iteration (which can be very frequent especially in diffusive media, see section 10.2). In order to avoid such difficulty, the scattering energy and angle cross-section $P_{V', W'}$ is often pretreated in order to ensure a sampling of the energy V' followed by a sampling of the angle W' conditional to the inner and outer energies (v, V') . For this, the marginal probability measure for the energy is introduced:

$$P_{v, \omega, V'}(v') dv' = \frac{\int P_{V', W'}(v', v, \omega', \omega) d\omega'}{\iint P_{V', W'}(v', v, \omega', \omega) d\omega' dv'} dv'.$$

It allows defining the conditional angular distribution by

$$P_{v', v, W'}(v', v, \omega', \omega) = \frac{P_{V', W'}(v', v, \omega', \omega)}{P_{V'}(v', v)}, \text{ which sums up to 1 with respect to } \omega' \text{ by definition.}$$

¹² γ is the order of the integration method and $\gamma = 2$ if we choose (9.59) as integration rule.

With the above definitions, the process consists in first sampling the energy V' knowing v from the probability measure

$$P_{v,V'}(v')dv',$$

and then sample the angle conditionally to having the triplet v, ω, V' from probability measure

$$P_{v,V',W'}(v', v, \omega', \omega)d\omega' = P_{v,V',W'}(v', v, \omega \cdot \omega')d\omega'.$$

Note that depending on the physics of interest and the shapes of the data σ_s , it may be more efficient from a computational point of view to build first the probability for the angle W' and sample the energy conditionally to the sampled angle W' . The pretreatments are almost the same and the above description is general.

In the particular case of monokinetic particles and isotropic scattering, i.e. $\sigma_s(v', v, \omega', \omega) = \sigma_s$, the direct and adjoint energy and angle samplings are the same.

Once the conditional distribution obtained...

Suppose now the physical data σ_s are pretreated as above, implying a first sampling of the energy V' then a sampling of W' conditionally to (v, V') . In practice, it is once again all about inverting the cdfs of the two probability measures. Let us introduce two independent samples $\mathcal{U}_1, \mathcal{U}_2$ of a uniform random variable, then (V', W') are defined by

$$\begin{cases} \mathcal{U}_1 = \int_{-\infty}^{V'} P_{v,V'}(v')dv', \\ \mathcal{U}_2 = \int_{-\infty}^{W'} P_{v,V',W'}(v, V', \omega \cdot \omega')d\omega'. \end{cases} \quad (9.61)$$

The complexity of the above samplings directly depends on the physical constants σ_s . For example, $\forall v$, the probability measure $P_{v,V'}(v')dv'$ may be discretised by a sequence of $G + 1$ points such that $\bigcup_{g=0}^G [v_g, v_{g+1}] = \mathbb{R}^+$ together with normalized¹³ basis functions $(\phi_g(v'))_{g \in \{0, \dots, G\}}$ defined on every subinterval $[v_g, v_{g+1}]$. Introducing $\mathbf{1}_{g'}(v') = \mathbf{1}_{[v_{g'}, v_{g'+1}]}(v')$, we get the discretised measure

$$P_{v,V'}(v')dv' \approx \sum_{g'=0}^G P_v^{g'} \phi_{g'}(v') \mathbf{1}_{g'}(v')dv'.$$

The above expression is general, it can both describe punctual cross-sections or multigroup ones [249, 71, 173, 268] depending on the shapes of the basis functions $(\phi_g)_{g \in \{0, \dots, G\}}$ and the definitions of the coefficients $(P_v^{g'})_{g' \in \{0, \dots, G\}}$. Independently of this choice, the sampling of the energy V' with discrete cross-sections can be made once again by conditional samplings. First, sample the outer subinterval \mathcal{G} by inverting a discrete probability measure: sample a uniform random variable $U_{\mathcal{G}}$ verifying

$$\mathcal{G} = \min_{h \in \{0, \dots, G\}} \left\{ \mathcal{U}_{\mathcal{G}} < \frac{\sum_{g'=0}^h P_v^{g'}}{\sum_{g'=0}^G P_v^{g'}} \right\}. \quad (9.62)$$

Then use the shape of the basis function $\phi_{\mathcal{G}}(v)$ in the sampled subinterval \mathcal{G} to obtain V' from another uniform random variable $\mathcal{U}_{V'}$ such that

$$\mathcal{U}_{V'} = \int_{-\infty}^{V'} \phi_{\mathcal{G}}(v')dv'. \quad (9.63)$$

¹³such that $\forall g \in \{0, \dots, G\}, \int \phi_g(v) \mathbf{1}_g(v)dv = 1$.

Sampling (9.63) is conditional to being in subinterval \mathcal{G} . Of course, in general, the basis functions $(\phi_g)_{g \in \{0, \dots, G\}}$ are chosen so that the computation of (9.63) can be carried out analytically (linear, log-linear, log-log on the subintervals $[v_g, v_{g+1}]_{g \in \{0, \dots, G\}}$). Regarding the conditional sampling for the angle, the process is quite the same. The main difference comes from the format strategies: for example Legendre polynomials may be used in order to represent the anisotropy of the scattering cross-sections [173, 268]. There exists many other ways depending on the physics of interest.

With the previous descriptions, the practical samplings with respect to energy and angles strongly depend on the format of the data available. It is up to the developer of the MC scheme to adapt the above material to its physics of interest and ensure performances with respect to memory and CPU consumption.

9.6.3 The modification of the weight of the particle $w_p(t)$

The aspect by which the three presented schemes, analog/semi-analog/non-analog (adjoint or direct), differ most may be the modification of the weights or not of the MC particles. We denote by τ the interaction time for an MC particle p , independently of the choice of the MC scheme. We then have:

- for the analog scheme ('full_analog' option), the weight $w_p(t) = w_p(0), \forall t \in [0, T]$ is kept constant. It does not change all along its flight path nor after a collision. It can be considered changed at a collision if the MC particle is captured/absorbed (reaction σ_0). In this case, the weight becomes zero $w_p(\tau) = 0$. In practice, it is more efficient 'killing' the MC particle, i.e. removing it from the list of MC particle to treat, than tracking a particle with zero weight.
- For the semi-analog scheme (or the analog one with the 'multiplicity' option), the weight $w_p(t)$ of an MC particle does not change all along its flight path. But it is multiplied, locally at position $\mathbf{x} + \mathbf{v}\tau$ and time τ , by the probability of being reemitted/scattered, i.e. $w_p(\tau) = \frac{\sigma_S}{\sigma_t}(\mathbf{x} + \mathbf{v}\tau, \tau, v)w_p(0)$ (direct form recalled here).
- For the non-analog scheme, the weight $w_p(t)$ of an MC particle changes all along the flight path of the particle. It is multiplied (direct form explicated here) by the solution of the punctual/homogeneous problem (9.83):

$$w_p(\tau) = e^{-\int_0^\tau v\sigma_A(\mathbf{x} + \mathbf{v}\alpha, \alpha, v)d\alpha}w_p(0).$$

Note that in this particular case, an MC particle crossing the interface between two cells has its weight affected whereas this was not the case for the previous schemes. Suppose the cross-sections are considered constant with respect to \mathbf{x} in each cell and assume the MC particle crosses two cells \mathcal{D}_1 and \mathcal{D}_2 (and recursively an arbitrary number of cells), then $\sigma_A(\mathbf{x}, t, \mathbf{v}) = \sigma_A^1(t, v)\mathbf{1}_{\mathcal{D}_1}(\mathbf{x}) + \sigma_A^2(t, v)\mathbf{1}_{\mathcal{D}_2}(\mathbf{x})$. Denote by $t_c \in [0, \tau]$ the time at which the MC particle reaches the interface position between those cells, then

$$\begin{aligned} w_p(\tau) &= w_p(0) & e^{-\int_0^{t_c} v\sigma_A^1(\alpha, v)d\alpha} & e^{-\int_{t_c}^\tau v\sigma_A^2(\alpha, v)d\alpha}, \\ w_p(\tau) &= & w_p(t_c) & e^{-\int_{t_c}^\tau v\sigma_A^2(\alpha, v)d\alpha}. \end{aligned}$$

As testifies the above expression, the weight modification can be expressed with respect to the weight of the particle at time t_c . The MC scheme by construction handles discontinuities of the cross-sections at the interface. In fact, what is important here is *consistency* between the samplings of the interaction time and the weight modification: the treatments must be made with the same hypothesis. For example, take the common approximations detailed in section 9.6.1 for the interaction time for σ_S , consistent approximations on σ_A must be made, i.e. $\sigma_A(\mathbf{x}, t, \mathbf{v}) = \sigma_A^{1,n}(v)\mathbf{1}_1(\mathbf{x}) + \sigma_A^{2,n}(v)\mathbf{1}_2(\mathbf{x})$. The resulting weight modification is given by

$$\begin{aligned} w_p(\tau) &= w_p(0) & e^{-v\sigma_A^{1,n}t_c}(v) & e^{-v\sigma_A^{2,n}(v)(\tau-t_c)}, \\ w_p(\tau) &= & w_p(t_c) & e^{-v\sigma_A^{2,n}(v)(\tau-t_c)}. \end{aligned} \tag{9.64}$$

On the fly computation strategies similar to the one presented in section 9.6.1 (cf. equation (9.60)) can also be applied for the weight modification. This is the case in [3] for example, and it will be studied in the next chapter 10.

Depending on the approximations chosen in the previous sections, the expressions of the weight modifications, samplings of the time interaction, energy and angle, for the different MC schemes can be considerably simplified. We insist *consistent* treatments for all the samplings must be made in order to ensure an MC convergence (i.e. depending only on N_{MC}).

9.7 Variance and moments of the MC schemes

In the previous sections, we described three¹⁴ converging MC schemes. Convergence implies they ensure obtaining asymptotically the same results for the mean, the first moment of the particle distribution. Obviously, the schemes differ (see the samplings). But it is hard *a priori* having any idea of their performances. The Central Limit theorem [256, 165] states that their performance differences can be expressed in term of convergence rate/variance¹⁵. They will also be compared *via* the physical regime they capture. The aim of this section is to give an idea of how we can choose an MC scheme having in mind a particular physical regime. For this, we study the asymptotic behaviours of the MC schemes with respect to the variance¹⁶ of the population of particles in a *monokinetic homogeneous configuration*: it corresponds to the case of an infinite medium with constant cross-sections with respect to time, space and energy. With these assumptions, the transport equation (9.1) resumes to

$$\partial_t \int u(t, \omega) d\omega + v\sigma_t \int u(t, \omega) d\omega = \sigma_s \int v P_s(\omega', \omega) u(t, \omega') d\omega'. \quad (9.65)$$

From the definition of P_s ensuring¹⁷ $\forall \omega \in \mathbb{S}^2, \int P_s(\omega', \omega) d\omega' = 1$, it even simplifies to the classical ODE

$$\partial_t U(t) + v\sigma_t U(t) = v\sigma_s U(t).$$

Its solution is $U(t) = U_0 e^{-v\sigma_a t}$ where $U(t) = \int u(t, \omega) d\omega$. In the following section, we verify the three MC schemes are converging for the mean solution $M_1(t) = U(t)$. We furthermore compute their asymptotical higher order moments. The latters will allow comparing their performances.

9.7.1 Asymptotic variance of the analog scheme (full_analog and multiplicity)

In order to compute the first two moments of any analog MC solution of (9.65), we have to come back to the expectation form of the transport equation from which the MC scheme is built. For the analog scheme, the recursive equation (9.14) simplifies to

$$U(t) = \mathbb{E} \left[\mathbf{1}_{[t, \infty]}(\tau) U_0 + \mathbf{1}_{[0, t]}(\tau) \sum_{r=0}^{N_R} U(t - \tau) \nu_r \delta_r(\mathcal{B}) \right]. \quad (9.66)$$

In the above expression, we recall $\tau \sim \mathcal{E}(v\sigma_t)$ and $\mathcal{B} \sim \mathcal{M}(r \in \{0, \dots, N_R\}, (\frac{\sigma_r}{\sigma_t})_{r \in \{0, \dots, N_R\}})$. Let us expand the recursive part into an infinite sum over the number of interactions. The first term in (9.66), with U_0 in factor, corresponds to the event 'there is no interaction between times 0 and t' . The second term of (9.66) corresponds to the event 'there is at least one interaction between times 0 and t' . We here introduce a new random variable, function of the already defined ones, $S_i = \sum_{k=0}^i \tau_k$ where $\tau_k \sim \mathcal{E}(v\sigma_t) \forall k \in \{1, \dots, i\}$ are independent identically distributed. It is well-known S_i follows a Gamma law of parameters $(v\sigma_t, i)$, denoted by $S_i \sim \Gamma(v\sigma_t, i)$ ¹⁸. Let us introduce X_t the stochastic process induced by the possible histories of any MC particles, it is given by

$$X_t = \sum_{k=0}^{\infty} U_0 \mathbf{1}_{[0, t]}(S_k) \mathbf{1}_{[t, \infty]}(S_k + \tau_{k+1}) \prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i) \right). \quad (9.67)$$

¹⁴Four if we count the 'full_analog' and the 'multiplicity' options of the analog scheme.

¹⁵the Central Limit theorem states that the variance (if obtained from an unbiased estimator, see [256]) is an error estimator.

¹⁶and even some high order moments for the semi-analog and non-analog schemes in sections 9.7.2–9.7.3.

¹⁷It only corresponds to a pretreatment of the cross-sections.

¹⁸this can be proven by first convoluting two pdfs of two exponential laws and then by recurrence.

In the above expression, k denotes the number of interactions encountered by any MC particles for times in $[0, t]$. The indicatrices

$$\mathbf{1}_{[0,t]}(S_k)\mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}),$$

express the fact an MC particle encounters exactly k interactions during the interval of time $[0, t]$. The product over i corresponds to the different scenarii/reactions a particle can encounter during any of these k interactions. In this section, we consider the 'multiplicity' option detailed in remark 9.1 of section 9.2. It consists in multiplying the weight of an MC particle enduring reaction i by ν_i . We choose here to study this option because it is less common than the 'full_analog' one (which is intensively studied [285, 200, 18, 144, 246, 199] in term of high order moments). The results for the 'full_analog' option are recalled in remark 9.2. The first moment of X_t , $M_1(t) = U(t)$, is defined by

$$M_1(t) = U(t) = \mathbb{E}[X_t] = \mathbb{E}\left[\sum_{k=0}^{\infty} U_0 \mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i)\right)\right]. \quad (9.68)$$

By linearity and independence of S_k with respect to the $(\mathcal{B}_i)_{i \in \{1, \dots, k\}}$ and of the $(\mathcal{B}_i)_{i \in \{1, \dots, k\}}$ two by two, the last expression becomes

$$\begin{aligned} M_1(t) = U(t) &= U_0 \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i)\right)\right], \\ &= U_0 \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1})\right] \mathbb{E}\left[\prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i)\right)\right], \\ &= U_0 \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1})\right] \prod_{i=1}^k \mathbb{E}\left[\left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i)\right)\right], \\ &= U_0 \sum_{k=0}^{\infty} \mathbb{P}(\tau_{k+1} > t - S_k | S_k < t) \left(\mathbb{E}\left[\left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B})\right)\right]\right)^k. \end{aligned} \quad (9.69)$$

By definition of $S_k \sim \Gamma(v\sigma_t, k)$ and $\tau_{k+1} \sim \mathcal{E}(v\sigma_t)$, the above conditional probability is equal to

$$\mathbb{P}(\tau_{k+1} > t - S_k | S_k < t) = e^{-v\sigma_t t} (v\sigma_t)^k \frac{s^k}{k!}. \quad (9.70)$$

The last expectation in (9.69) corresponds to the mean of the multinomial law \mathcal{B} :

$$\mathbb{E}\left[\left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B})\right)\right] = \sum_{r=0}^{N_R} \nu_r \frac{\sigma_r}{\sigma_t} = \frac{\sigma_s}{\sigma_t}.$$

Introducing the two previous expressions in $U(t)$ leads to

$$\begin{aligned} U(t) &= U_0 \sum_{k=0}^{\infty} e^{-v\sigma_t t} (v\sigma_t)^k \frac{t^k}{k!} \left(\frac{\sigma_s}{\sigma_t}\right)^k, \\ &= U_0 e^{-v\sigma_t t} \sum_{k=0}^{\infty} (v\sigma_s)^k \frac{t^k}{k!}, \\ &= U_0 e^{-v\sigma_a t}. \end{aligned} \quad (9.71)$$

With the few previous computations, we formally verified the convergence of the (multiplicity option of the) analog MC scheme for the mean of the stochastic process X_t . Let us now study the moment of

order 2 of the stochastic process X_t . It is defined as

$$M_2(t) = \mathbb{E}[X_t^2] = \mathbb{E} \left[U_0^2 \sum_{k,m=0}^{\infty} \frac{\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i) \right)}{\mathbf{1}_{[0,t]}(S_m) \mathbf{1}_{[t,\infty[}(S_m + \tau_{m+1}) \prod_{j=1}^m \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_j) \right)} \right]. \quad (9.72)$$

Using the fact that one MC particle has exactly k interactions, we have $\forall(k, m) \in \mathbb{N}^2$

$$\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \mathbf{1}_{[0,t]}(S_m) \mathbf{1}_{[t,\infty[}(S_m + \tau_{m+1}) = \delta_{k,m} \mathbf{1}_{[0,t]}(S_m) \mathbf{1}_{[t,\infty[}(S_m + \tau_{m+1}).$$

The independence of the random variables S_k, τ_{k+1} with respect to the random variables $(\mathcal{B}_i)_{i \in \{0, \dots, N_R\}}$ leads to

$$\begin{aligned} M_2(t) &= U_0^2 \mathbb{E} \left[\sum_{k=0}^{\infty} \mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i) \right) \prod_{j=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_j) \right) \right], \\ &= U_0^2 \sum_{k=0}^{\infty} \mathbb{P}(\tau_{k+1} > t - S_k | S_k < t) \mathbb{E} \left[\left(\prod_{i=1}^k \left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}_i) \right) \right)^2 \right], \\ &= U_0^2 \sum_{k=0}^{\infty} \mathbb{P}(\tau_{k+1} > t - S_k | S_k < t) \left(\mathbb{E} \left[\left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}) \right)^2 \right] \right)^k. \end{aligned} \quad (9.73)$$

The expectation is nothing more than the second moment of the multinomial law \mathcal{B} : we introduce the notation

$$\mathbb{E} \left[\left(\sum_{r=0}^{N_R} \nu_r \delta_r(\mathcal{B}) \right)^2 \right] = \sum_{r=0}^{N_R} (\nu_r)^2 \frac{\sigma_r}{\sigma_t} = \frac{\tilde{\sigma}_s}{\sigma_t}.$$

Plugging the expressions in (9.73) leads to

$$\begin{aligned} M_2(t) &= U_0^2 \sum_{k=0}^{\infty} e^{-v\sigma_t t} (v\sigma_t)^k \frac{s^k}{k!} \left(\frac{\tilde{\sigma}_s}{\sigma_t} \right)^k, \\ &= U_0^2 e^{-v\sigma_t t} \sum_{k=0}^{\infty} \frac{s^k}{k!} (v\tilde{\sigma}_s)^k, \\ &= U_0^2 e^{-v\sigma_t t + v\tilde{\sigma}_s t}. \end{aligned} \quad (9.74)$$

In term of variance, the analog scheme with the 'multiplicity' option verifies

$$\sigma_{\text{multiplicity}}^2(t) = U_0^2 \left[U_0^2 e^{-v\sigma_t t + v\tilde{\sigma}_s t} - e^{2(v\tilde{\sigma}_s - v\sigma_t)t} \right]. \quad (9.75)$$

Remark 9.2 In this section, we exhibited the asymptotic variance of the analog scheme with option 'multiplicity'. This option is much less¹⁹ studied than the 'full_analog' one for which the asymptotic variance may be found in many books and articles, see for example [285, 200, 199, 52]. We recall the full_analog scheme mimics the physics in the sense an MC particle can represent a physical particle. Applying both the material of papers [285, 200, 52] and the simplifications of this section leads to the following second moment for the analog MC scheme with 'full_analog' option:

$$M_2(t) = U_0^2 (2v\hat{\sigma}_s t + 1) e^{-2v\sigma_a t}. \quad (9.76)$$

¹⁹reasons will be given later in section 9.7.

In the above expression, $\hat{\sigma}_s$ is defined as

$$\hat{\sigma}_s = \sum_{k=0}^{N_R} \nu_k (\nu_k - 1) \sigma_k.$$

In such condition, the asymptotic variance is given by

$$\sigma_{full_analog}^2(t) = U_0^2 2v\hat{\sigma}_s t e^{-2v\sigma_a t}. \quad (9.77)$$

The variance of the full_analog scheme is of interest in many applications (see amongst others [200, 285, 246]) and is also referred to as the 'physical' variance. It is used for example in order to estimate the Feynman factor Y [222, 9, 18] (excess of variance) defined as

$$\frac{\sigma_{full_analog}^2(t)}{M_1(t)} = 1 + Y(t).$$

This physical variance may be interpreted as such: the higher $\sigma_{full_analog}^2(t)$ is, the less probable the particle population is to be close to the mean particle population $M_1(t)$. This is very well explained in [200] for example.

Comparisons between the different asymptotic variances will be made in section 9.7.4, we suggest first studying the two other schemes.

9.7.2 Asymptotic variance of the semi-analog scheme

Let us now come back to the expectation form of the transport equation obtained for the semi-analog scheme and compute the n^{th} order moment of any semi-analog MC solution of (9.65). With the simplifications detailed above, the recursive equation (9.22) becomes

$$U(t) = \mathbb{E} \left[\mathbf{1}_{[t,\infty]}(\tau) U_0 + \mathbf{1}_{[0,t]}(\tau) \frac{\sigma_s}{\sigma_t} U(t-\tau) \right]. \quad (9.78)$$

We recall we have $\tau \sim \mathcal{E}(v\sigma_t)$. We suggest expanding the recursive part into an infinite sum over the number of interactions. Let us introduce a new random variable $S_i = \sum_{k=0}^i \tau_k$ where $\tau_k \sim \mathcal{E}(v\sigma_t)$ $\forall k \in \{1, \dots, i\}$ are independent identically distributed. Once again S_i follows a Gamma law of parameters $(v\sigma_t, i)$, denoted by $S_i \sim \Gamma(v\sigma_t, i)$. Let us introduce X_t the stochastic process induced by the possible histories of any MC particles. It is given by

$$X_t = \sum_{k=0}^{\infty} \mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty]}(S_k + \tau_{k+1}) \left(\frac{\sigma_s}{\sigma_t} \right)^k U_0.$$

The indice k denotes the number of interactions encountered by any MC particles for times in $[0, t]$. The indicatrices

$$\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty]}(S_k + \tau_{k+1}),$$

express the fact an MC particle encounters exactly k interactions for times between $[0, t]$. The first moment is defined by $M_1(t) = U(t) = \mathbb{E}[X_t]$, and by linearity its expression becomes

$$\begin{aligned} M_1(t) = U(t) &= \mathbb{E}[X_t] = \mathbb{E} \left[\sum_{k=0}^{\infty} \mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty]}(S_k + \tau_{k+1}) \left(\frac{\sigma_s}{\sigma_t} \right)^k U_0 \right], \\ &= U_0 \sum_{k=0}^{\infty} \mathbb{E} [\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty]}(S_k + \tau_{k+1})] \left(\frac{\sigma_s}{\sigma_t} \right)^k, \\ &= U_0 \sum_{k=0}^{\infty} \mathbb{P}(\tau_{k+1} > t - S_k | S_k < t) \left(\frac{\sigma_s}{\sigma_t} \right)^k. \end{aligned} \quad (9.79)$$

Replacing the probability of having k interactions by its expression (9.70) leads to

$$\begin{aligned} U(t) &= U_0 \sum_{k=0}^{\infty} e^{-v\sigma_t t} (v\sigma_t)^k \frac{s^k}{k!} \left(\frac{v\sigma_s}{v\sigma_t} \right)^k, \\ &= U_0 e^{-v\sigma_t t} \sum_{k=0}^{\infty} (v\sigma_s)^k \frac{s^k}{k!}, \\ &= U_0 e^{-v\sigma_a t}. \end{aligned} \quad (9.80)$$

With the few previous computations, we formally verified the convergence of the semi-analog MC scheme for the mean of the stochastic process X_t . Let us now study the moment of order M of the stochastic process X_t . It is defined as $\mathbb{E}[X_t^M]$ with

$$X_t^M = U_0^M \sum_{i_1=0}^{\infty} \dots \sum_{i_M=0}^{\infty} \mathbf{1}_{[0,t]}(S_{i_1}) \mathbf{1}_{[t,\infty[}(S_{i_1} + \tau_{i_1+1}) \dots \mathbf{1}_{[0,t]}(S_{i_M}) \mathbf{1}_{[t,\infty[}(S_{i_M} + \tau_{i_M+1}) \left(\frac{\sigma_s}{\sigma_t} \right)^{i_1+\dots+i_M}.$$

In the previous expression, we expanded the exponent M into M summations over indices (i_1, \dots, i_M) . Using the generalization to M terms of the fact that $\forall (k, m) \in \mathbb{N}^2$, we have

$$\mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty[}(S_k + \tau_{k+1}) \mathbf{1}_{[0,t]}(S_m) \mathbf{1}_{[t,\infty[}(S_m + \tau_{m+1}) = \delta_{k,m} \mathbf{1}_{[0,t]}(S_m) \mathbf{1}_{[t,\infty[}(S_m + \tau_{m+1}),$$

we simplify the above expression of X_t^M into

$$X_t^M = U_0^M \sum_{i=0}^{\infty} \mathbf{1}_{[0,t]}(S_i) \mathbf{1}_{[t,\infty[}(S_i + \tau_{i+1}) \left(\frac{\sigma_s}{\sigma_t} \right)^{M \times i}.$$

Taking the expectation of X_t^M leads to

$$\begin{aligned} \mathbb{E}[X_t^M] &= U_0^M \sum_{i=0}^{\infty} \mathbb{P}(\tau_{i+1} > t - S_i | S_i < t) \left(\frac{\sigma_s}{\sigma_t} \right)^{M \times i}, \\ &= U_0^M e^{-v\sigma_t t} \sum_{i=0}^{\infty} (v\sigma_t)^i \frac{t^i}{i!} \left(\frac{\sigma_s}{\sigma_t} \right)^{M \times i}, \\ &= U_0^M \exp \left(\frac{(v\sigma_s)^M - (v\sigma_t)^M}{(v\sigma_t)^{M-1}} t \right). \end{aligned} \quad (9.81)$$

The latter expression is in agreement with the moment of order 1 and allows obtaining the asymptotic variance of the homogeneous process for the semi-analog scheme:

$$\sigma_{\text{semi-analog}}^2(t) = U_0^2 \left(e^{\frac{(v\sigma_s)^2 - (v\sigma_t)^2}{(v\sigma_t)} t} - e^{2(v\sigma_s - v\sigma_t)t} \right). \quad (9.82)$$

Let us now develop the same computations for the non-analog scheme before comparing the performances of the three schemes in section 9.7.4.

9.7.3 Asymptotic variance of the non-analog scheme

Let us apply the same methodology to the non-analog MC scheme and compute the M^{th} order moments of any non-analog-MC solution of (9.65). For this, we come back the expectation form of the transport equation from which the MC scheme is built. With the implications detailed above, the recursive equation (9.30) becomes

$$U(t) = \mathbb{E} \left[\mathbf{1}_{[t,\infty[}(\tau) U_0 e^{-v\sigma_a t} + \mathbf{1}_{[0,t]}(\tau) e^{-v\sigma_a(t-\tau)} U(t-\tau) \right]. \quad (9.83)$$

We recall $\tau \sim \mathcal{E}(v\sigma_s)$. Let us expand the recursive part into an infinite sum over the number of interactions thanks to $S_i = \sum_{k=0}^i \tau_k$ where $\tau_k \sim \mathcal{E}(v\sigma_s) \forall k \in \{1, \dots, i\}$ are independent identically distributed. Random variable S_i follows a Gamma law of parameters $(v\sigma_s, i)$, denoted by $S_i \sim \Gamma(v\sigma_s, i)$. Then (9.83), the equation for the mean (or the moment of order 1) rewrites

$$\begin{aligned} M_1(t) = U(t) &= \mathbb{E}[X_t] = \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbf{1}_{[0,t]}(S_k) \mathbf{1}_{[t,\infty]}(S_k + \tau_{k+1}) e^{-v\sigma_a t} U_0\right], \\ &= U_0 e^{-v\sigma_a t} \underbrace{\sum_{k=0}^{\infty} \mathbb{P}(\tau_{k+1} > t - S_k | S_k < t)}_{=1}, \\ &= U_0 e^{-v\sigma_a t}. \end{aligned} \quad (9.84)$$

We then recover the analytical solution of the homogeneous problem and formally verified the convergence of the non-analog scheme for the mean. Note that in this homogeneous configuration, the convergence of the non-analog scheme does not even depend on the probability measure of the interaction times τ_k, S_k as the sum over k always equals 1, whatever this choice. The interesting part concerns the moments of higher orders. Their computations are in fact very similar to the previous one, they are given by

$$\mathbb{E}[X_t^M] = U_0^M e^{-Mv\sigma_a t}.$$

The latter expression is in agreement with the moment of order 1 and allows showing the asymptotic variance of the homogeneous process for the non-analog scheme is given by:

$$\sigma_{\text{non-analog}}^2(t) = 0. \quad (9.85)$$

This property of the non-analog scheme is singular. Its interpretation in term of finite accuracy MC method (i.e. use of a finite number of MC particles as we recall the results are here obtained in the asymptotical limit $N_{MC} \rightarrow \infty$) will be studied in the next section.

9.7.4 Comparisons of the standard deviations of the MC schemes (homogeneous)

In this section, we briefly compare the four²⁰ previous MC schemes in term of asymptotical standard deviations (square root of the variance). The asymptotic standard deviations, in a homogeneous monokinetic configuration, are recalled here:

$$\begin{aligned} \sigma_{\text{multiplicity}}(t) &= U_0 e^{-v\sigma_a t} \sqrt{e^{-v\sigma_a t + v \left(\sum_{r=0}^{N_R} (\nu_r)^2 \sigma_r \right) t + 2v\sigma_a t} - 1}, \\ \sigma_{\text{full_analog}}(t) &= U_0 e^{-v\sigma_a t} \sqrt{\left(2v \sum_{k=0}^{N_R} \nu_k (\nu_k - 1) \sigma_k \right) t}, \\ \sigma_{\text{semi-analog}}(t) &= U_0 e^{-v\sigma_a t} \sqrt{e^{\frac{(v\sigma_s)^2 - (v\sigma_t)^2}{(v\sigma_t)} t + 2v\sigma_a t} - 1}, \\ \sigma_{\text{non-analog}}(t) &= 0. \end{aligned} \quad (9.86)$$

The study performed in this section is non exhaustive. It is only a pretext to briefly present how the analysis of standard deviations can help choose a particular MC schemes in a particular configuration.

The standard deviations are compared in figure 9.1 for a particular choice of the cross-sections and multiplicities. For both pictures of figure 9.1, $\sigma_0 = 0.3$, $\sigma_1 = 0.6$ and $\sigma_2 = 0.1$. The left picture presents the time evolutions of the standard deviations with $\nu_0 = 0, \nu_1 = 1, \nu_2 = 2$. The right picture shows the variations of the standard deviations with respect to ν_2 at time $t = 5$ for $\nu_0 = 0, \nu_1 = 1$. We begin by scheme to scheme comparisons: on figure 9.1 (left), it is easy verifying that $\forall t \in [0, 20]$, in this particular configuration,

$$\sigma_{\text{multiplicity}}^2(t) \geq \sigma_{\text{full_analog}}^2(t) \geq \sigma_{\text{semi-analog}}^2(t) \geq \sigma_{\text{non-analog}}^2(t).$$

If we focus on the 'multiplicity' option of the analog scheme, its standard deviation is higher than the

²⁰We distinguish the 'multiplicity' and the 'full_analog' options for the analog scheme here.

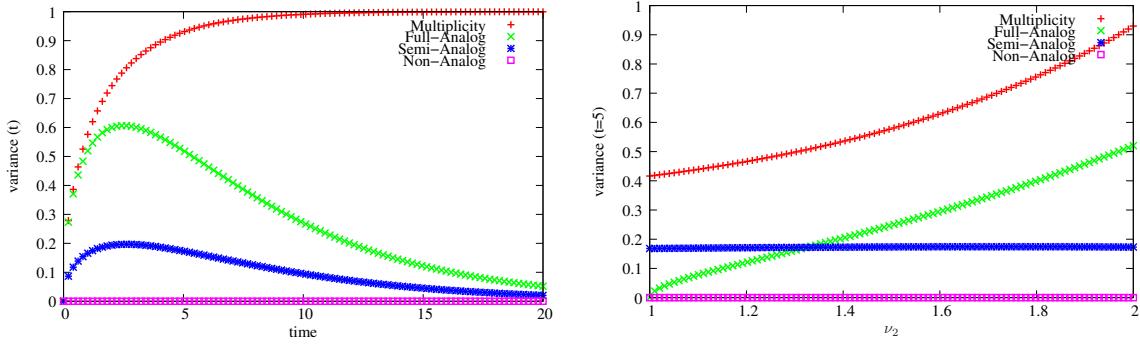


Figure 9.1: Comparison of the asymptotic standard deviations of the four presented MC schemes. Left: standard deviation with respect to time. Right: standard deviation at time $t = 5$ with respect to ν_2 .

ones of the other schemes, even higher than the one of the 'full_analog' one. This observation allows discarding the 'multiplicity' option as a relevant scheme:

- its standard deviation is different from the physical one²¹. Consequently it can not be used for computing and studying extinction probabilities etc. (see remark 9.2). Only its first moment M_1 is physically relevant.
- Its standard deviation, being different from the physical one, consequently only has a purely numerical interest. In this context, it exhibits a very poor convergence rate: it is higher than the other ones and even higher than the physical one.

For these two reasons, we can conclude the 'multiplicity' option has a very limited numerical and physical interest and is an example of bad MC scheme. It explains why it is scarcely used/studied in the literature. In fact, with this option, we mainly wanted to give the reader an example of a bad MC scheme together with how MC schemes can be compared and chosen/discard.

The full_analog scheme allows recovering the physical moments of the particle distribution, and not only the first one. On another hand,

- in very multiplicative media, it implies dealing with more and more MC particles. This can lead to intractable situations (memory consumption/explosion, unaffordable costs etc.).
- In very absorbing media, it implies dealing with a very small number of remaining MC particles leading to very important variances. It consequently needs intensive computations even if one is only interested in an accurate first order moment $M_1 = U$.

In many configurations of interest, the particle flows are fully characterised by their first order moment. In these cases, it is enough studying the mean of the particle distribution and consequently it is relevant looking for the scheme having the lowest (numerical) standard deviation. The semi-analog and the non-analog scheme both have, on the particular configuration of figure 9.1, standard deviations lower than the full_analog one. This makes them relevant alternatives in the limit of a high number of physical particles. Figure 9.1 (right) presents the standard deviations of the four schemes at $t = 5$ with respect to ν_2 , the multiplicity of the second reaction. We recall reaction 0 corresponds to absorption ($\nu_0 = 0$) and reaction 1 corresponds to diffusion ($\nu_1 = 1$). Once again, figure 9.1 (right) allows identifying the 'multiplicity' scheme as an inefficient one, having systematically a higher standard deviation than the physical one. For $1 \leq \nu_2 \leq 1.35$, the standard deviation of the semi-analog scheme is higher than the one of the full_analog one, lower for $1.35 \leq \nu_2$. On another hand, the standard deviation of the non-analog scheme is the lower one, equals to zero whatever the choices of the different reaction parameters $(\nu_r, \sigma_r)_{r \in \{0, \dots, N_R\}}$ in this monokinetic homogeneous configuration. This property of the non-analog scheme leads to the introduction of a new²² notion: *Asymptotic Preserving* (AP) scheme. It will be very useful all along the last sections and chapters of this document.

²¹obtained with the 'full_analog' option mimicing physics, see remark 9.2.

²²New in the document, not in the literature.

Definition 9.1 A converging numerical scheme of discretisation parameter $\mathcal{O}(\Delta)$ is Asymptotic Preserving in an identified regime of interest characterised by $\delta \rightarrow 0$, if its convergence rate weakly depends on Δ in this regime $\delta \rightarrow 0$. Another way to understand this definition is that the error is $\mathcal{O}(\Delta) = K_\delta \Delta$ with

$$K_\delta \underset{\delta \rightarrow 0}{\ll} 1, \text{ or at least } K_\delta = K,$$

but K_δ does not explode with $\delta \rightarrow 0$.

The above definition may appear foggy at this stage of the discussion but is also general enough to be reused (section 9.9 and chapter 10). For example, Δ may refer to $\frac{1}{\sqrt{N_{MC}}}$ for an MC scheme or $\Delta \mathbf{x}$ or Δt once a coupling is taken into account (see chapter 10). The regime of interest is characterised by $\delta \rightarrow 0$ with δ depending on the characteristic time, length, mean free path etc. of the problem. In order to clarify the above definition, we suggest emphasizing the AP character of the non-analog scheme in the monokinetic homogeneous regime. To do so, it is helpful to non-dimensionalize the monokinetic linear Boltzmann equation. Let us introduce

$$\begin{cases} \mathbf{x} = \mathbf{x}^* \mathcal{X}, \mathbf{v} = \mathbf{v}^* \mathcal{V}, t = t^* \mathcal{T}, \\ \sigma_\alpha = \sigma_\alpha^* \frac{1}{\lambda_\alpha}, \forall \alpha \in \{s, t, a\}, \end{cases} \quad (9.87)$$

where the superscript * denotes a nondimensional quantity. Let us introduce $u^*(\mathbf{x}^*, t^*, \omega) = u(\mathbf{x}, t, \omega)$, then

$$\frac{1}{\mathcal{T}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) = \partial_t u(\mathbf{x}, t, \omega), \quad \frac{1}{\mathcal{X}} \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) = \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega).$$

Using the above expressions in the transport equation yields

$$\frac{\mathcal{X}}{\mathcal{T}\mathcal{V}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_t^* \frac{\mathcal{X}}{\lambda_t} u^*(\mathbf{x}^*, t^*, \omega) = v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} \int u^*(\mathbf{x}^*, t^*, \omega) d\omega.$$

Let us decompose $\sigma_t = \sigma_a + \sigma_s$ to obtain

$$\begin{aligned} \frac{\mathcal{X}}{\mathcal{T}\mathcal{V}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) &+ v^* \sigma_a^* \frac{\mathcal{X}}{\lambda_a} u^*(\mathbf{x}^*, t^*, \omega) \\ &+ v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} u^*(\mathbf{x}^*, t^*, \omega) = v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} \int u^*(\mathbf{x}^*, t^*, \omega) d\omega. \end{aligned}$$

Now suppose $\frac{\mathcal{X}}{\mathcal{T}\mathcal{V}} = \mathcal{O}(\frac{1}{\delta}) = \frac{\mathcal{X}}{\lambda_a}$ and $\frac{\mathcal{X}}{\lambda_s} = \mathcal{O}(1)$, we have (we drop the superscript * for convenience)

$$\begin{aligned} \frac{1}{\delta} \partial_t u(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) &+ v \sigma_a \frac{1}{\delta} u(\mathbf{x}, t, \omega) \\ &+ v \sigma_s u(\mathbf{x}, t, \omega) = v \sigma_s \int u(\mathbf{x}, t, \omega) d\omega. \end{aligned}$$

Performing a Hilbert development, i.e. $u = u^0 + u^1 \delta + u^2 \delta^2 + \mathcal{O}(\delta^3)$ see [143], and considering only the first order (i.e. u^0) leads to

$$\partial_t u^0 = -v \sigma_a u^0.$$

It corresponds to the monokinetic homogeneous regime as $\delta \rightarrow 0$. In this regime, the non-analog scheme asymptotically ensures $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}}) = \frac{\sigma_{\text{non-analog}}(t)}{\sqrt{N_{MC}}} \ll 1$, whatever the number of MC particles, as asymptotically $\sigma_{\text{non-analog}}(t) = 0$, see (9.85). In practice, in such homogeneous configurations, if one performs a convergence study with respect to N_{MC} comparing the numerical solution to the analytical one $M_1(t)$, the study will exhibit an $\mathcal{O}(\frac{1}{N_{MC}})$ convergence rate instead of an $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$. This has been experimentally put forward in a study of [3].

Finally, the previous expressions for the standard deviations are also interesting for the developer willing to verify his implementations. Making sure the implementation of the MC scheme allows recovering the standard deviation and the high order moments up to numerical accuracy $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$ in several regimes can be a relevant and discriminating tests from a verification point of view (see V&V for Verification and Validation, see [13]). Indeed, the higher the order of the moment is, the more sensitive it is with respect to any inaccuracy, see sections 3.4.2–3.4.3 of chapter 3 of part II.

9.8 A general canvas for developing MC schemes

All along the previous sections, several MC schemes, direct or adjoint, together with their characteristics have been presented. Each scheme is interesting (except maybe the 'multiplicity' option, discarded in section 9.7.4) in the sense each scheme allows capturing more or less accurately different regimes or present development advantages (extinction probability, homogeneous case, low asymptotic variance, minimal amount of developments for adjoint/direct resolutions etc.). For these reasons, one may wish to be able to choose the relevant scheme in the relevant situation *with the same simulation device*. In this section, we insist on the fact that every one of these schemes can fit in the same canvas. This is of practical interest especially when developing a simulation code (a platform) in order to mutualize the more possible parts of the resolution schemes. In order to illustrate the matter, we rewrite in the same canvas

- the three MC schemes of sections 9.2–9.3–9.4,
- for a direct resolution,
- on a given grid $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$, with the classical approximations presented in section 9.6 with respect to space and time discretisation of the cross-sections (i.e. constant in each cell and in each time step),
- taking into account explicitly in the chart the fact that an MC particle goes from one cell to another.

With algorithm 9, detailed and commented later on in this section, we hope the reader willing to develop an MC simulation code in order to solve the linear Boltzmann equation with constant cross-sections in each cell and time step can follow the description below and achieve its purpose.

In this section, we deal with the direct resolution of the linear Boltzmann equation. As already briefly tackled in section 9.5.2 of this document, the direct resolution needs a first step, a pretreatment, ensuring the initial population of MC particles accurately represents the initial condition u_0 . This step is technical and crucial. The sampling phase resumes once again to identifying a set probability measures and, exactly as in the resolution phase, this set is not unique. In the next section, we detail some possibilities, the list is non exhaustive.

9.8.1 Sampling the initial MC particle population

The first step of a direct resolution corresponds to what is commonly called the *sampling phase*. This phase was hinted at in algorithm 5 in its most general form. It implies correlated samplings of the initial position \mathbf{x} , velocity \mathbf{v} from a quite complex probability measure

$$\begin{aligned} du_0(\mathbf{x}, \mathbf{v}) &= \frac{1}{\frac{1}{|\mathcal{D}|} \iint_{\mathcal{D}} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}, \\ &= \frac{1}{U_0} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}. \end{aligned} \tag{9.88}$$

At the end of such phase, the population of (N_{MC}) MC particles is supposed to *statistically accurately represent* the initial condition $u_0(\mathbf{x}, \mathbf{v})$. It means that we demand the initial MC particle population to verify

$$\sum_{p=1}^{N_{MC}} u_p(\mathbf{x}, 0, \mathbf{v}) \underset{N_{MC} \rightarrow \infty}{=} u_0(\mathbf{x}, \mathbf{v}), \forall \mathbf{x} \in \mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i, \mathbf{v} \in \mathbb{R}^3, \text{ and } \forall N_{\mathbf{x}} \in \mathbb{N}^*. \tag{9.89}$$

The important point in (9.89) concerns the grid and the condition $\forall N_{\mathbf{x}} \in \mathbb{N}^*$: we want the convergence of the initial MC particle population not to depend on the size of the grid N_x to ensure a convergence depending only on N_{MC} . In the following lines, we detail two possibilities to ensure the above property:

- The first one corresponds to the ‘MC solution’ in the sense it allows having uniform initial weights $w_p(0) = \frac{U_0}{N_{MC}}$, $\forall p \in \{1, \dots, N_{MC}\}$ for every MC particles. The other fields $(\mathbf{x}_p(0), \mathbf{v}_p(0), \mathbf{v}_p(0))$ are sampled from (9.88).
- The second, in opposition, would be closer to a ‘quadrature rule solution’ (see part II, chapter 5). It begins by choosing arbitrarily the distribution of the fields $(\mathbf{x}_p(0), \mathbf{v}_p(0), \mathbf{v}_p(0))$ of the MC particles before correcting consistently their weights $w_p(0)$.

Both methods are described and applied in a simple configuration at the end of this section. Care will be taken to put forward the particularities of the sampling strategies.

From a practical point of view, the initial sampling phase is almost the inverse of what is presented in part II regarding uncertainty quantification. In part II and especially in chapter 5, the idea is to determine/approximate u_0 given a particular discretisation of \mathbf{x}, \mathbf{v} . In this section, the problem is, knowing u_0 , discretise consistently \mathbf{x}, \mathbf{v} . All along the next paragraph, our aim is to determine the different fields $w_p(0), \mathbf{x}_p(0), \mathbf{v}_p(0)$ of any MC particle p in order to ensure that

$$\sum_{p=1}^{N_{MC}} u_p(\mathbf{x}, 0, \mathbf{v}) \stackrel{a.s.}{=} u_0(\mathbf{x}, \mathbf{v}), \quad (9.90)$$

The overset a.s. is for *almost surely*, see part II. For convenience in the following lines, we use the abusive notations $w_p(0) = w_p, \mathbf{x}_p(0) = \mathbf{x}_p, \mathbf{v}_p(0) = \mathbf{v}_p$. The first paragraph concerns how the MC particles are initially distributed amongst some cells $i \in \{1, \dots, N_{\mathbf{x}}\}$ tesselating the simulation domain $\mathcal{D} = \bigcup_{i=1}^{N_{\mathbf{x}}} \mathcal{D}_i$. The two following ones correspond to the descriptions of the two above solutions (MC and quadrature rule) described within a given cell \mathcal{D}_i .

Sampling i_p , the initial cell of p amongst the different cells $i \in \{1, \dots, N_{\mathbf{x}}\}$

Let us consider particle p . It is common to first determine the initial cell i_p of particle p before determining its other fields. We could have determined \mathbf{x}_p first but finding i_p afterward would have involved a loop over cells and MC particles to determine i_p from \mathbf{x}_p , hence possible performance losses (increasing with the number of MC particle and the number of cells). Sampling the initial cell i_p resumes to sampling from a discrete law. We introduce

$$U_0 = \frac{1}{|\mathcal{D}|} \iint_{\mathcal{D}} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} = \sum_{i=1}^{N_{\mathbf{x}}} U_0^i.$$

The quantity U_0 is the initial amount of particles in the whole simulation domain \mathcal{D} . The quantities $(U_0^i)_{i \in \{1, \dots, N_{\mathbf{x}}\}}$ are the amounts of particles in each cells \mathcal{D}_i of the simulation domains $\mathcal{D} = \bigcup_{i=1}^{N_{\mathbf{x}}} \mathcal{D}_i$ defined by

$$\forall i \in \{1, \dots, N_{\mathbf{x}}\}, U_0^i = \frac{1}{|\mathcal{D}_i|} \iint_{\mathcal{D}_i} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}.$$

Algorithm 6: Sampling of the initial cell i_p

```

1 Function sample_cell()
2   #More generally, this kind of function allows sampling from any discrete probability measure
   with  $i \in \{1, \dots, N_x\}$  states of probability  $\frac{U_0^i}{U_0}$ .
3    $\mathcal{U} = \text{sample\_uniform\_law}()$ 
4   set  $Proba = 0$ 
5   set  $i_p = N_x$ 
6   for  $i \in \{1, \dots, N_x - 1\}$  do
7      $Proba \leftarrow Proba + U_0^i$ 
8     if  $Proba > \mathcal{U} \times U_0$  then
9        $i_p \leftarrow i$ 
10      break;
11    end
12  end
13  return  $i_p$ 

```

With the previous notations, the probability of having a particle in cell i is $\frac{U_0^i}{U_0}$. The initial cell i_p of particle p is sampled from the following discrete probability measure

$$d\mathcal{P}_{\mathcal{I}}(I) = \sum_{i=1}^{N_x} \frac{U_0^i}{U_0} \delta_i(I). \quad (9.91)$$

Practically, this resumes to sampling from a uniform law in $[0, 1]$, $\mathcal{U}_{\mathcal{I}} \sim \mathcal{U}([0, 1])$, and inverse the piecewise constant cdf of (9.91). The initial cell i_p then verifies

$$i_p = \min_{h \in \{1, \dots, N_x\}} \left\{ \mathcal{U}_{\mathcal{I}} < \sum_{i=0}^h \frac{U_0^i}{U_0} \right\}. \quad (9.92)$$

Algorithm 6 presents the operations to obtain i_p as in (9.92). Once the cell \mathcal{D}_{i_p} identified, it remains to sample the position and velocity of particle p within cell i_p (i.e. conditionally to being in cell i_p). There are several ways to do so, two different ones are presented in the two next paragraphs. From now on, we suppose the initial cell i_p known. In order to alleviate the notations, the indice i_p for cell \mathcal{D}_{i_p} is omitted, i.e. we determine $w_p, \mathbf{x}_p, \mathbf{v}_p$ within cell \mathcal{D} . In other words, the following sampling strategies can be applied even without the introduction of a grid for \mathcal{D} .

The 'MC solution': uniformly distributed weights and consistently sampled fields

Suppose now the cell \mathcal{D} in which p is sampled identified, it remains to determine $w_p, \mathbf{x}_p, \mathbf{v}_p$ consistently with probability measure

$$du_0(\mathbf{x}, \mathbf{v}) = \frac{1}{\frac{1}{|\mathcal{D}|} \iint_{\mathcal{D}} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} = \frac{1}{U_0} u_0(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}. \quad (9.93)$$

The structure of (9.93) has already been encountered in section 9.6.2. It must be compared to the expression of $dP_s = P_s(\mathbf{v}, \mathbf{v}') d\mathbf{v}'$, from (9.25) for example, for the correlated samplings of the scattering energy and angle. The sampling strategy described for dP_s also applies to du_0 : it is possible to perform successive conditional samplings from the marginals of du_0 . The order of the samplings of the different fields can be arbitrary. In the following paragraphs, we first sample \mathbf{x}_p , then \mathbf{v}_p conditionally to being

at position \mathbf{x}_p . Those successive samplings are described below.

Algorithm 7: Sampling step to represent the initial condition $u_0(\mathbf{x}, \mathbf{v})$ with N_{MC} particles.

```

1 Function sampling( $N_{MC}$ )
2   for  $i \in \{1, \dots, N_x\}$  do
3     | set  $N_{MC}^i = 0$ 
4   end
5   for  $p \in \{1, \dots, N_{MC}\}$  do
6     #Sample the cell  $\mathcal{D}_{i_p}$  of the MC particle  $p$  according to (9.92) and algorithm 6
7     set  $i_p = \text{sample\_cell}()$ 
8     do  $N_{MC}^{i_p} \leftarrow N_{MC}^{i_p} + 1$ 
9     #Sample the position  $\mathbf{x}_p$  of particle  $p$  in cell  $i_p$  according to (9.94)
10    set  $\mathbf{x}_p = \text{sample\_position.in\_cell}(i_p)$ 
11    #Sample the velocity  $\mathbf{v}_p$  of particle  $p$  at position  $\mathbf{x}_p$  according to (9.95)
12    set  $\mathbf{v}_p = \text{sample_velocity_at\_position}(\mathbf{x}_p)$ 
13  end
14  #Once the number of particles per cell  $N_{MC}^i$  known  $\forall i \in \{1, \dots, N_x\}$ 
15  #Renormalization
16  for  $p \in \{1, \dots, N_{MC}\}$  do
17    #Set the weights of the particle to ensure exactly  $\sum_{p \in \mathcal{D}_i} w_p = U_0^i$ 
18    set  $w_p \leftarrow \frac{U_0^{i_p}}{N_{MC}^{i_p}}$ 
19  end
20 return A population of MC particles representing almost surely  $u_0(\mathbf{x}, \mathbf{v})$ 
```

Let us introduce the marginal probability measure $U_0(\mathbf{x})d\mathbf{x}$ defined by

$$U_0(\mathbf{x})d\mathbf{x} = \frac{1}{U_0} \left[\int u_0(\mathbf{x}, \mathbf{v})d\mathbf{v} \right] \frac{\mathbf{1}_{\mathcal{D}}(\mathbf{x})}{|\mathcal{D}|} d\mathbf{x}.$$

Introduce furthermore a new sample from a uniform law $\mathcal{U}_{\mathcal{X}} \sim \mathcal{U}([0, 1])$. Then the sampled initial position \mathbf{x}_p satisfies

$$\mathcal{U}_{\mathcal{X}} = \int_{-\infty}^{\mathbf{x}_p} U_0(\mathbf{x})d\mathbf{x}. \quad (9.94)$$

It only remains to inverse the above expression in order to obtain explicitly the position \mathbf{x}_p of the particle in the cell \mathcal{D} . The velocity \mathbf{v}_p is then sampled conditionally to being at position \mathbf{x}_p . Introducing the conditional probability measure

$$U_0^{\mathbf{x}_p}(\mathbf{v})d\mathbf{v} = \frac{1}{U_0(\mathbf{x}_p)} u_0(\mathbf{x}_p, \mathbf{v})d\mathbf{v}, \quad \mathcal{U}_{\mathcal{V}} \sim \mathcal{U}([0, 1]), \quad \mathcal{U}_{\mathcal{V}} = \int_{-\infty}^{\mathbf{v}_p} U_0^{\mathbf{x}_p}(\mathbf{v})d\mathbf{v}. \quad (9.95)$$

Suppose the above probability measure is known, then the initial velocity of the particle is once again obtained inverting its cdf as done with the position \mathbf{x}_p .

Once every samplings done for particle p , it only remains to set its weight w_p . The above process ensures equiprobable MC particles, i.e. all MC particles must have the same weight within the same cell for consistency. At the end of the sampling phase, during a *renormalization* step once we know the number of MC particles N_{MC}^i in each cell $i \in \{1, \dots, N_x\}$, the weight w_p is set to

$$w_p = \frac{U_0^i}{N_{MC}^i}, \quad \forall p \in \mathcal{D}_i. \quad (9.96)$$

Such process ensures, by construction, the population of particles represents the initial condition asymptotically with $N_{MC} \rightarrow \infty$. To verify it, it is enough verifying every moments of the cell population tend to the moments of u_0 . In fact, it is even exact for the first moment (thanks to the renormalization step).

The higher order moments are approximated up to $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}^i}}\right)$ accuracy. In practice, the different presented marginal probability measures may be complex to inverse. We usually have resort to approximations (linear, logarithmic etc.), converging with respect to the size of the cell $|\mathcal{D}|$: if an approximation is introduced, the convergence in (9.90) may depend, strongly or not, on $\mathcal{O}\left(\max_{i \in \{1, \dots, N_x\}} |\mathcal{D}_i|\right)$. As a result, a compromise between the simplicity of the cdf inversions and the cell size must be made. The algorithmic description for the MC sampling phase in order to represent the initial condition is given in algorithm 7. A practical example applying the above strategy is presented at the end of the section.

In the next paragraph, we present another way to make sure the cell population statistically describes the initial condition: it does not involve inverting (potentially complex) cdfs.

The 'quadrature rule solution': uniformly distributed fields and weight corrections

We still assume we already determined the cell in which the MC particle is sampled. We here describe another way to make sure the MC population of the cell almost surely represent the initial condition. This second method, closer to a quadrature rule than an MC discretisation (see section 5.2.3), allows avoiding the potentially complex cdf inversions of the previous paragraph. It supposes $\mathbf{x}_p, \mathbf{v}_p$ are sampled independently according to known probability measures. These probability measures can be arbitrary. In practice, it is common choosing the uniform ones with respect to space, energy and angle, i.e. respectively $\frac{\mathbf{1}_{\mathcal{D}_i}(\mathbf{x})}{|\mathcal{D}_i|} d\mathbf{x}$, $\mathbf{1}_{\mathbb{R}^+}(v) dv$, $\mathbf{1}_{\mathbb{S}^2}(\omega) d\omega$. The samplings are then followed by a consistent correction of the weights in order to ensure the convergence of the MC population toward u_0 almost surely. Note that the notation $\mathbf{1}_{\mathbb{R}^+}(v) dv$ for the sampling of the energy field may appear unconventional. This is done on purpose for the sake of genericity as depending on the physics of interest, the measure $\mathbf{1}_{\mathbb{R}^+}(v) dv$ may be different. Of course, we suppose it sums up to 1, i.e. $\int_{\mathbb{R}^+} dv = 1$.

With the previous choice, $\mathbf{x}_p, \mathbf{v}_p, \omega_p$ are sampled *independently and uniformly* according to the (simple to inverse) cdfs:

$$\begin{cases} \mathcal{U}_{\mathcal{X}} \sim \mathcal{U}([0, 1]), & \mathcal{U}_{\mathcal{X}} = \int_{-\infty}^{\mathbf{x}_p} \frac{\mathbf{1}_{\mathcal{D}_i}(\mathbf{x})}{|\mathcal{D}_i|} d\mathbf{x}, \\ \mathcal{U}_{\mathcal{V}} \sim \mathcal{U}([0, 1]), & \mathcal{U}_{\mathcal{V}} = \int_{-\infty}^{v_p} \mathbf{1}_{\mathbb{R}^+}(v) dv, \\ \mathcal{U}_{\mathcal{W}} \sim \mathcal{U}([0, 1]), & \mathcal{U}_{\mathcal{W}} = \int_{-\infty}^{\omega_p} \mathbf{1}_{\mathbb{S}^2}(\omega) d\omega. \end{cases} \quad (9.97)$$

Once the position, energy and angle obtained, it remains to make sure the weights w_p ensure the convergence property: setting the weight to $w_p = u_0(\mathbf{x}_p, \mathbf{v}_p, \omega_p)$ induces an $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}^i}}\right)$ convergence. It is common to correct it during a *renormalization* phase, once N_{MC}^i is known, in order to enforce exactness for the first moment, i.e. such that $\sum_{p \in \mathcal{D}_i} w_p = U_0^i$. The above procedure is detailed in algorithm 8.

The latter sampling strategy is simpler than the one presented in the previous paragraph and is probably the most common one. It still has an important drawback: suppose there exists a volume $\mathbb{V} \subset \mathcal{D} \times \mathbb{R}^3$ in cell \mathcal{D} such that

$$\iint_{\mathbb{V}} u_0(\mathbf{x}, \mathbf{v}) \frac{\mathbf{1}_{\mathcal{D}}(\mathbf{x})}{|\mathcal{D}|} d\mathbf{x} d\mathbf{v} = 0.$$

If this volume \mathbb{V} is too important, many $(\mathbf{x}_p, \mathbf{v}_p)_{p \in \{1, \dots, N_{MC}\}}$ can be sampled in \mathbb{V} and many particles would see their weights assigned to zero for consistency. In practice, to avoid such case, a rejection method is applied (see algorithm 8) but the performance of the sampling phase may suffer from it. In fact, the larger the size $|\mathbb{V}|$ of volume \mathbb{V} and the number of dimensions of the problem (i.e. $3(\mathbf{x}) + 1(v) + 2(\omega) = 6$ here), the more the rejection method is inefficient, see [161] for a pedagogical example. The latter sampling strategy, with and without rejection, is applied in a simplified configuration at the end of this section.

Of course, it is possible to mix both of the previously described methods and obtain a good compro-

mise. This is in fact what is commonly done. We also insist on the fact that in practice, approximations can be made without strong consequences on the convergence rate of the MC resolution on the observables of interest. For example, it is common to sample uniformly the fields $\mathbf{x}_p, \mathbf{v}_p$ and set uniform weights equal to $\frac{U_0^i}{N_{MC}^i}$ to the particles of cell i . This induces a $\mathcal{O}(\max_i |\mathcal{D}_i|)$ error which can be negligible in comparison to the $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$ error of the resolution scheme. If it is not, it is still always possible to increase $N_{\mathbf{x}}$ to obtain the result. Nevertheless, care has to be taken so that the constant in the $\mathcal{O}(\max_i |\mathcal{D}_i|)$ error is not too important: the property may not hold for example in photonic applications in which the *source sampling* is critical (teleportation error [301]). More details are given in section 9.9.1 and later in section 10.2.1 of chapter 10.

In the next paragraph, we apply the above material on a simple initial condition u_0 . The aim is to illustrate the points tackled above.

Algorithm 8: Sampling step to represent the initial condition $u_0(\mathbf{x}, \mathbf{v})$ with N_{MC} particles.

```

1 Function sampling( $N_{MC}$ )
2   for  $i \in \{1, \dots, N_{\mathbf{x}}\}$  do
3     set  $N_{MC}^i = 0$ 
4     set  $U_0^{i,N_{MC}^i} = 0$ 
5   end
6   for  $p \in \{1, \dots, N_{MC}\}$  do
7     #Sample the cell  $\mathcal{D}_{i_p}$  of the MC particle  $p$  according to (9.92)
8     set  $i_p = \text{sample\_cell}()$ 
9     set  $w_p = 0$ 
10    #Rejection method
11    while  $w_p == 0$  do
12      #Sample the position  $\mathbf{x}_p$  of particle  $p$  in cell  $i_p$  according to (9.97)
13      set  $\mathbf{x}_p = \text{sample\_unif\_position\_in\_cell}(i_p)$ 
14      #Sample the velocity  $\mathbf{v}_p$  of particle  $p$  at position  $\mathbf{x}_p$  according to (9.97)
15      set  $\mathbf{v}_p = \text{sample\_unif\_velocity}()$ 
16      #Set the initial weight of particle  $p$  according to (9.96)
17      set  $w_p = u_0(\mathbf{x}_p, \mathbf{v}_p)$ 
18      if  $w_p \neq 0$  then
19        do  $U_0^{i,N_{MC}^i} \leftarrow U_0^{i,N_{MC}^i} + w_p$ 
20        do  $N_{MC}^{i_p} \leftarrow N_{MC}^{i_p} + 1$ 
21      end
22    end
23  end
24  #Once the number of particles per cell  $N_{MC}^i$  known together with  $U_0^{i,N_{MC}^i}$ ,  $\forall i \in \{1, \dots, N_{\mathbf{x}}\}$ 
25  #Renormalization
26  for  $p \in \{1, \dots, N_{MC}\}$  do
27    #Correct the weights of the particle to ensure exactly  $\sum_{p \in \mathcal{D}_i} w_p = U_0^i$ 
28    set  $w_p \leftarrow w_p \times \frac{1}{U_0^{i_p,N_{MC}^{i_p}}} \frac{U_0^i}{N_{MC}^{i_p}}$ 
29  end
30 return A population of MC particles representing  $u_0(\mathbf{x}, \mathbf{v})$ 
```

Example of initial MC sampling

We consider a simple example and illustrate the two sampling methods briefly described above. We focus on the sampling in a 1-dimensional cell $\mathcal{D} = [-1, 1]$. Without loss of generality, we consider an isotropic and monokinetic initial condition, i.e. we assume $u_0(\mathbf{x}, \mathbf{v}) = u_0(x)$. It simplifies the computations and

remains relevant. Besides, we suppose

$$u_0(x) = \mathbf{1}_{[-a,a]}(x) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma}}, \quad (9.98)$$

i.e. u_0 has a truncated gaussian (mean 0, variance σ^2) form with $[-a, a] \subset \mathcal{D} = [-1, 1]$. Figure 9.2 presents the results obtained with the previous sampling methods for initial condition (9.98) (reference in red) for different values of parameters σ and a . These parameters allow controlling the size of volume \mathbb{V} (especially thanks to a) together with the range of probable positions within the cell (especially thanks to σ). The smaller a is, the larger $|\mathbb{V}|$ is. The larger σ is, the more the initial condition tends to a uniform sampling in interval $[-a, a]$.

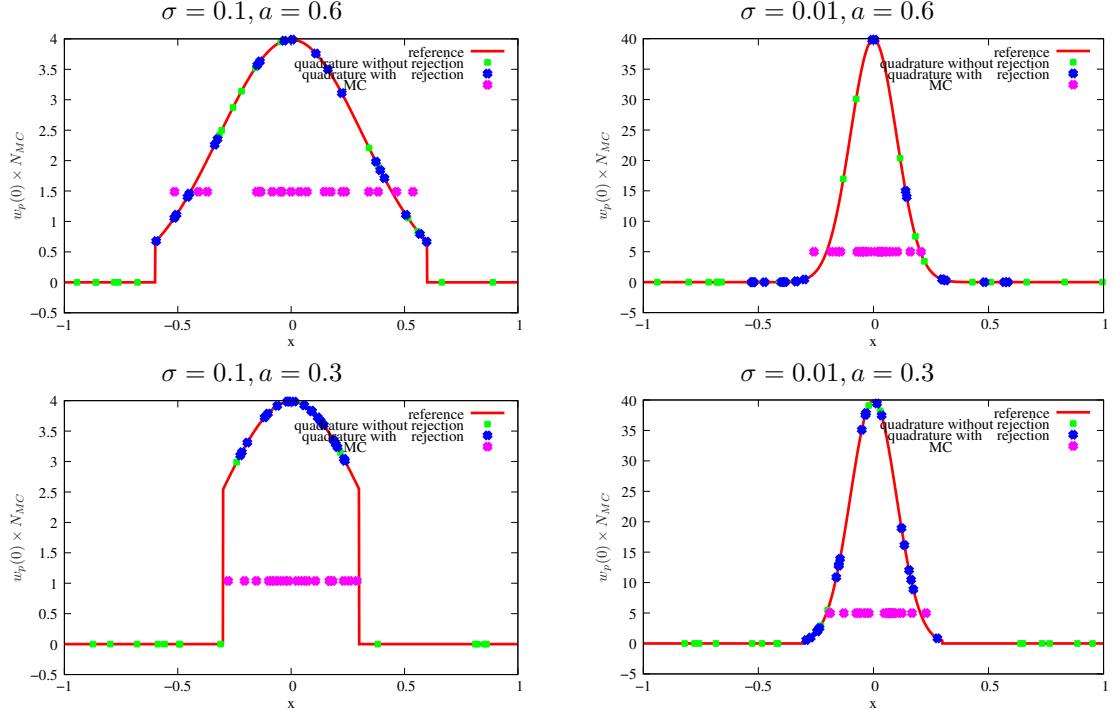


Figure 9.2: Comparison of the behaviours of the different kinds of initial samplings for u_0 as in (9.98) for $\sigma = 0.1, a = 0.6$ (top left), $\sigma = 0.01, a = 0.6$ (top right), $\sigma = 0.1, a = 0.3$ (bottom left), $\sigma = 0.01, a = 0.3$ (bottom right). The red curve is the reference (9.98). The dots are obtained with different samplings (with $N_{MC} = 20$ here). The dots are not exactly the weights of the MC particles but the weights multiplied by N_{MC} , i.e. $w_p(0) \times N_{MC}$. The green dots are obtained with the 'quadrature without rejection' method, the blue dots with the 'quadrature with rejection' one and the magenta are obtained with the MC one.

In figure 9.2, three initial sampling strategies are compared. The 'MC solution', the 'quadrature' one and the 'quadrature' one with rejection. We briefly go through few implementation details before commenting the results:

- Let us begin with the 'MC solution', the initial sampling ensuring uniform weights. In order to apply this method, we first need to build the equivalent of (9.93) based on (9.98). Integrating (9.98) on \mathcal{D} leads to

$$U_0 = \operatorname{erf}\left(\frac{1}{2}\sqrt{\frac{2}{\sigma}}a\right), \text{ so that the probability measure is given by } du_0 = \frac{1}{\sqrt{2\pi}\sigma\operatorname{erf}\left(\frac{1}{2}\sqrt{\frac{2}{\sigma}}a\right)} e^{-\frac{x^2}{2\sigma}} dx.$$

Its cdf is given by $\forall x \in [-a, a]$

$$F_{u_0}(x) = \frac{1}{2} + \frac{1}{2} \frac{\operatorname{erf}\left(\frac{1}{2}\sqrt{\frac{2}{\sigma}}x\right)}{\operatorname{erf}\left(\frac{1}{2}\sqrt{\frac{2}{\sigma}}a\right)}.$$

The position x_p of the MC particles can be sampled introducing $\mathcal{U}_{\mathcal{X}} \sim \mathcal{U}([0, 1])$ and inverting $\mathcal{U}_{\mathcal{X}} = F_{u_0}(x_p)$. In practice, we rely on a Newton algorithm in order to perform the inversion and obtain the magenta dots of figure 9.2. By construction, the magenta dots are within interval $[-a, a]$. The weights of the MC particles are the same (note that we display $w_p \times N_{MC}$ for a better readability of the picture mainly). There are more MC particles in the area of higher probability, i.e. in the vicinity of $x = 0$, than elsewhere.

- The 'quadrature rule solution' (with or without rejection) is much simpler to describe. It consists in sampling the MC particle position x_p uniformly within $\mathcal{D} = [-1, 1]$ and setting the weights of the MC particles to $w_p = \frac{u_0(x_p)}{N_{MC}}$. The rejection only supposes that if $w_p = 0$, x_p is resampled according to the same probability measure (here uniformly within $[-1, 1]$) until $w_p \neq 0$, see algorithm 8. The quadrature sampling method without rejection is presented in figure 9.2 with the green dots for $N_{MC} = 20$ samples. The left column of figure 9.2 presents the results for $a = 0.6$ and $a = 0.3$. The smaller a is, the larger the size $|\mathbb{V}|$ is: as a consequence, between the top picture and the bottom one of this column, more MC particles are sampled in the "zero weight" area and wasted. With the rejection method (blue dots), every MC particles are sampled within interval $[-a, a]$ but at the cost of several rejections. The blue dots are then uniformly sampled within $[-a, a]$ and their weights corrected accordingly. The quadrature method without rejection is very inefficient: in figure 9.2 there are only 60% ($a = 0.6$) and 30% ($a = 0.3$) of MC particles sampled within the area of non-zero probabilities. But the rejection does not always increase drastically its efficiency: consider the blue dots in the top-right picture of figure 9.2. For such value of $\sigma = 0.01$, the interesting area (of high probability) is narrow: many blue dots are assigned a very small weight due to the steep gradient of probability of occurrence of some positions. This may induce slow convergence rates for the MC resolution.
- To end the description of the sampling phase, let us comment on the consequence of a non consistent sampling (and prepare the discussion concerning the teleportation error of section 10.2). Suppose the positions are sampled uniformly within the cell $[-1, 1]$ (as for the 'quadrature' strategy) but the weights are uniformly set to ensure the exactness of the first moment only. Basically, in figure 9.2, this would imply having the same positions as the blue and green dots but the magenta weights. For the green dots, MC particles in the "zero-weight" area (or in the 'nearly-zero-weight' one for the blue dots) would have a non-negligible weight. This can significantly affect the solution u at later times (see teleportation error [301]).

We here presented two solutions for the sampling of the initial condition, we insist lots of other possibilities exist. For example, the material of part II regarding quadrature rules can be used. In fact, Gauss quadrature rules are even applied for the sampling of the initial condition in [298] for the MC resolution of the Euler system with a BGK model. The method is called 'Quiet MC' as the use of Gauss quadrature rules in each cell for the sampling (both initial and source sampling in fact, see section 9.9) allows having smoother profiles than with the classical MC sampling. The numerical strategy described in [298] is *more an original sampling strategy than a new MC scheme*. In this sense, paper [298] shows the sampling phase can be crucial and considerably contribute to an acceleration of the MC resolution.

9.8.2 A general skeleton in order to develop each scheme in the same platform

Algorithm 9: The general canvas for the different MC schemes described in term of algorithmic operations in order to compute (direct) $U(\mathbf{x}, t) = \int u(\mathbf{x}, t, \mathbf{v}) d\omega dv$.

```

1 #SAMPLING described in algorithm 7 or algorithm 8
2 call sampling( $N_{MC}$ )
3 set  $t = \Delta t$ 
4 #Time step loop
5 while  $t < T$  do
6     #Initialize to zero the array of the quantity of interest on the whole simulation domain  $\mathcal{D}$ 
7     set  $U(\mathbf{x}, t) = 0 \forall \mathbf{x} \in \mathcal{D}$ 
8     #TRACKING: make sure each  $u_p$  is an MC particles
9     for  $p \in \{1, \dots, N_{MC}\}$  do
10        set  $s_p = t - \Delta t$  #this will be the current time of particle  $p$ 
11        while  $s_p < t$  and  $w_p > 0$  do
12            if  $x_p \notin \mathcal{D}$  then
13                #here a general function for the application of arbitrary boundary conditions
14                apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
15            end
16            sample  $\tau_{inter} = \text{sample\_interaction\_time}(\mathbf{v}_p, i_p)$ 
17            compute  $\tau_{exit} = \text{compute\_cell\_exit\_time}(\mathbf{x}_p, \mathbf{v}_p, i_p)$ 
18            compute  $\tau_{census} = \max(t - \tau, 0)$ 
19            set  $\tau = \min(\tau_{exit}, \tau_{census}, \tau_{inter})$ 
20            #move the particle  $p$ 
21             $\mathbf{x}_p \leftarrow \mathbf{x}_p - \mathbf{v}_p \tau$ ,
22            #change its weight
23             $(K, r) = \text{compute\_weight\_modif}(\mathbf{v}_p, \tau, \tau_{census}, \tau_{exit}, \tau_{inter}, i_p)$ 
24             $w_p \leftarrow K \times w_p$ 
25            if  $\tau == \tau_{census}$  then
26                #set the life time of particle  $p$  to zero:
27                 $s_p \leftarrow t$ 
28                #tally the contribution of particle  $p$ 
29                 $U(\mathbf{x}_p, t) += w_p$ 
30            end
31            if  $\tau == \tau_{exit}$  then
32                #The particle  $p$  changes of cell: find its new cell number
33                 $i_p = \text{find_neighbouring\_cell}(i_p, \mathbf{v}_p)$ 
34                #set the life time of particle  $p$  to:
35                 $s_p \leftarrow s_p + \tau < t$ 
36            end
37            if  $\tau == \tau_{inter}$  then
38                #Sample the velocity of particle  $p$ 
39                 $\mathbf{V}' = \text{sample\_velocity}(\mathbf{v}_p, r, i_p)$ 
40                set  $\mathbf{v}_p = \mathbf{V}'$ 
41                #set the life time of particle  $p$  to:
42                 $s_p \leftarrow s_p + \tau < t$ 
43            end
44        end
45    end
46     $t \leftarrow t + \Delta t$ 
47 end

```

Now we detailed the sampling phase, common to every direct resolutions, we suggest tackling the

tracking one, see algorithm 9. This tracking phase is independent of the sampling of the previous section. It describes the 'tracking' of the population of particles²³ in the discretised simulation domain $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$, with the hypothesis of having constant cross-sections in each cell and time step

$$\sigma_\alpha(\mathbf{x}, v) = \sum_{i=1}^{N_x} \sigma_\alpha^i(v) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}), \alpha \in \{S, t\}.$$

More details are given in section 9.6. The structure of the tracking phase of the particles is very close to the one presented in the description of the direct non-analog scheme (algorithm 4). We only encapsulated some key parts in several functions: sample_interaction_time, compute_weight_modif, sample_velocity²⁴. The three latter key functions are described in algorithms 10–11–12 but for the moment we focus on the common canvas (algo. 9).

Algorithm 10: The sampling of the interaction time function

```

1 Function sample_interaction_time(real v, int i)
2   set  $\tau = \text{REAL\_MAX}$ 
3   #Sampling of the interaction time depending on the choice of the MC scheme
4   U =sample_uniform_law()
5   if MC_scheme == analog or MC_scheme == semi – analog then
6      $\tau = -\frac{\ln(U)}{v\sigma_t^i(v)}$ 
7   end
8   if MC_scheme == non – analog then
9      $\tau = -\frac{\ln(U)}{v\sigma_S^i(v)}$ 
10  end
11  return  $\tau$ 
```

Algorithm 11: Sampling of the velocity

```

1 Function sample_velocity(real v, int r, int i)
2   if MC_scheme == analog then
3     if full_analog then
4       #The function returns a list of  $\nu_r$  particles
5       ( $p'_1, \dots, p'_{\nu_r}$ )=split_the_particle_into( $\nu_r$ )
6     end
7     for  $j \in \{1, \dots, \nu_r\}$  do
8       #Call the probability measure for reaction r in cell i for each split MC particles
9        $\mathbf{V}_j = \text{sample\_from\_} P_S^{r,i}(\mathbf{v})$ 
10       $\mathbf{v}_{p'_j} = \mathbf{V}_j$ 
11    end
12  end
13  if MC_scheme == semi-analog or MC_scheme == non-analog then
14    #Averaged over the set of reactions in cell i
15     $\mathbf{V}' = \text{sample\_from\_} P_S^i(\mathbf{v})$ 
16  end
17  return  $\mathbf{V}'$ 
```

Each presented scheme relies on comparing three times, τ_{inter} the interaction time, τ_{exit} the time at which an MC particle p would get out of the cell i_p , τ_{census} the time before ending the time step. For each scheme, the particle moves along $\mathbf{v}_p \tau$ where τ is the minimum of the three above times. Its weight is modified or not (in compute_weight_modif) depending on the scheme. Furthermore, depending on the minimum of $\tau_{census}, \tau_{exit}, \tau_{inter}$, the particle sees its life time updated and finishes its treatment (*census*)

²³initially sampled according to algorithm 7 or algorithm 8.

²⁴We do not detail the functions compute_cell_exit_time and find_neighbouring_cell as they depend more on the type of grid (cartesian, structured, unstructured) than on the MC resolution scheme.

or crosses the interface between two cells (*exit*) or encounters an interaction (*inter*). In the latter case, its velocity change.

Let us now focus on the encapsulated functions. First, note that they all only depend on particle fields ($\mathbf{x}_p, \mathbf{v}_p, i_p, \dots$). The first one, to sample the interaction time, only needs the particle energy \mathbf{v}_p and is detailed in algorithm 10. Depending on the chosen scheme, the interaction time is sampled from the total cross-section σ_t (analog and semi-analog) or from the scattering one σ_S in the current cell i_p . Both are obtained inversing the cdf of an exponential law, see section 9.6.1.

The second corresponds to the modification of the weight of the particle, detailed in algorithm 12. For this function, the event the particle encounters explicitly appears in the treatment. The non-analog scheme is the only one having a treatment independent of the event. The weight of a particle remains unchanged for the analog and semi-analog schemes for the *census* and *cell exit* events. It changes in the case of an interaction: for the semi-analog scheme, the weight is multiplied by the probability of being scattered $\frac{\sigma_S}{\sigma_t}$. For the analog scheme, once the indice r of the reaction sampled²⁵, the weight of the particle is multiplied by the multiplicity ν_r of reaction r or will be split later (at the interaction position and time) for the 'full_analog' option.

Algorithm 12: The weight modification depending on the MC scheme

```

1 Function compute_weight_modif(real v, real  $\tau_{\min}$ , real  $\tau_{census}$ , real  $\tau_{exit}$ , real  $\tau_{inter}$ , real i)
2   set K = 1
3   if MC_scheme == analog then
4     if  $\tau_{\min} == \tau_{exit}$  or  $\tau_{\min} == \tau_{census}$  then
5       | K = 1
6     end
7     if  $\tau_{\min} == \tau_{inter}$  then
8       | r=sample_reaction_number( $\sigma_S^i$ )
9       | if multiplicity then
10      |   | K =  $\nu_r$ 
11      | end
12      | if full_analog then
13      |   | K = 1
14      | end
15    end
16  end
17  if MC_scheme == semi - analog then
18    if  $\tau_{\min} == \tau_{exit}$  or  $\tau_{\min} == \tau_{census}$  then
19      | K = 1
20    end
21    if  $\tau_{\min} == \tau_{inter}$  then
22      | | K =  $\frac{\sigma_S^i(v)}{\sigma_t^i(v)}$ 
23    end
24  end
25  if MC_scheme == non - analog then
26    | K =  $e^{-v(\sigma_t^i(v)-\sigma_S^i(v))\tau_{\min}}$ 
27  end
28  return K, r

```

At the interaction time, each scheme needs the sampling of the outer velocity \mathbf{V}' , summed up in algorithm 11. The semi-analog and the non-analog schemes share the same procedure, using P_S , averaged over the set of reactions $r \in \{0, \dots, N_R\}$. The analog scheme implies keeping in memory the previously sampled reaction number r in order to:

- first, split particle p into ν_r new particles $(p'_1, \dots, p'_{\nu_r})$,
- and sample for each of them the outer velocity \mathbf{V}'_j from $P_S^{r,i}$ in cell i .

²⁵The function sample_reaction_number is not detailed. In fact, it is very similar to the function of algorithm 6: one must only replace N_x by N_R , U_0^i by σ_r and U_0 by σ_t .

Note that the functions `split_the_particle_into` and `sample_from_PSi` are not detailed. They depend more on the choice of the data structures and physical data format (see section 9.6.2) than on the MC resolution schemes.

In algorithm 9, time steps are explicitly detailed but for the linear Boltzmann equation, time steps may coincide with the times of interest (MC methods are unconditionally stable for the linear Boltzmann equation). In other words, if one is only interested in time T , it is possible choosing $\Delta t = T$. This will not be the case in chapter 10 in which the coupling with additional equations will induce restrictions on the time step. Note that the above algorithm description still applies for the MC resolution of the linear Boltzmann equation coupled to other equations but needs additional instrumentations (track length estimator for example). They will be detailed in the next chapter.

To finish, we add that the tracking phase as described in algorithm 9 is commonly denoted 'history-based'. It refers to the fact that during one time step, an MC particle is followed from $s_p = t - \Delta t$, i.e. the beginning of the new time step, until $s_p = t$, i.e. the end of the current time step (if, of course, the MC particle is not killed during its history²⁶). Another possibility would be to apply the events one by one to the whole population of particles until they all reach census or die, this is what is commonly called an 'event-based' tracking phase. These considerations are practical ones and do not explicitly depend on the applied MC scheme. Nevertheless, the discussion on the choice of a 'history-based' tracking or an 'event-based' one is far from being irrelevant as the target computation device (one may have access to a station, a computation cluster, a supercomputer, with homogeneous nodes or hybrid ones...) may be sensitive to the operations induced by the two possibilities. Hybrid architectures (classical nodes, GP-GPU units, vectorization) becoming more and more common, hybrid strategies mixing history-and-event-based tracking phases may become more and more relevant. The discussion is beyond the scope of this document but is very interesting and we refer to [99] for some examples of fine-coarse grain parallel strategies for the linear Boltzmann equation.

9.9 Taking into account a source term

Just as the set of random variables for the MC resolution of the transport equation (or the initial sampling) is not unique, there are many different ways to deal with a source term. In fact, its presence just adds combinatorial possibilities. In this section, we focus on two different MC strategies. We first focus on the most common treatment. It is based on Duhammel's principle²⁷. It is very convenient in practice as it does not need to rewrite a complete set of MC treatments if one has already access to

- an MC solver for the linear Boltzmann equation without source terms,
- together with an initial sampling phase.

We show the treatment of the source term is closely related to the initial sampling. In section 9.9.2, we present a different way to take it into account. It corresponds to a new MC scheme, rather than only a modification of the sampling phase. The scheme is original and to our knowledge is not commonly applied in the literature (at least for neutronics or photonic applications). It is designed to be efficient in presence of a stiff source term (Asymptotic Preserving scheme in this regime, see remark 9.1). It is presented here because its construction is based on some important practical steps which are intensively applied in chapter 10 in a more complex context (nonlinear Boltzmann equation due to a coupling).

As we aim at focusing on the treatment of the source term, in order to alleviate the notations and avoid redundancies, we make some assumptions. We consider the monokinetic transport equation with constant cross-sections with respect to time, space and velocity, i.e. $\sigma_t(\mathbf{x}, t, \mathbf{v}) = \sigma_t$ and $\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}') = \sigma_s P_s(\omega', \omega)$. This corresponds to a very simple case but the generalizations are straightforward thanks to the already described material. The previous hypothesis lead to the following transport equation

$$-\partial_t u(\mathbf{x}, t, \omega) - \mathbf{v} \partial_x u(\mathbf{x}, t, \omega) + v \sigma_t u(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega' - S(\mathbf{x}, t, \omega). \quad (9.99)$$

²⁶depending on the chosen MC scheme.

²⁷Applied in order to prove the existence and unicity of the solution of linear Boltzmann equation with source term.

The source term has not been treated previously only because of a small subtlety of the structure of (9.99). Consider first equation

$$-\partial_t u(\mathbf{x}, t, \omega) - \mathbf{v} \partial_x u(\mathbf{x}, t, \omega) + v \sigma_t u(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega', \quad (9.100)$$

which corresponds to (9.99) where $S = 0$. Then if $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ are solutions of (9.100), it is easy verifying $\sum_{p=1}^{N_{MC}} u_p$ is also a solution of (9.100). On another hand, if $(u_p)_{p \in \{1, \dots, N_{MC}\}}$ are solutions of (9.99), then $\phi = \sum_{p=1}^{N_{MC}} u_p$ is not a solution of (9.99) but rather a solution of

$$-\partial_t \phi(\mathbf{x}, t, \omega) - \mathbf{v} \partial_x \phi(\mathbf{x}, t, \omega) + v \sigma_t \phi(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' - \textcolor{red}{N_{MC}} S(\mathbf{x}, t, \omega). \quad (9.101)$$

In other words, the quantity $\frac{1}{N_{MC}} \sum_{p=1}^{N_{MC}} u_p$ is the solution of the linear Boltzmann equation with source term (9.99). If the number of particles depends on time²⁸, i.e. if $N_{MC}(t)$, some additional treatments must be added to cancel the $\partial_t N_{MC}(t)$ term for $\frac{1}{N_{MC}(t)} \sum_{p=1}^{N_{MC}} u_p$ to solve (9.99). The latter property may appear irrelevant and obvious at this stage of the discussion but it becomes really important when designing, developing and implementing an MC scheme. In the case of a source term, depending on the choice of the scheme, the weight of the MC particles, mainly in the *sampling phase*, have to be defined in a consistent way (with respect to N_{MC}) together with every numerical algorithm implying a change in the number of MC particles during their tracking (russian-roulette, splitting, window screening etc.).

In the following sections, we describe two ways to take into account a source term:

- the first one resumes to an enrichment of the initial sampling phase (hence its denomination *source sampling*).
- The second corresponds to a new MC scheme.

For both resolutions, we rely on an integral form before identifying different sets of random variables allowing to rewrite it as an expectation.

9.9.1 Application of Duhammel's principle: source sampling (direct)

The most common MC strategy allowing to take into account an external source term S relies on Duhammel's principle. It can be stated as follow: let u_1 be the solution of the following Cauchy problem implying the linear Boltzmann equation without source term

$$\begin{cases} -\partial_t u_1(\mathbf{x}, t, \omega) - \mathbf{v} \partial_x u_1(\mathbf{x}, t, \omega) + v \sigma_t u_1(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) u_1(\mathbf{x}, t, \omega') d\omega', \\ u_1(\mathbf{x}, 0, \omega) = u_1^0(\mathbf{x}, \omega). \end{cases} \quad (9.102)$$

Let u_2 be the solution of the following Cauchy problem with

$$\begin{cases} -\partial_t u_2(\mathbf{x}, t, \omega) - \mathbf{v} \partial_x u_2(\mathbf{x}, t, \omega) + v \sigma_t u_2(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) u_2(\mathbf{x}, t, \omega') d\omega' - S(\mathbf{x}, t, \omega), \\ u_2(\mathbf{x}, 0, \omega) = 0. \end{cases} \quad (9.103)$$

Then $u = u_1 + u_2$ is solution of (9.99). The application of Duhammel's principle allows decoupling the treatment of the sources from the resolution of (9.102). The direct consequence is that one can choose its favorite solver for (9.102) and add the resolution of (9.103) to its MC resolution code almost transparently. From now on in this section, we focus on the MC resolution of (9.103). Equation (9.103)

²⁸It is the case when splitting or russian-roulette is activated.

can be rewritten in a integral form (direct formulation with $u_2(\mathbf{x}, 0, \omega) = 0$)

$$\int_0^t e^{-v\sigma_t s} S(\mathbf{x} + \mathbf{v}s, s, \omega) ds = \int_0^t \left[+\mathbf{1}_{[t, \infty]}(s) e^{-v\sigma_A t} u_2(\mathbf{x} + \mathbf{v}t, t, \omega) + \mathbf{1}_{[0, t]}(s) \int P_S(\omega', \omega) e^{-v\sigma_A s} u_2(\mathbf{x} + \mathbf{v}s, s, \omega') \right] v\sigma_S e^{-v\sigma_S s} ds. \quad (9.104)$$

It must be rewritten as an expectation over a set of identified random variables in order to introduce an MC discretisation. Once again, this set of random variables is not unique. Note that we paved the path toward a non-analog treatment as we already introduced $\sigma_A = \sigma_t - \sigma_S$. Let us assume equation (9.102) is solved with a non-analog MC scheme and that we want to apply similar treatments in order to take into account the source term. Introduce $\tau \sim \mathcal{E}(v\sigma_S)$ and $\tau_U \sim \mathcal{U}([0, t = \Delta t])$, $W' \sim P_S(\omega', \omega)$, then (9.104) becomes

$$\Delta t \mathbb{E} [e^{-v\sigma_S \tau_U} e^{-v\sigma_A \tau_U} S(\mathbf{x} + \mathbf{v}\tau_U, \tau_U, \omega)] = \mathbb{E} [\mathbf{1}_{[t, \infty]}(\tau) e^{-v\sigma_A t} u_2(\mathbf{x} + \mathbf{v}t, t, \omega) + \mathbf{1}_{[0, t]}(\tau) e^{-v\sigma_A \tau} u_2(\mathbf{x} + \mathbf{v}\tau, \tau, W')] \quad (9.105)$$

Introduce furthermore $\tau' \sim \mathcal{E}(v\sigma_S)$, we get for the left hand side

$$\Delta t \mathbb{E} [\mathbf{1}_{[\tau_U, \infty]}(\tau') e^{-v\sigma_A \tau_U} S(\mathbf{x} + \mathbf{v}\tau_U, \tau_U, \omega)] = \mathbb{E} [\mathbf{1}_{[t, \infty]}(s) e^{-v\sigma_A t} u_2(\mathbf{x} + \mathbf{v}t, t, \omega) + \mathbf{1}_{[0, t]}(s) e^{-v\sigma_A \tau} u_2(\mathbf{x} + \mathbf{v}\tau, \tau, W')] \quad (9.106)$$

At this stage, expression (9.106) must be compared to equation (9.44) (with some simplification hypothesis, such as monokinetic etc.) where the initial condition in (9.44) has been replaced by

$$\Delta t \mathbb{E} [\mathbf{1}_{[\tau_U, \infty]}(\tau') e^{-v\sigma_A \tau_U} S(\mathbf{x} + \mathbf{v}\tau_U, \tau_U, \omega)]. \quad (9.107)$$

An MC resolution of (9.106) comes naturally with the MC sampling of (9.107), which is similar and can benefit the strategies applied in order to ensure the initial population of MC particles accurately represent the initial condition²⁹. It resumes to working on the sampling of (9.107) instead of $u_0(\mathbf{x}, \omega)$, hence its denomination *source sampling*. Typically, the source sampling can be done during the initial sampling phase (see for example algorithm 4). The main difference comes from the fact a time discretisation must be taken into account for the source particles. Once the source sampling done, the MC treatments are the same as in section 9.5.2. In the following section, we focus on the MC discretisation of (9.107) with $p \in \{1, \dots, N_{MC}^S\}$ source particles. We consider $N_{MC} = N_{MC}^S + N_{MC}^0$ where N_{MC}^0 denotes the number of MC particles representing the initial condition. Let us define a source MC particle $S_p(\mathbf{x}, t, \omega) = w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\omega}(\omega_p(t))$ and determine the treatments one must apply in order to ensure it is an MC solution of (9.106). Plugging the expression of $(S_p)_{p \in \{1, \dots, N_{MC}^S\}}$ into (9.106) leads to

$$\Delta t \mathbf{1}_{[\tau_U, \infty]}(\tau') e^{-v\sigma_A \tau_U} S_p(\mathbf{x} + \mathbf{v}\tau_U, \tau_U, \omega) = \mathbf{1}_{[t, \infty]}(\tau) e^{-v\sigma_A t} S_p(\mathbf{x} + \mathbf{v}t, t, \omega) + \mathbf{1}_{[0, t]}(\tau) e^{-v\sigma_A \tau} S_p(\mathbf{x} + \mathbf{v}\tau, \tau, W'). \quad (9.108)$$

After a change of variable, we have

$$\Delta t \mathbf{1}_{[\tau_U, \infty]}(\tau') S_p(\mathbf{x}, \tau_U, \omega) = \mathbf{1}_{[t, \infty]}(\tau) e^{-v\sigma_A(t - \tau_U)} S_p(\mathbf{x} + \mathbf{v}(t - \tau_U), t, \omega) + \mathbf{1}_{[0, t]}(\tau) e^{-v\sigma_A(\tau - \tau_U)} S_p(\mathbf{x} + \mathbf{v}(\tau - \tau_U), \tau, W'). \quad (9.109)$$

Now, let us first consider the case $\tau' < \tau_U$: in this case, the indicatrix on the left hand side is zero and we have

$$0 = \mathbf{1}_{[t, \infty]}(\tau) e^{-v\sigma_A(t - \tau_U)} S_p(\mathbf{x} + \mathbf{v}(t - \tau_U), t, \omega) + \mathbf{1}_{[0, t]}(\tau) e^{-v\sigma_A(\tau - \tau_U)} S_p(\mathbf{x} + \mathbf{v}(\tau - \tau_U), \tau, W'), \quad (9.110)$$

which implies $S_p = 0 \forall t \in [0, \Delta t]$. In other words, only sources emitted after τ_U contribute to u_2 . In practice, in order to make sure every source MC particle contribute to u_2 (and avoid rejections), their *birth times* are sampled from a uniform law on $\tau_U \sim \mathcal{U}([0, \Delta t])$ so that we have $\tau' > \tau_U$ by construction.

²⁹presented in sections 9.5.2–9.8 and in algorithms 5–7–8.

From now on we suppose $\tau' > \tau_U$. We obtain the following recursive treatment for particle p

$$\begin{aligned} w_p(\tau_U) \delta_{\mathbf{x}}(\mathbf{x}_p(\tau_U)) \delta_{\omega}(\omega_p(\tau_U)) = \\ + \mathbf{1}_{[t, \infty]}(\tau) e^{-v\sigma_A(t-\tau_U)} w_p(t) \delta_{x+\mathbf{v}(t-\tau_U)}(\mathbf{x}_p(t)) \delta_{\omega}(\omega_p(t)) \\ + \mathbf{1}_{[0, t]}(\tau) e^{-v\sigma_A(\tau-\tau_U)} w_p(\tau) \delta_{x+\mathbf{v}(\tau-\tau_U)}(\mathbf{x}_p(\tau)) \delta_{W'}(\omega_p(\tau)). \end{aligned} \quad (9.111)$$

It is exactly the same treatment as the direct one presented in section 9.5.2 (in the monokinetic case) except the initial condition begins at time τ_U rather than 0. Of course, the above treatment implicitly calls for a consistent sampling of the source term, i.e. such that $\sum_{p=1}^{N_{MC}^S} S_p(\mathbf{x}, \tau_U^p, \omega) \stackrel{a.s.}{=} \Delta t \mathbb{E}[S(\mathbf{x}, \tau_U, \omega)]$. The next paragraph is dedicated to a brief description (complementary to section 9.8.1) together with difficulties one must have in mind when dealing with S .

Let us now focus on some technical details (complementary to the ones of the initial sampling) of the source sampling phase. First, this sampling must come with a relevant choice of N_{MC}^S with respect to N_{MC}^0 . This choice can be arbitrary, the convergence being ensured by an increase in both numbers of initial (N_{MC}^0) and source particles (N_{MC}^S). With an arbitrary choice, the risk is only to waste computational time. For example, it is possible sampling too many sources with respect to the number of initial particles whereas the source term does not really contribute to the observable of interest (but the initial condition does). In practice, it is common building a binomial law of states 'source' and 'initial' of parameters respectively

$$\frac{U_S}{U_S + U_0} \text{ and } \frac{U_0}{U_S + U_0},$$

where

$$U_S = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \int_0^t \int S(\mathbf{x}, s, \omega) d\mathbf{x} ds d\omega \text{ and } U_0 = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \int u_0(\mathbf{x}, \omega) d\mathbf{x} d\omega.$$

The above expressions are similar to (9.88) for the initial sampling. Then one must determine the state of the MC particle ('initial' or 'source') during the sampling phase. Such choice ensures $\frac{N_{MC}^S}{N_{MC}^0} \approx \frac{U_S}{U_S + U_0}$ so that the number of sources N_{MC}^S is sampled relatively to the global weight of every particles. It kind of implicitly assumes the sources are contributing to the observable of interest with ratio $\frac{U_S}{U_S + U_0}$ and the initial condition with ratio $\frac{U_0}{U_0 + U_S}$. Consequently, this does not really guaranty avoiding a waste of computational time but it ensures a convergence of the MC resolution with N_{MC} , i.e. without choosing an additional parameter.

Independently of the relative amount of source vs. initial MC particles, the sampling of the N_{MC}^S sources must ensure $\sum_{p=1}^{N_{MC}^S} S_p(\mathbf{x}, \tau_U^p, \omega) \stackrel{a.s.}{=} \Delta t \mathbb{E}[S(\mathbf{x}, \tau_U, \omega)] = \int_0^t S(\mathbf{x}, s, \omega) ds = \Delta t S^n(\mathbf{x}, \omega)$. Just as for the initial sampling in section 9.5.3, source sampling can be described generally introducing the probability measure

$$\begin{aligned} dS(\mathbf{x}, s, \omega) &= \frac{1}{\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \int_0^t \int S(\mathbf{x}, s, \omega) d\mathbf{x} ds d\omega} S(\mathbf{x}, s, \omega) d\mathbf{x} d\omega ds, \\ &= \frac{1}{U_S} S(\mathbf{x}, s, \omega) d\mathbf{x} d\omega ds. \end{aligned} \quad (9.112)$$

The latter must be compared to expression (9.88) for the initial condition. Once the analogy noticed, the strategies hinted at in section 9.8.1 can be applied. We only end this paragraph on an important remark: the consistent samplings and weight corrections, presented in section 9.8.1, may be more critical than for the initial condition. Indeed, for example, arbitrarily setting uniform weights for the source particles, if not consistent, can lead to intractable approximations, see the *teleportation error* in [301, 69, 197, 70, 146] or the material of section 10.2.

9.9.2 Quasi-Static method for the transport equation with source term

In this section, we present a different way to take into account a source term. The previous one only implied a modification of the sampling phase. The one presented here implies a new and original MC

scheme. In fact, the scheme description is almost secondary: we use it as a pretext in order to make a parallel and an analogy between Quasi-Static methods, well-known and commonly applied in (coupled/nonlinear³⁰) neutronics mainly (see [139, 140, 216, 73, 221]), and Asymptotic-Preserving schemes (see section 9.7.4).

We consider the linear Boltzmann equation with source term

$$-\partial_t u(\mathbf{x}, t, \omega) - \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + v \sigma_t u(\mathbf{x}, t, \omega) = v \sigma_S \int P_S(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega' - S(\mathbf{x}, t). \quad (9.113)$$

We first non-dimensionalize it, exactly as in section 9.7.4, by introducing

$$\begin{cases} \mathbf{x} = \mathbf{x}^* \mathcal{X}, v = v^* \mathcal{V}, t = t^* \mathcal{T}, \\ \sigma_\alpha = \sigma_\alpha^* \frac{1}{\lambda_\alpha}, \forall \alpha \in \{S, t, A\}. \end{cases} \quad (9.114)$$

The upperscript * is used to denote nondimensional quantities. Let us introduce $u^*(\mathbf{x}^*, t^*, \omega) = u(\mathbf{x}, t, \omega)$ and $S^*(\mathbf{x}^*, t^*) = S(\mathbf{x}, t)$, then

$$\frac{1}{\mathcal{T}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) = \partial_t u(\mathbf{x}, t, \omega), \quad \frac{1}{\mathcal{X}} \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) = \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega).$$

Using the above expressions in the transport equation yields

$$-\frac{\mathcal{X}}{\mathcal{T} \mathcal{V}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) - v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_t^* \frac{\mathcal{X}}{\lambda_t} u^*(\mathbf{x}^*, t^*, \omega) = v^* \sigma_A^* \frac{\mathcal{X}}{\lambda_A} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} u^*(\mathbf{x}^*, t^*, \omega) - \frac{\mathcal{X}}{\mathcal{V}} S^*(\mathbf{x}^*, t^*).$$

Now, we decompose $\sigma_t = \sigma_A + \sigma_s$ to obtain

$$-\frac{\mathcal{X}}{\mathcal{T} \mathcal{V}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) - v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_A^* \frac{\mathcal{X}}{\lambda_A} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} u^*(\mathbf{x}^*, t^*, \omega) = v^* \sigma_s^* \frac{\mathcal{X}}{\lambda_s} \int u^*(\mathbf{x}^*, t^*, \omega) d\omega - \frac{\mathcal{X}}{\mathcal{V}} S^*(\mathbf{x}^*, t^*).$$

Suppose $\frac{\mathcal{X}}{\mathcal{T} \mathcal{V}} = \mathcal{O}(\frac{1}{\delta}) = \frac{\mathcal{X}}{\lambda_A} = \frac{\mathcal{X}}{\mathcal{V}}$ and $\frac{\mathcal{X}}{\lambda_s} = \mathcal{O}(1)$, we obtain (we drop the upperscript for convenience) the following asymptotic linear Boltzmann equation

$$-\frac{1}{\delta} \partial_t u(\mathbf{x}, t, \omega) - v \omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + v \sigma_A \frac{1}{\delta} u(\mathbf{x}, t, \omega) + v \sigma_s u(\mathbf{x}, t, \omega) = v \sigma_s \int u(\mathbf{x}, t, \omega) d\omega - \frac{1}{\delta} S(\mathbf{x}, t).$$

Performing a Hilbert development, i.e. $u = u^0 + u^1 \delta + u^2 \delta^2 + \mathcal{O}(\delta^3)$ see [143], and considering only the first order $u^0 = U$ leads to

$$\partial_t U(\mathbf{x}, t) = v \sigma_A U(\mathbf{x}, t) + S(\mathbf{x}, t). \quad (9.115)$$

It corresponds to the monokinetic homogeneous regime. It can be solved analytically, if completed by an initial condition $U_0(\mathbf{x}, \omega) = u(\mathbf{x}, 0, \omega)$ and considering \mathbf{x} as a parameter:

$$U(\mathbf{x}, t) = \left(U_0(\mathbf{x}) + \int_0^t S(\mathbf{x}, s) e^{-v \sigma_A s} ds \right) e^{v \sigma_A t}. \quad (9.116)$$

Suppose we now want to build an Asymptotic-Preserving MC scheme for the regime $\delta \rightarrow 0$. To do so, we follow the methodology described in [3] and use (9.116) as a reduced model preserving and solving the stiffness. We assume it is relevant along the flight path of the particles (i.e. applied on a characteristic). We want the MC scheme to focus on fluctuation around the stiff regime of interest. To make sure

³⁰It also allows introducing the methodology described in [3] for the linear Boltzmann equation coupled to the Bateman system in a simpler configuration and shows that the methodology also applies for the linear Boltzmann equation.

having an MC discretisation of these fluctuations, we introduce $f(\mathbf{x}, t, \omega)$ and the change of variable $u(\mathbf{x}, t, \omega) = U(\mathbf{x}, t)f(\mathbf{x}, t, \omega)$ with U as in (9.116) (on a characteristic). The new unknown f solves (multiplicative) fluctuations around the asymptotic regime U . Let us now identify the equation satisfied by f . For this, we plug $u = Uf$ into (9.113) to obtain

$$\begin{aligned} -\partial_t f(\mathbf{x}, t, \omega) - \mathbf{v} \partial_{\mathbf{x}} f(\mathbf{x}, t, \omega) &+ (v\sigma_t - \partial_t \ln(U(\mathbf{x}, t)) - \mathbf{v} \partial_{\mathbf{x}} \ln(U(\mathbf{x}, t))) f(\mathbf{x}, t, \omega) = \\ &+ v\sigma_S \int P_S(\omega', \omega) f(\mathbf{x}, t, \omega') d\omega' - \frac{S(\mathbf{x}, t)}{U(\mathbf{x}, t)}. \end{aligned} \quad (9.117)$$

It can be rewritten on a characteristic

$$\begin{aligned} -\partial_s f(\mathbf{x} + \mathbf{v}s, s, \omega) &+ (v\sigma_t - \partial_s \ln(U(\mathbf{x} + \mathbf{v}s, s))) f(\mathbf{x} + \mathbf{v}s, s, \omega) = \\ &+ v\sigma_S \int P_S(\omega', \omega) f(\mathbf{x} + \mathbf{v}s, s, \omega') d\omega' - \frac{S(\mathbf{x} + \mathbf{v}s, s)}{U(\mathbf{x} + \mathbf{v}s, s)}. \end{aligned} \quad (9.118)$$

Along a characteristic, according to (9.116), we have

$$\partial_s \ln(U(\mathbf{x} + \mathbf{v}s, s)) = v\sigma_A + \frac{S(\mathbf{x} + \mathbf{v}s, s)}{U(\mathbf{x} + \mathbf{v}s, s)}.$$

Plugging the above expression in (9.118) leads to

$$\begin{aligned} -\partial_s f(\mathbf{x} + \mathbf{v}s, s, \omega) &+ \left(v\sigma_S - \frac{S(\mathbf{x} + \mathbf{v}s, s)}{U(\mathbf{x} + \mathbf{v}s, s)} \right) f(\mathbf{x} + \mathbf{v}s, s, \omega) = \\ &+ v\sigma_S \int P_S(\omega', \omega) f(\mathbf{x} + \mathbf{v}s, s, \omega') d\omega' - \frac{S(\mathbf{x} + \mathbf{v}s, s)}{U(\mathbf{x} + \mathbf{v}s, s)}. \end{aligned} \quad (9.119)$$

Now suppose a time step discretisation $[0, \Delta t]$ and assume $\int f(\mathbf{x}, t, \omega) d\omega \approx 1 \forall \mathbf{x} \in \mathcal{D}, \forall t \in [0, \Delta t]$. The latter hypothesis is classical in Quasi-Static (QS) methods, see [139, 140, 216, 73, 221] and is referred as the *main hypothesis* in [3]. Then (9.119) can be rewritten as a balanced emission-absorption linear Boltzmann equation

$$\begin{aligned} -\partial_s f(\mathbf{x} + \mathbf{v}s, s, \omega) &+ v\Sigma_S(\mathbf{x} + \mathbf{v}s, s) f(\mathbf{x} + \mathbf{v}s, s, \omega) = \\ &+ v\Sigma_S(\mathbf{x} + \mathbf{v}s, s) \int \bar{P}_S(\mathbf{x} + \mathbf{v}s, s, \omega', \omega) f(\mathbf{x} + \mathbf{v}s, s, \omega') d\omega'. \end{aligned} \quad (9.120)$$

It is such that $\partial_s \int f(\mathbf{x} + \mathbf{v}s, s, \omega) d\omega = 0 \forall s \in [0, \Delta t]$. In (9.120), we introduced $\forall \mathbf{x} \in \mathcal{D}, \forall t \in [0, \Delta t]$:

$$\begin{aligned} v\Sigma_S(\mathbf{x}, t) &= v\sigma_S - \frac{S(\mathbf{x}, s)}{U(\mathbf{x}, s)}, \\ \bar{P}_S(\mathbf{x}, t, \omega', \omega) &= \frac{v\sigma_S P_S(\omega', \omega) - \frac{S(\mathbf{x}, s)}{U(\mathbf{x}, s)}}{v\Sigma_S(\mathbf{x}, t)}. \end{aligned}$$

Equation (9.120) has a form which has already been intensively encountered all along this document and we know how to apply an MC discretisation to such linear Boltzmann equation with space and time dependent cross-sections (see the material of section 9.5.2). We do not detail all the MC treatments but would like to focus on the weight modification of the MC particles. Its expression is given by

$$\frac{U(\mathbf{x} + \mathbf{v}t, t)}{U_0(\mathbf{x})} = \left(\frac{U_0(\mathbf{x} + \mathbf{v}t)}{U_0(\mathbf{x})} + \int_0^t \frac{S(\mathbf{x} + \mathbf{v}t, s)}{U_0(\mathbf{x})} e^{-v\sigma_A s} ds \right) e^{v\sigma_A t}.$$

It corresponds to the solution of the asymptotic regime $\delta \rightarrow 0$ of interest, along any characteristics. If $S = 0$, we recover the classical expression of the weight modification (and all the other MC treatments, even if not recalled here) for the non-analog scheme. In the asymptotic regime of interest $\delta \rightarrow 0$, the AP/QS scheme ensures by construction $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}}) = \frac{\sigma_{AP/QS}(t)}{\sqrt{N_{MC}}} \ll 1$, whatever the number of MC particles, as asymptotically³¹ $\sigma_{AP/QS}(t) = 0$ when solving (9.115).

³¹the proof is very similar to the one detailed in section 9.7.3 for the non-analog scheme and is not performed here.

The description of the above AP/QS MC scheme for taking into account a source term allowed exhibiting strong analogies between

- Quasi-Static methods for the linear Boltzmann equation, intensively applied in neutronics [139, 140, 216, 73, 221],
- and Asymptotic Preserving schemes encountered in many recent publications [108, 49, 113, 50, 51] and different application fields.

Both terminologies are closely related if not equivalent and we wanted to emphasize this point. The description of the above scheme also introduced the methodology described in [3] in a simpler configuration (in the sense there is no coupling with any other system of equation here, S is a known external field).

9.10 Taking into account an acceleration term in MC resolution schemes

So far, we have tackled the linear Boltzmann equation with source term but without taking into account an external acceleration. In this section, we explain in which sense the discussions of the previous sections are more general than it appears. An acceleration term can, for example, allow taking into account external forces on the physical particles. In this case, its expression is of the form $a(\mathbf{x}, t, \mathbf{q}) = \frac{F(\mathbf{x}, t, \mathbf{q})}{m}$ where $F(\mathbf{x}, t, \mathbf{q})$ corresponds to the force applied to the particles at position $\mathbf{x} \in \mathcal{D}$, time $t \in [0, T]$, velocity $\mathbf{q} \in \mathbb{R}^3$ (or energy $q = |\mathbf{q}| \in \mathbb{R}^+$ and angle $\Omega = \frac{\mathbf{q}}{q} \in \mathbb{S}^2$). The scalar m corresponds to the mass of the particles of interest. In this section, we slightly change our notations as the velocity component is now denoted by \mathbf{q} instead of \mathbf{v} . We will have $\mathbf{q} = \mathbf{v} \iff a = 0$. The expression of \mathbf{q} with respect to \mathbf{v} when $a \neq 0$ is the purpose of the following material. The acceleration applied to the particles can be general and come from different physics: it can be gravitational forces, electromagnetic forces, it can also be introduced in order to deal with Doppler and aberration effects, refractive and dispersive media in photonics³² (see [59, 245, 203]) etc. Its expression can be considered an external (known) field or may be induced by another physic of interest (Maxwell equations for example in the case of electromagnetic field forces) hence implies a coupling. In this chapter, we consider the acceleration term is known, i.e. is a known external field.

The linear Boltzmann equation with acceleration term generally rewrites

$$\begin{aligned} \partial_t f(\mathbf{x}, t, \mathbf{q}) + \mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) + a(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) = \\ -q\sigma_t(q)f(\mathbf{x}, t, \mathbf{q}) + \int q\sigma_s(\mathbf{q}, \mathbf{q}')f(\mathbf{x}, t, \mathbf{q}')d\mathbf{q}' + S(\mathbf{q}). \end{aligned} \quad (9.121)$$

The density of particles at position $\mathbf{x} \in \mathcal{D}$, time $t \in [0, T]$, velocity³³ $q \in \mathbb{R}^+$ and angle $\Omega \in \mathbb{S}^2$ is denoted by $f(\mathbf{x}, t, q, \Omega) = f(\mathbf{x}, t, q\Omega) = f(\mathbf{x}, t, \mathbf{q})$. The left hand side of the equation is called the *streaming part* of (9.121) and the right hand side the *collisional part* of (9.121). For the purpose of this section, we consider the total cross-section σ_t , the scattering cross-section σ_s and the source term S only depend on, respectively, q , $(\mathbf{q}, \mathbf{q}')$ and \mathbf{q} as the treatment of the position and the time dependences can benefit the descriptions of the previous sections.

There are two main ways to treat the acceleration term in MC computations:

- the first one implies curved trajectories of the MC particles for the streaming part of (9.121) and collisions in the *comobile frame*. This solution is described in section 9.10.1.
- The second one implies straight trajectories for the MC particles but corrections of the cross-sections and source term expressed in a *new referential*. It will be described in section 9.10.2.

³²in this case, q denotes a frequency rather than an energy or a velocity.

³³or energy or frequency.

These possibilities are well-known and applied in many physical applications (see plasma [189, 25, 24, 29], photonics [59, 203], neutronics [296, 297] etc.), but we here present them in a general and common way focusing on the practical implications on an MC resolution.

9.10.1 An MC resolution with curved trajectories in the comobile frame

A first possibility in order to build an MC scheme taking into account an acceleration term in the transport equation consists in rewriting the transport equation on a curved/accelerated characteristic. In the previous sections, a characteristic was (implicitly) defined by the change of variables:

$$\begin{cases} d_t \mathbf{v}(t) = 0, \\ d_t \bar{\mathbf{x}}(t) = \mathbf{v}(t), \\ \mathbf{v}(0) = \mathbf{v}, \\ \bar{\mathbf{x}}(0) = \mathbf{x}. \end{cases} \quad (9.122)$$

Its resolution leads to $\mathbf{v}(t) = \mathbf{v}(0) = \mathbf{v}$ and $\bar{\mathbf{x}}(t) = \mathbf{x} + \mathbf{v}t$. In this section, the acceleration term modifies the velocity³⁴ and the angular distribution of the particles along their flight path as it induces the following change of variables:

$$\begin{cases} d_t \mathbf{q}(t) = a(\mathbf{x}(t), t, \mathbf{q}(t)), \\ d_t \mathbf{x}(t) = \mathbf{q}(t), \\ \mathbf{q}(0) = \mathbf{q}, \\ \mathbf{x}(0) = \mathbf{x}. \end{cases} \quad (9.123)$$

Of course, in this case, a characteristic is much more complex: solving the above system leads to

$$\begin{cases} \mathbf{q}(t) = \mathbf{q} + \int_0^t a(\mathbf{x}(s), s, \mathbf{q}(s)) ds, \\ \mathbf{x}(t) = \mathbf{x} + \int_0^t \mathbf{q}(s) ds. \end{cases} \quad (9.124)$$

Plugging the expression of $\mathbf{x}(t)$ into the acceleration term we obtain the weakly³⁵ coupled system

$$\begin{cases} \mathbf{q}(t) = \mathbf{q} + \int_0^t a\left(\mathbf{x} + \int_0^s \mathbf{q}(\alpha) d\alpha, s, \mathbf{q}(s)\right) ds, \\ \mathbf{x}(t) = \mathbf{x} + \int_0^t \mathbf{q}(s) ds. \end{cases} \quad (9.125)$$

The equation satisfied by the velocity $\mathbf{q}(t)$ may be a nonlinear integro-differential equation and could be hard to solve depending on the shape of the acceleration term a . In the following discussions, we assume existence and unicity³⁶ of $\mathbf{q}(t)$ and $\mathbf{x}(t)$ and that their analytical expressions are available. In practice, one may have to rely on approximations such as the ones of section 9.6 for example (constant acceleration in each cell \mathcal{D}_i such that $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$, independence of a with respect to $\mathbf{q}(t)$ etc.). Of course, care has to be taken so that the asymptotic regime of interest is not strongly affected by the approximations. If a varies a lot on small characteristic distances, one may rely on AP methods or high order reconstructions within cells (Splines for example for [189, 25, 24, 29]) for triggered stiff regime of interest.

Let us introduce the norm of the velocity $q(s) = |\mathbf{q}(s)|$ evolving with respect to time s . We can rewrite (9.121) on one characteristic

$$\begin{aligned} \partial_s f(\mathbf{x}(s), s, \mathbf{q}(s)) &= \\ -q(s)\sigma_t(\mathbf{q}(s))f(\mathbf{x}(s), s, \mathbf{q}(s)) + \iint q(s)\sigma_s(\mathbf{q}(s), \mathbf{q}')f(\mathbf{x}(s), s, \mathbf{q}')d\mathbf{q}' &+ S(\mathbf{q}(s)). \end{aligned} \quad (9.126)$$

³⁴hence the energy/frequency.

³⁵weakly coupled in the sense once solved with respect to \mathbf{q} , we consider the solution for $\mathbf{x}(t)$ is straightforward.

³⁶This suggest the definition of F does not contradict/trigger the wellposedness of (9.121) beforehand.

Let us introduce the abusive notations $q(s)\sigma_t(\mathbf{q}(s)) = q\sigma_t(s)$, $S(\mathbf{q}(s)) = S(s)$ and $q(s)\sigma_s(\mathbf{q}(s), \mathbf{q}') = q\sigma_s(s, \mathbf{q}, \mathbf{q}')$: the dependence with respect to time is recalled in the first argument and to the initial conditions $\mathbf{q}(0) = \mathbf{q}$ of the velocity vector in the second one. It allows rewriting (9.126) as

$$\partial_s f(\mathbf{x}(s), s, \mathbf{q}(s)) = \iint q\sigma_s(s, \mathbf{q}, \mathbf{q}')f(\mathbf{x}(s), s, \mathbf{q}')d\mathbf{q}' - q\sigma_t(s)f(\mathbf{x}(s), s, \mathbf{q}(s)) + S(s). \quad (9.127)$$

With the above expression, the linear Boltzmann equation is rewritten in an already encountered form: it remains to build an MC scheme for the linear Boltzmann equation for cross sections/sources having dependences with respect to time. Everything we already presented in the previous sections applies. The integral form of equation (9.121) is then given by

$$f(\mathbf{x}(t), t, \mathbf{q}(t)) = \int \begin{pmatrix} f(\mathbf{x}(t), t, \mathbf{q}(t)) \\ +\mathbf{1}_{[t, \infty[}(s) & f_0(\mathbf{x}, \mathbf{q}) \\ +\mathbf{1}_{[0, t]}(s) & \left[\iint \frac{q\sigma_s(s, \mathbf{q}, \mathbf{q}')}{q\sigma_t(s)} f(\mathbf{x}(s), s, \mathbf{q}')d\mathbf{q}' + \frac{S(s)}{q\sigma_t(s)} \right] \end{pmatrix} q\sigma_t(s)e^{-\int_s^t q\sigma_t(\alpha)d\alpha}ds. \quad (9.128)$$

From (9.128) can be deduced the analog (section 9.2), semi-analog (section 9.3), non-analog (section 9.4) schemes in adjoint or direct (section 9.5) forms with source term (section 9.9). To sum-up, it is possible taking into account the acceleration term by considering *curved trajectories* and *time-dependent cross-sections* expressed in the *comobile frame*.

9.10.2 An MC resolution with straight trajectories in a new frame

Being able to solve the accelerated transport equation with an MC resolution scheme with particles having straight trajectories can be convenient in practice. For example

- because we already have access to a simulation code taking into account straight trajectories.
- Or because of the complexity of computing distances to events on curved trajectories in complex cells (non-conform, unstructured ones).

In this section, we briefly explain how, *via* a well chosen change of variables, it is possible to solve (9.121) with straight trajectories at the condition of performing some *corrections* to the cross-sections and source terms. It implies expressing the linear Boltzmann equation in a *new frame*. Depending on the physics of interest, those *corrections* and this *new frame* take different denominations. The term *effective* cross-sections/source term is used to describe the corrections in neutronics, see [296, 297]. In photonics, there are as many denominations as authors: they are called *covariant* relations in [59], *effective* ones in [174], *unadorned frame* ones in [245]. Particular terminologies are also used to denote the *new frame*: the corrections describe transformation laws between the comobile frame (commonly accepted terminology) to the lab frame [296], the local proper frame [174], the unadorned (as seen by an observer in the zero) frame [245], the fixed frame in [59] or the noninertial frame in [203]. With this (non-exhaustive) list of terms, we indirectly insist on the fact the material of this section can be found in several books and publications [296, 59, 203, 245, 174] depending on the physics of interest. More important than the terminology differences in these publications, the corrections are not deduced from the same computations. For example, in neutronics, they are deduced from infinitesimal physical analysis [296, 297]. In photonics [174, 59, 245, 203], they are deduced from the Lorentz invariants. In plasma physics, to our knowledge, the solution of the previous section 9.10.1 is commonly applied³⁷. The main interest of this section is to highlight that independently of the physics of interest, the corrections come from the same change of variable (on the velocity for neutronics, the frequency for photonics,...). In order to make the description the more general possible (and to treat every physics at the same time), we suggest an original (to our knowledge) way to identify the consistent corrections to the cross-sections and source term. It consists in dealing with an arbitrary acceleration term and performing (once again for an MC resolution) a particular change of variables.

³⁷This is probably due to the fact that the acceleration term is central for this physics and the linear Boltzmann equation without collision term, also denoted as Vlasov equation, is already physically relevant.

In order to describe the forementioned change of variable, let us adopt a step-by-step approach. We first consider the accelerated transport equation without collisional part and progressively introduce the different terms (total cross-section, scattering one to finish with the source term). This progressive methodology is commonly used, for example in order to prove the wellposedness of the linear Boltzmann equation in [127] or for the construction of the Boltzmann equation in [18].

The accelerated transport equation without collisional part

Let us first consider the accelerated transport equation without collisional part: its expression is

$$\partial_t f(\mathbf{x}, t, \mathbf{q}) + \mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) + a(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) = 0. \quad (9.129)$$

Let us introduce an artificial quantity, homogeneous to a velocity, such that $a(\mathbf{x}, t, \mathbf{q}) = -\partial_t \mathcal{V}(\mathbf{x}, t, \mathbf{q}), \forall \mathbf{x} \in \mathcal{D}, \mathbf{q} \in \mathbb{R}^3$. Then (9.129) rewrites

$$\partial_t f(\mathbf{x}, t, \mathbf{q}) + \mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) - \partial_t \mathcal{V}(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) = 0. \quad (9.130)$$

In this section we aim at exhibiting a change of variables allowing to rewrite (9.129) under the form

$$\partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) = 0. \quad (9.131)$$

In (9.131), we implicitly have $d_t \mathbf{v}(t) = 0$, leading to a resolution with straight trajectories. The solutions of (9.129) and (9.131) are respectively given by

$$\begin{aligned} f(\mathbf{x}(t), t, \mathbf{q}(t)) &= f_0(\mathbf{x}, \mathbf{q}), \\ u(\bar{\mathbf{x}}(t), t, \mathbf{v}(t)) &= u_0(\mathbf{x}, \mathbf{v}), \end{aligned} \quad (9.132)$$

where $\mathbf{x}(t), \mathbf{q}(t)$ are the solutions (9.123) and $\bar{\mathbf{x}}(t), \mathbf{v}(t)$ of (9.122)³⁸. We now aim at defining $\mathbf{v}(t)$ as a function of $\mathbf{q}(t)$ and \mathcal{V} so that $\frac{d\mathbf{v}(t)}{dt} = 0$. To do so, recall³⁹

$$\frac{d\mathbf{q}(t)}{dt} = a(\mathbf{x}(t), t, \mathbf{q}(t)) = -\partial_t \mathcal{V}(\mathbf{x}(t), t, \mathbf{q}(t)).$$

Consequently, it is possible to define $\mathbf{v}(t)$ as wanted noticing that

$$\frac{d\mathbf{q}(t)}{dt} + \partial_t \mathcal{V}(\mathbf{x}(t), t, \mathbf{q}(t)) = 0 = \frac{d\mathbf{v}(t)}{dt}.$$

Then $\mathbf{v}(t)$ can be defined as

$$\mathbf{v} = \mathbf{q}(t) + \int_0^t \partial_t \mathcal{V}(\mathbf{x}(s), s, \mathbf{q}(s)) ds,$$

ensuring $d_t \mathbf{v} = 0$. Recall $\mathbf{x}(t)$ depends explicitly on $\mathbf{q}(t)$, and we can rewrite without loss of generality in a more concise way⁴⁰

$$\mathbf{v}(t) = \mathbf{q}(t) + \int_0^t \partial_t \mathcal{V}(\mathbf{q}(s), s) ds = \mathbf{q}(t) + V(t).$$

Depending on the shape of \mathcal{V} with respect to the dependences, the above equation may be complex to solve without further hypothesis. In practice, assumptions (see section 9.6) can considerably simplify the computations together with capturing the regime of interest. Note that for some physics, \mathcal{V} may depend only on t and not on $\mathbf{q}(t)$ (see [297, 59]). For others, see [29, 24, 25], the velocity dependences are important. In the following paragraphs, we keep the computations the more general possible by considering $\mathbf{v}(t) = \mathbf{q}(t) + V(t)$.

Let us study the jacobian of the previous change of variable and its determinant. Note that when there is no ambiguity, the time dependence is omitted for the sake of conciseness. The easiest way to

³⁸Recall that we have $\mathbf{q}(0) = \mathbf{q}, \mathbf{v}(0) = \mathbf{v}$.

³⁹We intensively use computations of the previous section.

⁴⁰abusively dropping the dependence in $x(0) = x$ as we think there are no ambiguities here.

express the jacobian⁴¹ consists in formally introducing cartesian coordinates (v_x, v_y, v_z) of the velocity. We have

$$\mathbf{v} = \begin{pmatrix} v \cos(\theta) \\ v \sin(\theta) \cos(\phi) \\ v \sin(\theta) \sin(\phi) \end{pmatrix}, \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} q \cos(\Theta) \\ q \sin(\Theta) \cos(\Phi) \\ q \sin(\Theta) \sin(\Phi) \end{pmatrix},$$

so that

$$\frac{\partial(v_x, v_y, v_z)}{\partial \mathbf{v}} = \begin{pmatrix} \cos(\theta) & -v \sin(\theta) & 0 \\ \sin(\theta) \cos(\phi) & v \cos(\theta) \cos(\phi) & -v \sin(\theta) \sin(\phi) \\ \sin(\theta) \sin(\phi) & v \cos(\theta) \sin(\phi) & v \cos(\theta) \cos(\phi) \end{pmatrix}.$$

Furthermore, we have

$$\frac{\partial(v_x, v_y, v_z)}{\partial \mathbf{q}} = \begin{pmatrix} \cos(\Theta) & -q \sin(\Theta) & 0 \\ \sin(\Theta) \cos(\Phi) & q \cos(\Theta) \cos(\Phi) & -q \sin(\Theta) \sin(\Phi) \\ \sin(\Theta) \sin(\Phi) & q \cos(\Theta) \sin(\Phi) & q \cos(\Theta) \cos(\Phi) \end{pmatrix}.$$

We finally get

$$\left| \frac{\partial \mathbf{v}}{\partial \mathbf{q}} \right| = \left| \frac{\partial \mathbf{v}}{\partial(v_x, v_y, v_z)} \right| \times \left| \frac{\partial(v_x, v_y, v_z)}{\partial \mathbf{q}} \right| = \frac{q^2}{v^2}. \quad (9.133)$$

Exhibiting the above change of variable and its jacobian (9.133) will be convenient to track MC particles with fields expressed in a new frame. But relation (9.133) can also be particularly useful in order to compute quantities in the new frame from quantities in the comobile one: for every functional F , we have

$$\begin{aligned} \int_0^t \int F(u(\bar{\mathbf{x}}(s), s, \mathbf{v}(s))) ds d\mathbf{v} &= \int_0^t \int F(u(\bar{\mathbf{x}}(s), s, \mathbf{v}(\mathbf{q}(s)))) \left| \frac{\partial \mathbf{v}}{\partial \mathbf{q}} \right|(s) ds d\mathbf{q}, \\ &= \int_0^t \int F(f(\mathbf{x}(s), s, \mathbf{q}(s))) d\mathbf{q} ds. \end{aligned} \quad (9.134)$$

This last relation will be applied throughout the following sections. In the latter, we express the different cross-sections and source term in the new frame from their expression in the comobile one.

Total cross-section in the new frame

Now we highlighted a particular change of variable allowing to solve the transport counterpart of (9.121) with straight trajectories, it remains to identify the expressions of the cross-sections in the new frame from the ones in the comobile one given the change of variable (9.134). Let us first focus on the total cross-section. The methodology is pretty similar to the one adopted in the previous section: we aim at solving

$$\partial_t f(\mathbf{x}, t, \mathbf{q}) + \mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) + a(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) + |\mathbf{q}| \sigma_t(|\mathbf{q}|) f(\mathbf{x}, t, \mathbf{q}) = 0, \quad (9.135)$$

with the change of variable (9.134) leading to an equation of the form

$$\partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + |\mathbf{v}| \bar{\sigma}_t(|\mathbf{v}|) u(\mathbf{x}, t, \mathbf{v}) = 0. \quad (9.136)$$

We recall $|\mathbf{v}| = v$ and $|\mathbf{q}| = q$, so that $|\mathbf{q}| \sigma_t(|\mathbf{q}|) = q \sigma_t(q)$ and $|\mathbf{v}| \sigma_t(|\mathbf{v}|) = v \bar{\sigma}_t(v)$. The change of variable of this section operates on \mathbf{q} and \mathbf{v} and not only on q, v , hence the above unconventional notations. These echo the remark in [245] concerning the angular dependence of the absorption cross-section expressed in the unadorned frame in the very first chapters. In order to make sure (9.136) allows an equivalent resolution of (9.135) in a new frame, the *effective* total cross-section $\bar{\sigma}_t$ has an imposed expression. To

⁴¹Here, this is obvious for non-relativistic mechanics, less in the photonic case for example but it is possible convincing yourself by studying [245], p. 270 with expression (A.45).

identify it, we first rewrite both equations in an integral form

$$\begin{aligned} f(\mathbf{x}(t), t, \mathbf{q}(t)) &= f_0(\mathbf{x}, \mathbf{q}) e^{-\int_0^t |\mathbf{q}(s)| \sigma_t(|\mathbf{q}(s)|) ds}, \\ u(\bar{\mathbf{x}}(t), t, \mathbf{v}) &= u_0(\mathbf{x}, \mathbf{v}) e^{-\int_0^t |\mathbf{v}(s)| \bar{\sigma}_t(|\mathbf{v}(s)|) ds}. \end{aligned}$$

Now introduce an arbitrary test function $\psi(\mathbf{q}) \in \mathcal{C}_b(\mathbb{R}^3)$ so that we have (particular case of (9.134))

$$\int_0^t \int f(\mathbf{x}(s), s, \mathbf{q}(s)) \psi(\mathbf{q}(s)) ds d\mathbf{q} = \int_0^t \int u(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) \psi(\mathbf{q}(\mathbf{v}(s))) ds d\mathbf{v}.$$

Plugging the previous expressions of u and f with respect to u_0 and f_0 leads to

$$\begin{aligned} \int_0^t \int f_0(\mathbf{x}, \mathbf{q}) e^{-\int_0^s |\mathbf{q}(\alpha)| \sigma_t(|\mathbf{q}(\alpha)|) d\alpha} \psi(\mathbf{q}(s)) ds d\mathbf{q} \\ = \int_0^t \iint u(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) \psi(\mathbf{q}(\mathbf{v}(s))) \left| \frac{\partial \mathbf{v}}{\partial \mathbf{q}} \right| (s) ds d\mathbf{v}, \\ = \int_0^t \iint f_0(\mathbf{x}, \mathbf{q}) e^{-\int_0^s |\mathbf{v}(\mathbf{q}(\alpha))| \bar{\sigma}_t(|\mathbf{v}(\mathbf{q}(\alpha))|) d\alpha} \psi(\mathbf{q}(s)) ds d\mathbf{q}. \end{aligned}$$

We consequently have $\forall \psi \in \mathcal{C}_b(\mathbb{R}^3)$ with $f_0 \geq 0$,

$$\int_0^t \int f_0(\mathbf{x}, \mathbf{q}) \psi(\mathbf{q}(s)) \left(e^{-\int_0^s |\mathbf{v}(\mathbf{q}(\alpha))| \bar{\sigma}_t(|\mathbf{v}(\mathbf{q}(\alpha))|) d\alpha} - e^{-\int_0^s |\mathbf{q}(\alpha)| \sigma_t(|\mathbf{q}(\alpha)|) d\alpha} \right) ds d\mathbf{q} = 0.$$

This relation is ensured if and only if

$$v \bar{\sigma}_t(v) = q \sigma_t(q). \quad (9.137)$$

Equation (9.137) expresses the *effective* total cross-section in the new frame from the one in the comobile frame. The expression is in agreement with the one obtained by [296, 297] (effective total cross-section in neutronics) and by [59, 245, 174, 203] (covariant transformation in photonics).

Scattering cross-section in the new frame

We here aim at identifying the expression of the *effective* scattering cross-section. The idea is similar to the above calculations: we add a scattering term and use the previous relations in order to identify the consistent expression of the transformed scattering cross-section. Let us consider

$$\begin{aligned} \partial_t f(\mathbf{x}, t, \mathbf{q}) + +\mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) + a(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) = \\ -|\mathbf{q}| \sigma_t(|\mathbf{q}|) f(\mathbf{x}, t, \mathbf{q}) + \int |\mathbf{q}| \sigma_s(\mathbf{q}, \mathbf{q}') f(\mathbf{x}, t, \mathbf{q}') d\mathbf{q}', \end{aligned} \quad (9.138)$$

which we want to solve with the change of variable (9.122) leading to an equation of the form

$$\partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + |\mathbf{v}| \bar{\sigma}_t(|\mathbf{v}|) u(\mathbf{x}, t, \mathbf{v}) = \int |\mathbf{v}| \bar{\sigma}_s(\mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}'. \quad (9.139)$$

In order to make sure (9.139) allows an equivalent resolution of (9.138) in a new frame, the *effective* scattering cross-section $\bar{\sigma}_s$ has also an imposed expression. To identify it, we rewrite the two above expressions in an integral form

$$\begin{aligned} f(\mathbf{x}(t), t, \mathbf{q}(t)) &= f_0(\mathbf{x}, \mathbf{q}) e^{-\int_0^t |\mathbf{q}(s)| \sigma_t(|\mathbf{q}(s)|) ds} \\ &\quad + \int_0^t \int e^{-\int_0^s |\mathbf{q}(\alpha)| \sigma_t(|\mathbf{q}(\alpha)|) d\alpha} |\mathbf{q}(s)| \sigma_s(\mathbf{q}(s), \mathbf{q}'(s)) f(\mathbf{x}(s), s, \mathbf{q}'(s)) d\mathbf{q}' ds, \\ u(\bar{\mathbf{x}}(t), t, \mathbf{v}(t)) &= u_0(\mathbf{x}, \mathbf{v}) e^{-\int |\mathbf{v}(s)| \bar{\sigma}_t(|\mathbf{v}(s)|) ds} \\ &\quad + \int_0^t \int e^{-\int_0^s |\mathbf{v}(\alpha)| \bar{\sigma}_t(|\mathbf{v}(\alpha)|) d\alpha} |\mathbf{v}(s)| \bar{\sigma}_s(\mathbf{v}(s)) u(\bar{\mathbf{x}}(s), s, \mathbf{v}'(s)) d\mathbf{v}' ds. \end{aligned}$$

The starting point of the next computations is the same as above. For every test-function $\psi(\mathbf{q}) \in \mathcal{C}_b(\mathbb{R}^3)$, we have (particular case of (9.134))

$$\int_0^t \int f(\mathbf{x}(s), s, \mathbf{q}(s)) \psi(\mathbf{q}(s)) ds d\mathbf{q} = \int_0^t \int u(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) \psi(\mathbf{q}(\mathbf{v}(s))) ds d\mathbf{v}.$$

The idea now is to plug the expressions of the integral solutions into the previous relation, use the expression of the effective total cross-section (9.137) to get after few computations

$$\int \int_0^t \int \left[\begin{array}{l} (|\mathbf{q}(s)|\sigma_s(\mathbf{q}(s), \mathbf{q}'(s)) - |\mathbf{v}(s)|\bar{\sigma}_s(\mathbf{v}(\mathbf{q}(s)), \mathbf{v}'(\mathbf{q}'(s)))) \\ \times f(\mathbf{x}(s), s, \mathbf{q}'(s)) e^{-\int_0^s |\mathbf{q}(\alpha)|\sigma_t(|\mathbf{q}(\alpha)|) d\alpha} \psi(\mathbf{q}(s)) \end{array} \right] d\mathbf{q}' d\mathbf{q} ds = 0.$$

The above expression is true $\forall \psi(\mathbf{q}) \in \mathcal{C}_b(\mathbb{R}^3)$, hence if and only if

$$\bar{\sigma}_s(\mathbf{v}, \mathbf{v}') = \frac{q}{v} \sigma_s(\mathbf{q}, \mathbf{q}'). \quad (9.140)$$

Now it only remains to identify the corrections on the source term. It is dealt with in the next section.

Source term in the new frame

In order to express the correction one has to operate on the source term, it is enough considering equation

$$\partial_t f(\mathbf{x}, t, \mathbf{q}) + \mathbf{q} \partial_{\mathbf{x}} f(\mathbf{x}, t, \mathbf{q}) + a(\mathbf{x}, t, \mathbf{q}) \partial_{\mathbf{q}} f(\mathbf{x}, t, \mathbf{q}) = S(\mathbf{x}, t, \mathbf{q}), \quad (9.141)$$

and

$$\partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) = \bar{S}(\mathbf{x}, t, \mathbf{v}). \quad (9.142)$$

In order to make sure (9.142) allows an equivalent resolution of (9.141) in a new frame, the *effective* source term \bar{S} has an imposed expression which can be deduced from the integral forms of both equations

$$\begin{aligned} f(\mathbf{x}(t), t, \mathbf{q}(t)) &= f_0(\mathbf{x}, \mathbf{q}) + \int_0^t S(\mathbf{x}(s), s, \mathbf{q}(s)) ds, \\ u(\bar{\mathbf{x}}(t), t, \mathbf{v}(t)) &= u_0(\mathbf{x}, \mathbf{v}) + \int_0^t \bar{S}(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) ds. \end{aligned}$$

Now introduce an arbitrary test function $\psi(\mathbf{q}) \in \mathcal{C}_b(\mathbb{R}^3)$, we have (particular case of (9.134))

$$\int_0^t \int f(\mathbf{x}(s), s, \mathbf{q}(s)) \psi(\mathbf{q}(s)) ds d\mathbf{q} = \int_0^t \int u(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) \psi(\mathbf{q}(\mathbf{v}(s))) d\mathbf{v} ds.$$

Plugging the two integral expressions of u and f with respect to u_0, \bar{S} and f_0, S into the above relation together with few computations lead to

$$\begin{aligned} \int_0^t \int S(\mathbf{x}(s), s, \mathbf{q}(s)) \psi(\mathbf{q}(s)) ds d\mathbf{q} &= \int_0^t \int \bar{S}(\bar{\mathbf{x}}(s), s, \mathbf{v}(s)) \psi(\mathbf{q}(\mathbf{v}(s))) ds d\mathbf{v}, \\ &= \int \int_0^t \bar{S}(\bar{\mathbf{x}}(s), s, \mathbf{v}(\mathbf{q}(s))) \psi(\mathbf{q}(s)) \left| \frac{\partial \mathbf{v}}{\partial \mathbf{q}} \right| ds d\mathbf{q}, \\ &= \int \int_0^t \bar{S}(\bar{\mathbf{x}}(s), s, \mathbf{v}(\mathbf{q}(s))) \psi(\mathbf{q}(s)) \frac{|\mathbf{q}(s)|^2}{|\mathbf{v}(\mathbf{q}(s))|^2} ds d\mathbf{q}. \end{aligned}$$

The above expression is equivalent to

$$\int_0^t \int \left[S(\mathbf{x}(s), s, \mathbf{q}(s)) - \bar{S}(\bar{\mathbf{x}}(s), s, \mathbf{v}(\mathbf{q}(s))) \frac{|\mathbf{q}(s)|^2}{|\mathbf{v}(\mathbf{q}(s))|^2} \right] \psi(\mathbf{q}) ds d\mathbf{q} = 0.$$

This relation is ensured $\forall \psi \in \mathcal{C}_b(\mathbb{R}^3)$ if and only if

$$\bar{S}(\mathbf{v}) = S(\mathbf{q}) \frac{v^2}{q^2}. \quad (9.143)$$

This expression is in agreement with the one obtained by [296, 296] (neutronics) and by [59, 203] (photronics).

Implications of the above corrections on an MC resolution

In the previous section 9.10.1, dealing with curved trajectories, we did not details the MC resolution⁴² as the treatments were very similar as the one highlighted in the previous sections 9.2–9.3–9.4–9.5–9.9. It depended more on the choice of the MC scheme, with a backward or a forward resolution, than on taking into account the acceleration term. The only difference came from the fact that the MC particles had curves trajectories defined by (9.123) instead of straight ones (9.122).

The treatments described in this section deserve some more practical details for its MC resolution. The MC particles have straight trajectories, this is the simple part, but the cross-sections and source term must be corrected *on-the-fly* in order to take into account the acceleration term. The aim of this paragraph is to present the algorithmic implications of the above corrections on the cross-sections and sources (9.137)–(9.140)–(9.143) for an MC resolution.

Our aim here is to solve the following transport equation of unknown u having its dependences in a new frame $(\mathbf{x}, t, \mathbf{v})$

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + v \bar{\sigma}_t(v) u(\mathbf{x}, t, \mathbf{v}) = \int v \bar{\sigma}_s(\mathbf{v}, \mathbf{v}') u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}' + \bar{S}(\mathbf{x}, t, \mathbf{v}), \\ \left\{ \begin{array}{l} \mathbf{q} = q\omega(\mathbf{v}, V), \mathbf{v} = \mathbf{v}(\mathbf{q}, V) \\ v \bar{\sigma}_t(v) = q\sigma_t(q), \\ v \bar{\sigma}_s(\mathbf{v}, \mathbf{v}') = q\sigma_s(\mathbf{q}, \mathbf{q}'), \\ \bar{S}(\mathbf{x}, t, \mathbf{v}) = \frac{v^2}{q^2} S(\mathbf{x}, t, \mathbf{q}). \end{array} \right. \end{array} \right. \quad (9.144)$$

In the above expression, V is a given external field and induces corrections to the cross-sections and sources which are only known in the comobile frame. In other words, (9.144) could be rewritten in a closed form

$$\begin{aligned} \partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + |\mathbf{q}(\mathbf{v}, V)| \sigma_t(|\mathbf{q}(\mathbf{v}, V)|) u(\mathbf{x}, t, \mathbf{v}) = \\ \int |\mathbf{q}(\mathbf{v}, V)| \sigma_s(\mathbf{q}(\mathbf{v}, V), \mathbf{q}'(\mathbf{v}', V)) u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}' + \frac{v^2}{|\mathbf{q}(\mathbf{v}, V)|^2} S(\mathbf{x}, t, \mathbf{q}(\mathbf{v}, V)). \end{aligned}$$

By closed form, we mean every quantities except our unknown u are known fields $(V, \sigma_t, \sigma_s, S)$. The above equation, now, has a form which has already intensively been encountered throughout the document. It can be rewritten in an integral form, as an expectation over a set of random variables, which determine the chosen resolution scheme, in a backward or a forward manner, for its MC resolution. We do not provide the details of the computations, they can be deduced from the previous ones in the different sections (depending on the favorite MC scheme of the reader). We only highlight, in algorithm 13, where the corrections must be carried on and comment on them.

⁴²no algorithmic details presented for example.

Algorithm 13: The general canvas for the different MC schemes described in term of algorithmic operations in order to compute (direct) $U(\mathbf{x}, t) = \int u(\mathbf{x}, t, \mathbf{v}) d\mathbf{v}$ with an acceleration term. The velocity V is supposed constant in each cell in the time step (forward form).

```

1 call sampling( $N_{MC}$ ) #SAMPLING described in algorithm 5
2 set  $t = \Delta t$ 
3 while  $t < T$  do
4   #Initialize to zero the array of the quantity of interest on the whole simulation domain  $\mathcal{D}$ 
5   set  $U(\mathbf{x}, t) = 0 \forall \mathbf{x} \in \mathcal{D}$ 
6   #TRACKING: make sure each  $u_p$  is an MC particles
7   for  $p \in \{1, \dots, N_{MC}\}$  do
8     set  $s_p = t - \Delta t$  #this will be the current time of particle p
9     #corrections on the MC source particles
10    if  $p == \text{source}$  then
11      #During the sampling phase, sources emitted in the comobile frame  $\mathbf{q}_p$ 
12       $\mathbf{v}_p = \mathbf{v}(\mathbf{q}_p, V_{i_p}^n)$ 
13       $w_p \leftarrow \frac{|\mathbf{v}_p|^2}{|\mathbf{q}_p|^2} w_p$ 
14    end
15    while  $s_p < t$  and  $w_p > 0$  do
16      if  $x_p \notin \mathcal{D}$  then
17        #here a general function for the application of arbitrary boundary conditions
18        apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
19      end
20      compute  $\mathbf{q} = \mathbf{q}(\mathbf{v}_p, V_{i_p}^n)$ 
21      sample  $\tau_{inter} = \text{sample\_interaction\_time}(|\mathbf{q}|, i_p)$ 
22      compute  $\tau_{exit} = \text{compute\_cell\_exit\_time}(\mathbf{x}_p, \mathbf{v}_p, i_p)$ 
23      compute  $\tau_{census} = \max(t - \tau, 0)$ 
24      set  $\tau = \min(\tau_{exit}, \tau_{census}, \tau_{inter})$ 
25      #move the particle p
26       $\mathbf{x}_p \leftarrow \mathbf{x}_p - \mathbf{v}_p \tau$ ,
27      #change the particle weight
28      ( $K, r$ ) = compute_weight_modif( $|\mathbf{q}|, \tau, \tau_{census}, \tau_{exit}, \tau_{inter}, i_p$ )
29       $w_p \leftarrow K \times w_p$ 
30      if  $\tau == \tau_{census}$  then
31        #set the life time of particle p to zero:
32         $s_p \leftarrow t$ 
33        #tally the contribution of particle p
34         $U(\mathbf{x}_p, t) += w_p$ 
35      end
36      if  $\tau == \tau_{exit}$  then
37        #The particle p changes of cell: find its new cell number
38         $i_p = \text{find_neighbouring\_cell}(i_p, \mathbf{v}_p)$ 
39        #set the life time of particle p to:
40         $s_p \leftarrow s_p + \tau < t$ 
41      end
42      if  $\tau == \tau_{inter}$  then
43        #Sample the velocity of particle p
44         $\mathbf{Q}' = \text{sample\_velocity}(\mathbf{q}, r, i_p)$ 
45         $\mathbf{V}' = \mathbf{v}(\mathbf{Q}', V_{i_p}^n)$ 
46        set  $\mathbf{v}_p = \mathbf{V}'$ 
47        #set the life time of particle p to:
48         $s_p \leftarrow s_p + \tau < t$ 
49      end
50    end
51  end
52   $t \leftarrow t + \Delta t$ 
53 end

```

Algorithm 13 is very similar to algorithm 9 where a general canvas for MC resolutions was presented. Their main differences are highlighted (in blue) in algorithm 13. The first blue lines concern the source term corrections. The population of source MC particles is supposed to be built in the sampling phase, in the comobile frame (we only know S in this frame). The energies, angles and weights of the source particles must be corrected to represent an emitted particle in the new frame (in order to represent \bar{S}). The transformation is performed by correcting the energy q_p and angle Ω_p using the relation $\mathbf{v}_p = \mathbf{q}(\mathbf{q}_p, V_{i_p}^n)$ where $V_{i_p}^n$ is an estimation of V at time s_p , position \mathbf{x}_p in cell i_p . The weight of the source particle is multiplied by the factor $\frac{v_p^2}{q_p^2}$ in agreement with (9.143).

The described source corrections are the heaviest modifications to perform. The others resume to applying the change of variable

$$\mathbf{v} = \mathbf{v}(\mathbf{q}, V), \quad (9.145)$$

before the sampling of the interaction time and the computation of the weight modification and its inverse counterpart

$$\mathbf{q} = \mathbf{q}(\mathbf{v}, V), \quad (9.146)$$

after having sampled the scattered velocity in the case of an interaction. The different functions described in algorithms 10–11–12 are only called with argument \mathbf{q}_p instead of \mathbf{v}_p . Of course, if $a = 0$, algorithm 13 degenerates toward algorithm 9 as the change of variable $\mathbf{v} = \mathbf{v}(\mathbf{q}, V)$ and its inverse become identities.

9.11 The Uncertain Linear Boltzmann equation

In this section, we would like to tackle the combined problem of uncertainty quantification (as in part II) and the resolution of the linear Boltzmann equation (current part III). More precisely, we tackle it from two different point of views:

- in section 9.11.1 we question the necessity to eventually design numerical schemes⁴³ adapted to the uncertainty analysis we aim at carrying on.
- The matter of section 9.11.2, fully adressed in [241], shows it is sometimes very efficient *opening the black-box*, i.e. being intrusive, to perform an uncertainty analysis.

We assume the reader went through the material of part II regarding uncertainty quantification: this allows going straight to interesting uncertainty analysis considerations. As explained in part II, we explicitly introduce the dependence of the solution $u(\mathbf{x}, t, \mathbf{v})$ of the PDE of interest with respect to a random vector $X = (X_1, \dots, X_Q)^t$ modeling the uncertainty. Vector $X \in \Omega \subset \mathbb{R}^Q$ may depend on the position, the time or the velocity of the physical particles, i.e. $X = X(\mathbf{x}, t, \mathbf{v})$, $\forall \mathbf{x} \in \mathcal{D}, t \in [0, T], \mathbf{v} = v\omega \in \mathbb{R}^3$. We assume its components are independent⁴⁴ and that its probability measure is known $\forall \mathbf{x} \in \mathcal{D}, t \in [0, T], \mathbf{v} = v\omega \in \mathbb{R}^3$, denoted by $d\mathcal{P}_X = \prod_{i=1}^Q d\mathcal{P}_{X_i}$. The aim of this section is to solve the general SPDE (Cauchy problem)

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \mathbf{v}, X) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}, X) + v\sigma_t(\mathbf{x}, t, \mathbf{v}, X)u(\mathbf{x}, t, \mathbf{v}, X) = \\ \quad + \int v\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X)u(\mathbf{x}, t, \mathbf{v}', X)d\mathbf{v}' + S(\mathbf{x}, t, \mathbf{v}, X), \\ u(\mathbf{x}, 0, \mathbf{v}, X) = u_0(\mathbf{x}, \mathbf{v}, X). \end{array} \right. \quad (9.147)$$

With such general formulation, X can describe uncertainties in the initial condition, the source term, the cross-sections⁴⁵ without more distinctions.

⁴³for the black-box code.

⁴⁴We recall this is not a strong hypothesis as we can assume the correlations have been pretreated, see part II and [169, 100] for example.

⁴⁵and even in the boundary conditions which are not recalled here for the sake of conciseness.

9.11.1 Non-intrusive resolution of the uncertain linear Boltzmann equation

The non-intrusive resolution of (9.147) uses a simulation code⁴⁶ solving the linear Boltzmann equation as a black-box, run at several identified points $(X_i, w_i)_{i \in \{1, \dots, N\}}$ of an experimental design depending on the measure $d\mathcal{P}_X$, see chapter 5. Once the simulation code available, this corresponds to the most direct, simple and fast way to tackle the uncertain counterpart of the numerical resolution of (9.147).

While performing an uncertainty quantification problem, we tacitly demand conditions to hold:

- From a mathematical point of view, we suppose $\forall X \in \Omega \subset \mathbb{R}^Q$, the wellposedness of the linear Boltzmann equation is not questionned.
- From a numerical point of view, we assume $\forall X \in \Omega \subset \mathbb{R}^Q$ the resolution scheme of the black-box code can accurately capture the regime of interest.

Without any of these two conditions, the uncertainty analysis may be pointless [153]. The first condition and its implications have been intensively studied in chapters 4–5 (for systems of conservation laws). The second may deserve an example, this is the matter of this section. Suppose there exists a subspace $\mathcal{D}_{\delta \rightarrow 0} \subset \Omega$ such that the probability measure of $\{X \in \mathcal{D}_{\delta \rightarrow 0}\}$ is non-zero. Assume furthermore that every realisations of X in $\mathcal{D}_{\delta \rightarrow 0}$ induce the exploration of a stiff regime characterised by $\delta \rightarrow 0$ (see remark 9.1) for the numerical (MC) resolution. If the (deterministic) resolution scheme of the black-box code is too sensitive with respect to its discretisation parameter in this regime (i.e for example not Asymptotic Preserving, see definition 9.1), the numerical error for some realisations of $X \in \mathcal{D}_{\delta \rightarrow 0}$ may overcome the variability of the uncertain parameters⁴⁷. The interpretations and the conclusions of the uncertainty analysis may then be wrong.

Let us apply non-intrusive gPC to solve (9.147) in a simple configuration. It is monokinetic and homogeneous, so that an analytical solution is available. We assume the uncertainty, one-dimensional here for the sake of simplicity, affects the scattering cross-sections

$$\sigma_s = \bar{\sigma}_s + \hat{\sigma}_s X.$$

In the above expression, $X \sim \mathcal{U}[-1, 1]$ and $\hat{\sigma}_s = \frac{1}{10}\bar{\sigma}_s$: the uncertainty affects the scattering cross-section by a factor 10% relative to its mean. Let us build the analytical solution for this test-problem. In the monokinetic configuration, the uncertain linear Boltzmann equation resumes to

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega, X) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega, X) + v \sigma_t(\mathbf{x}, t) u(\mathbf{x}, t, \omega, X) = \int v \sigma_s(\mathbf{x}, t, X) u(\mathbf{x}, t, \omega', X) d\omega', \\ u(\mathbf{x}, 0, \omega) = u_0(\mathbf{x}, \omega). \end{cases} \quad (9.148)$$

In a homogeneous test-problem, the solution is given by $U(t, X)$ solution of the uncertain ODE

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega, X) + \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega, X) + v \sigma_t(\mathbf{x}, t) u(\mathbf{x}, t, \omega, X) = \\ v \sigma_s(\mathbf{x}, t, X) \int P_s(\mathbf{x}, t, \omega, \omega', X) u(\mathbf{x}, t, \omega', X) d\omega', \\ u(\mathbf{x}, 0, \omega) = u_0(\mathbf{x}, \omega). \end{cases} \quad (9.149)$$

It is given by

$$U(t, X) = U_0 e^{-v\sigma_a(X)t} = U_0 e^{-v(\sigma_t - \bar{\sigma}_s - \hat{\sigma}_s X)t} = U_0 e^{-v(\bar{\sigma}_a - \hat{\sigma}_s X)t}. \quad (9.150)$$

The quantity $U(t, X)$ is a random variable indexed by time t , i.e. it is a stochastic process. In this case, mean and variance of the stochastic process (9.150) can be computed analytically and are given by

$$\begin{aligned} M_1^U(t) &= \mathbb{E}[U(t, X)] = \frac{1}{2} U_0 e^{-v\bar{\sigma}_a t} \frac{e^{v\hat{\sigma}_s t} - e^{-v\hat{\sigma}_s t}}{\hat{\sigma}_s t v}, \\ M_2^U(t) &= \mathbb{E}[U^2(t, X)] = \frac{1}{4} U_0^2 e^{-2v\bar{\sigma}_a t} \frac{e^{2v\hat{\sigma}_s t} - e^{-2v\hat{\sigma}_s t}}{\hat{\sigma}_s t v}, \\ \mathbb{V}[U](t) &= M_2^U(t) - (M_1^U(t))^2. \end{aligned} \quad (9.151)$$

⁴⁶based on an MC resolution with one of the MC scheme described in this part.

⁴⁷i.e. some realisations may need finer discretisations for the uncertainty analysis to be relevant.

Of course, higher order moments, probability of failure, complete characterisation of the pdf of the stochastic process can be made but we first focus on the two first moments, the mean $\mathbb{E}[U](t)$ and the variance $\mathbb{V}[U](t)$. Let us discretise $(X, d\mathcal{P}_X)$ with N_{MC}^{UQ} Monte-Carlo points. Every runs of the black-box code solving the deterministic linear Boltzmann equation (with the non-analog or the semi-analog MC scheme) have N_{MC} MC particles. Figure 9.3 presents a convergence study with respect to

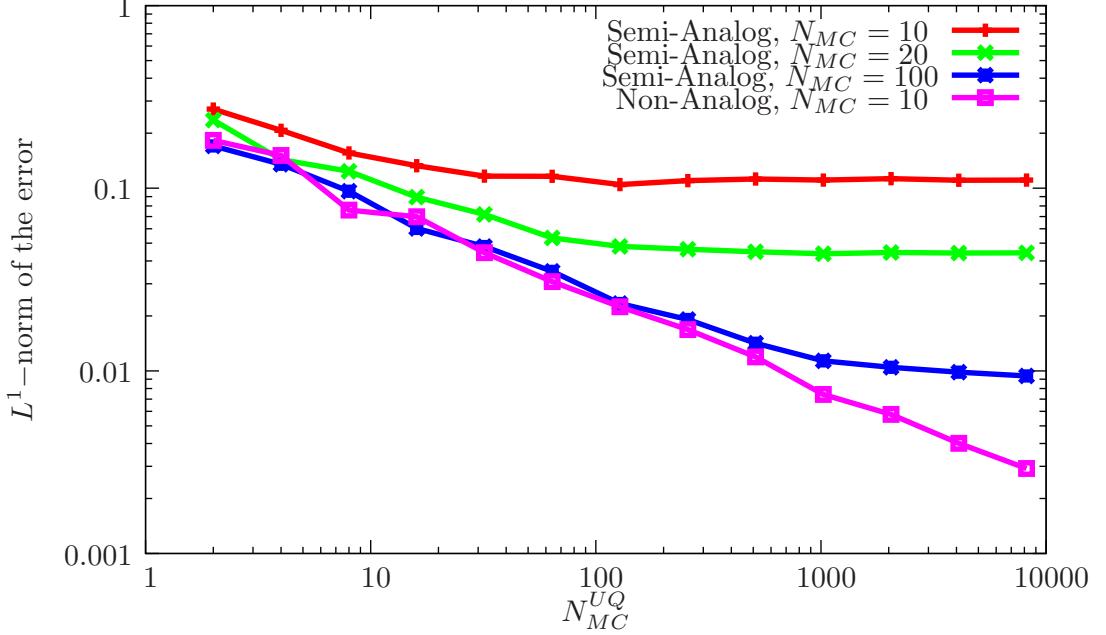


Figure 9.3: Convergence studies with respect to N_{MC}^{UQ} for the resolution of the uncertain linear Boltzmann equation in a homogeneous configuration. Comparison of the convergence studies for the semi-analog scheme for several N_{MC} and of the non-analog scheme.

$N_{MC}^{UQ} = 1, 2, \dots, 2^{13} = 8192$ performed in the previous configuration with the semi-analog MC scheme of section 9.3 and the non-analog one of section 9.4 for a fixed discretisation⁴⁸ $N_{MC} = 10, 20, 100$. The L^1 -norm of the error (i.e. on the variance) is averaged on 10 times of interest uniformly distributed (i.e. $\Delta t = 1$) in the time interval $[0, T = 10]$.

The three first curves (red, green, blue) are obtained with the semi-analog MC scheme for different N_{MC} (i.e $N_{MC} = 10, 20$ and 100). The curves all present two regimes:

- as N_{MC}^{UQ} increases, the L^1 -norm of the error decreases with the characteristic slope $-\frac{1}{2}$ of the MC method in the log-log plot of figure 9.3,
- until it reaches a plateau, after $N_{MC}^{UQ} \approx 32$ for the red curve ($N_{MC} = 10$), $N_{MC}^{UQ} \approx 128$ for the green one ($N_{MC} = 20$) and $N_{MC}^{UQ} \approx 1024$ for the blue one ($N_{MC} = 100$).

For $N_{MC} = 20$, for example, for the semi-analog MC scheme, using $N_{MC}^{UQ} = 128$ MC points or 8192 does not improve the accuracy of the approximation. This is due to the too important N_{MC} -error of the semi-analog scheme for $N_{MC}^{UQ} > 128$. It is even possible to compare the errors of the three plateaus and recover roughly an $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$ convergence rate. In other words we experimentally recover the fact the error has the general form

$$e_{L^1} = \mathcal{O}\left(\frac{1}{\sqrt{N_{MC}}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{N_{MC}^{UQ}}}\right). \quad (9.152)$$

⁴⁸ $N_{MC} = 10, 20, 100$ may seem very low but this is enough in order to illustrate our purpose in this configuration.

Once N_{MC}^{UQ} is important enough, it becomes negligible with respect to the error of the semi-analog scheme. For the green curve for example, when $N_{MC}^{UQ} > 128$, the computed variance is numerical⁴⁹ and not representative of the variability due to the uncertainty. This is closely related to the fact the constant $\sigma_{\text{semi-analog}}$, given by (9.82) in this homogeneous regime, strongly depends on the values of $\sigma_s(X)$ which fluctuate due to the uncertainty. To make sure the accuracy with N_{MC}^{UQ} for the semi-analog scheme is improved, N_{MC} must depend on the uncertain parameter X . On another hand, with the non-analog MC scheme, Asymptotic-Preserving in this same configuration (homogeneous, see remark 9.1), increasing N_{MC}^{UQ} always improves the quality of the approximation. With this scheme, care has been taken to have a numerical scheme adapted to the uncertainty quantification problem of interest as $\frac{\sigma_{\text{non-analog}}(X)}{\sqrt{N_{MC}}} = \frac{\sigma_{\text{non-analog}}}{\sqrt{N_{MC}}}$. This example emphasizes the *importance of identifying a relevant scheme for the regime of interest before tackling any uncertainty quantification problem*. Of course, the problem here is very simple but is representative of the difficulties encountered on real industrial simulations: the different numerical errors must be balanced for an efficient study.

The above conclusion of this study has indirectly already been emphasized in previous examples and convergence studies of part II. For example, replace

- $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}}}\right)$ by $\mathcal{O}(\Delta x)$,
- and $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}^{UQ}}}\right)$ by $\mathcal{O}(\exp(-P^k))$,

and we obtain the same content as remark 4.2. In other words, we can expect a non-intrusive resolution of the uncertain linear Boltzmann equation to depend on three discretisation parameters, N , N_{MC} and P (the size of the obtained P -truncated reduced model). This point is more developed in [241]. In the context of this chapter, there is one difference with the examples of part II. *We may be able to take advantage of the MC resolution to treat the uncertainties on-the-fly during the MC tracking.* The idea is to capitalize on the fact MC methods are insensitive to an increase of dimensions (the uncertain ones) and weaken the sensitivity to dimension of gPC. The idea would be to obtain a resolution method depending on one less discretisation parameter (dependence only with respect to N_{MC}, P). The description of such resolution scheme is the purpose of the next section.

9.11.2 A gPC-intrusive Monte-Carlo scheme for the uncertain linear Boltzmann equation

The previous section showed a non-intrusive application has one important drawback. The dependence of the accuracy of the computations to three intricated parameters. This is especially true if the uncertainty analyst has not designed an Asymptotic-Preserving scheme in agreement with the (uncertain) regime he wants to capture in his study. The idea here is: suppose one has to design a new adapted Asymptotic-Preserving MC scheme, is it possible to make the MC scheme solve *on-the-fly* the uncertain counterpart of the linear Boltzmann equation? The answer, positive, is summed up below and fully addressed in [241]⁵⁰. Let us consider the following uncertain transport equation

$$\begin{aligned} \partial_t u(\mathbf{x}, t, \mathbf{v}, X) + \mathbf{v} \cdot \nabla u(\mathbf{x}, t, \mathbf{v}, X) = & -v\sigma_t(\mathbf{x}, t, \mathbf{v}, X)u(\mathbf{x}, t, \mathbf{v}, X) \\ & + v\sigma_s(\mathbf{x}, t, \mathbf{v}, X) \int P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X)u(\mathbf{x}, t, \mathbf{v}', X) d\mathbf{v}'. \end{aligned} \quad (9.153)$$

Recall we have the notations

$$\begin{aligned} \sigma_s(\mathbf{x}, t, \mathbf{v}, X) &= \int \sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X) d\mathbf{v}', \\ P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X) &= \frac{\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X)}{\sigma_s(\mathbf{x}, t, \mathbf{v}, X)}. \end{aligned} \quad (9.154)$$

⁴⁹given by (9.86) with cross-sections depending on X .

⁵⁰In [241], we presented the construction of the gPC based semi-analog MC scheme, we here build the gPC based non-analog one. This way, this document and [241] remain complementary.

Let us go through the same steps as in sections 9.2–9.3–9.4 in which we built MC schemes but having in mind the quantities depend also on X . We aim at identifying the changes one must perform to the different samplings to take X into account *on-the-fly* during the MC resolution. First, we introduce

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s, X) ds = \mathbf{1}_{[0, \infty]}(s) v \sigma_s(\mathbf{x} - \mathbf{v}s, t - s, v, X) \exp\left(-\int_0^s v \sigma_\alpha(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, \mathbf{v}, X) d\alpha\right) ds. \quad (9.155)$$

Under some boundedness conditions⁵¹ $\forall X \in \text{Supp}(X)$, where $\text{Supp}(X)$ denotes the support of the random variable, (9.155) remains an exponential probability measure [219]. The uncertain counterpart of (9.28) is then given by

$$u(\mathbf{x}, t, \mathbf{v}, X) = \int \begin{bmatrix} \mathbf{1}_{[t, \infty]}(s) & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}, X) & e^{-\int_0^s v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} \\ \mathbf{1}_{[0, t]}(s) & u(\mathbf{x} - \mathbf{v}s, t - s, \mathbf{v}', X) & e^{-\int_0^s v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} \\ \times f_\tau(\mathbf{x}, t, \mathbf{v}, s, X) d\mathbf{v}' ds. \end{bmatrix} \quad (9.156)$$

Introduce the set of random variables τ_X, \mathcal{V}_X sampled from the probability measures $\tau_X \sim f_\tau(\mathbf{x}, t, \mathbf{v}, s, X) ds$ and $\mathcal{V}_X \sim P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X) d\mathbf{v}'$. The above integral equation rewritten as a recursive expectation becomes

$$u(\mathbf{x}, t, \mathbf{v}, X) = \mathbb{E} \begin{bmatrix} +\mathbf{1}_{[t, \infty]}(\tau_X) & e^{-\int_0^t v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} & u_0(\mathbf{x} - \mathbf{v}t, \mathbf{v}, X) \\ +\mathbf{1}_{[0, t]}(\tau_X) & e^{-\int_0^{\tau_X} v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} & u(\mathbf{x} - \mathbf{v}\tau_X, t - \tau_X, \mathcal{V}_X, X) \end{bmatrix}. \quad (9.157)$$

The next step consists in introducing an MC discretization allowing to take into account the uncertain variables. Let us introduce an 'uncertain MC particle' u_p defined as

$$u_p(\mathbf{x}, t, \mathbf{v}, X) = u_p(\mathbf{x}, t, \mathbf{v}) \delta_X(X_p(t)) = w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)) \delta_X(X_p(t)). \quad (9.158)$$

We are now going to identify the operations we must perform to ensure (9.158) is solution of (9.153). For this, we plug (9.158) into (9.157) and make sure $u_p(\mathbf{x}, t, \mathbf{v}, X)$ is a particular solution of (9.153). Plugging u_p into (9.157) leads to the construction of a (compatible) system of equations of unknowns $w_p(t), \mathbf{x}_p(t), \mathbf{v}_p(t), X_p(t)$ given by

$$\begin{cases} w_p(t) = \mathbf{1}_{[t, \infty]}(\tau_X) e^{-\int_0^t v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} w_p(0) & +\mathbf{1}_{[0, t]}(\tau_X) e^{-\int_0^{\tau_X} v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X) d\alpha} w_p(t - \tau_X), \\ \mathbf{x}_p(t) = \mathbf{1}_{[t, \infty]}(\tau_X) (\mathbf{x}(0) + \mathbf{v}t) & +\mathbf{1}_{[0, t]}(\tau_X) (\mathbf{x}_p(t - \tau_X) + \mathbf{v}\tau_X), \\ \mathbf{v}_p(t) = \mathbf{1}_{[t, \infty]}(\tau_X) \mathbf{v} & +\mathbf{1}_{[0, t]}(\tau_X) (\mathbf{v}_p(t - \tau_X) = \mathcal{V}_X), \\ X_p(t) = \mathbf{1}_{[t, \infty]}(\tau_X) X_p(0) & +\mathbf{1}_{[0, t]}(\tau_X) (X_p(t - \tau_X)). \end{cases} \quad (9.159)$$

Let us focus on the last equation: inconditionally with respect to time t , $X_p(t)$ is not modified. Indeed, if $\tau_X < t$ we have $X_p(t) = X_p(t - \tau_X)$ until, events after events, the initial condition is reached leading to $X_p(t) = X_p(0) = X_p$.

Remark 9.3 *The latter result tells the uncertain variable must be sampled initially for every MC particles and remain unchanged. It also implies an MC particle must transport amongst its attributes the realisation of a random vector of size Q . This has some impact on the memory consumption of the algorithm.*

Now we know $X_p(t) = X_p$, (9.159) reduces to

$$\begin{cases} w_p(t) = \mathbf{1}_{[t, \infty]}(\tau_{X_p}) e^{-\int_0^t v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X_p) d\alpha} w_p(0) & +\mathbf{1}_{[0, t]}(\tau_{X_p}) e^{-\int_0^{\tau_{X_p}} v \sigma_a(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, v, X_p) d\alpha} w_p(t - \tau_{X_p}), \\ \mathbf{x}_p(t) = \mathbf{1}_{[t, \infty]}(\tau_{X_p}) (\mathbf{x}(0) + \mathbf{v}t) & +\mathbf{1}_{[0, t]}(\tau_{X_p}) (\mathbf{x}_p(t - \tau_{X_p}) + \mathbf{v}\tau_{X_p}), \\ \mathbf{v}_p(t) = \mathbf{1}_{[t, \infty]}(\tau_{X_p}) \mathbf{v} & +\mathbf{1}_{[0, t]}(\tau_{X_p}) (\mathbf{v}_p(t - \tau_{X_p}) = \mathcal{V}_{X_p}). \end{cases} \quad (9.160)$$

Recall $\tau_{X_p}, \mathcal{V}_{X_p}$ are sampled from the probability measures $\tau_{X_p} \sim f_\tau(\mathbf{x}, t, \mathbf{v}, s, X_p) ds$ and $\mathcal{V}_{X_p} \sim P_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X_p) d\mathbf{v}'$. System (9.160) is similar to system (9.24) but the samplings depending on x_p may need few comments. Assume, for the sake of simplicity, the cross-sections do not depend on \mathbf{x}, t

⁵¹Here, the hypothesis $\forall X \in \text{Supp}X$ is certainly not optimal but sufficient for the property to hold.

locally⁵² (i.e. within a cell or an element of geometry). Then the probability measure (9.155) for the sampling of the interaction time resumes to

$$f_\tau(\mathbf{v}_p, s, X_p) ds = \mathbf{1}_{[0, \infty]}(s) v \sigma_s(\mathbf{v}_p, X_p) e^{-\mathbf{v}_p \sigma_s(\mathbf{v}_p, X_p)s} ds. \quad (9.161)$$

In practice, this implies sampling τ_{X_p} according to⁵³

$$\tau_{X_p} = -\frac{\ln(\mathcal{U})}{v \sigma_s(\mathbf{v}_p, X_p)} \text{ where } \mathcal{U} \sim \mathcal{U}([0, 1]) \text{ and } X_p \text{ is an embedded particle field just as } \mathbf{v}_p. \quad (9.162)$$

Expression (9.162) echoes (9.55). The same apply to the sampling of the outer⁵⁴ velocity \mathcal{V}_{X_p} and to the weight modification of the uncertain MC particles. The cross-sections at play in (9.154)–(9.160) must be used at both the physical ($\mathbf{x}_p, \mathbf{v}_p$) and uncertain (X_p) fields of the uncertain MC particle.

Now the gPC coefficients can easily be estimated thanks to the uncertain MC particles: the scheme once again intensively uses the linearity of equation (9.153) together with the linearity of the P –truncated gPC approximation defined *via* in corollary 3.2. Indeed, $(u_p(\mathbf{x}, t, \mathbf{v}, X) \phi_k^X(X))_{p \in \{1, \dots, N_{MC}\}}, \forall k \in \{0, \dots, P\}$ are independent solutions of the projection of the solution of (9.153) onto a P –truncated gPC basis. This implies the sum over the number of uncertain MC particles verifies $\forall k \in \{0, \dots, P\}$

$$\sum_{p=1}^{N_{MC}} u_p(\mathbf{x}, t, \mathbf{v}, X) \phi_k^X(X) \approx u_k^X(\mathbf{x}, t, \mathbf{v}).$$

Applying the operations related to (9.159) to any given uncertain MC particles ensures, by construction (see theorem 3.2.1 of [165]), the convergence of the MC solver toward the projection of the solution of (9.153) onto the truncated gPC basis in the limit $N_{MC} \rightarrow \infty$. The overall cost remains $\mathcal{O}(N_{MC})$ together with an $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$ accuracy on the gPC coefficients to compute. Note that the computation of

⁵²This assumption is commonly done.

⁵³The next expression is obtained inverting the cumulative density function of an exponential law, this is common in MC computations see [165].

⁵⁴Inner would be more appropriate as we are here identifying the backward samplings.

the gPC coefficients explicitly appears in the algorithmic presentation 14 of the new MC scheme.

Algorithm 14: The gPC-intrusive MC non-analog scheme described in term of algorithmic operations in order to compute (backward) the gPC coefficients of $u(\mathbf{x}, t, \mathbf{v}, X)$. The differences with the classical canvas are highlighted in blue: they concern the initial sampling (X_p from $d\mathcal{P}_X$), the **tallying** (estimating the gPC coefficients $(u_k^X(\mathbf{x}, t, \mathbf{v}))_{k \in \{0, \dots, P\}}$ intrusively) and the different calls to cross-sections.

```

1 for  $k \in \{0, \dots, P\}$  do
2   | set  $u_k^X(\mathbf{x}, t, \mathbf{v}) = 0$ 
3 end
4 for  $p \in \{1, \dots, N_{MC}\}$  do
5   | set  $s_p = t$  #this will be the remaining life time of particle  $p$ , it must go down to zero
     (backward)
6   | set  $\mathbf{x}_p = \mathbf{x}$ 
7   | set  $\mathbf{v}_p = \mathbf{v}$ 
8   | set  $w_p = \frac{1}{N_{MC}}$ 
9   | Sample  $X$  from the distribution having probability measure  $d\mathcal{P}_X$ .
10  | set  $X_p = X$ 
11  while  $s_p > 0$  and  $w_p > 0$  do
12    | if  $x_p \notin \mathcal{D}$  then
13      |   | #here a general function for the application of arbitrary uncertain boundary
         |   | conditions
14      |   | apply_boundary_conditions( $x_p, s_p, \mathbf{v}_p, X_p$ )
15    | end
16    | Sample  $\tau$  from the distribution having probability measure  $f_\tau(\mathbf{x}_p, s_p, s, \mathbf{v}_p, X_p)ds$ .
17    | if  $\tau > s_p$  then
18      |       | #see the treatment in factor of  $\mathbf{1}_{[t, \infty[}(\tau)$  in (9.160)
19      |       | #change the particle weight
20      |       |  $w_p \leftarrow e^{-\int_0^{s_p} v_p \sigma_a(\mathbf{x}_p - \mathbf{v}_p \alpha, s_p - \alpha, \mathbf{v}_p, X_p) d\alpha} w_p$ 
21      |       | #move the particle  $p$ 
22      |       |  $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p s_p$ ,
23      |       | #set the life time of particle  $p$  to zero:
24      |       |  $s_p \leftarrow 0$ 
25      |       | #tally the contribution of particle  $p$ 
26      |       | for  $k \in \{0, \dots, P\}$  do
27        |       |   |  $u_k^X(\mathbf{x}, t, \mathbf{v}) += w_p \times u_0(\mathbf{x}_p, \mathbf{v}_p, X_p) \phi_k^X(X_p)$ 
28      |       | end
29    | end
30  else
31    |       | #see the recursive treatment in factor of  $\mathbf{1}_{[0, t]}(\tau)$  in (9.160)
32    |       | #move the particle  $p$ 
33    |       |  $\mathbf{x}_p \leftarrow \mathbf{x}_p + \mathbf{v}_p \tau$ ,
34    |       | #change its weight
35    |       |  $w_p \leftarrow e^{-\int_0^\tau v_p \sigma_a(\mathbf{x}_p - \mathbf{v}_p \alpha, s_p - \alpha, \mathbf{v}_p, X_p) d\alpha} w_p$ 
36    |       | Sample the velocity of particle  $p$  from  $P_s(\mathbf{x}_p, s_p, \tau, \mathbf{v}_p, \mathbf{v}', X_p) d\mathbf{v}'$ 
37    |       |  $\mathbf{v}_p = \mathbf{v}'$ 
38    |       | #set the life time of particle  $p$  to:
39    |       |  $s_p \leftarrow s_p - \tau > 0$ 
40  end
41 end
42 end

```

From the previous description, one must understand the basic idea is to try to avoid a tensorisation between the N experimental design points concerning the random variable X and the N_{MC} samplings related to the physical variables $(\mathbf{x}, t, \mathbf{v})$ for the MC particles. This imposes some identified operations

to perform a full MC approximation of the gPC coefficient with N_{MC} samplings in the whole space of variables $(\mathbf{x}, t, \mathbf{v}, X)$. We intensively make use of the insensitiveness of an MC integration with respect to dimension to compute the gPC coefficient of any given output of interest. The new MC scheme is intrusive in the sense one must modify⁵⁵

- the attributes of the MC particles to take into account a discretisation $(X_p)_{p \in \{1, \dots, N_{MC}\}}$ of $(X, d\mathcal{P}_X)$,
- the call to the cross-sections at those points $(X_p)_{p \in \{1, \dots, N_{MC}\}}$ to sample the interaction time, the outer velocity and modify the weight of any uncertain MC particle,
- the tallies (to embed the computations of the gPC coefficients and other outputs of interest).

The last point may deserve few more details: to approximate any ouput of interest, the post-treatment must also be embedded in the MC resolution. Any other quantity of interest will not directly be available unless every fields of the uncertain MC particles are *tracked in some files to be post-treated*. We clearly want to avoid such solution because tracking down information with such frequency (many tallies⁵⁶ of MC particles per seconds leading to an important volume of I/O⁵⁷) slows drastically down the computations and can easily make a filesystem collapse. More details are given in the numerical examples of [241].

At this stage of the discussion, one may also wonder why relying on gPC and consequently remaining sensitive to the dimension Q via the increasing number⁵⁸ of coefficient $(u_k^X)_{k \in \{0, \dots, P\}}$ to be evaluated. To give an element of answer, let us build the PDE satisfied by the moment of order 2 of u , solution of (9.147). It is defined by

$$\begin{aligned} M_2(\mathbf{x}, t, \mathbf{v}) &= \int u^2(\mathbf{x}, t, \mathbf{v}, X) d\mathcal{P}_X, \\ &= \int m_2(\mathbf{x}, t, \mathbf{v}, X) d\mathcal{P}_X. \end{aligned}$$

It certainly corresponds to one of the simplest statistical observable. In this case, quantity m_2 is solution⁵⁹ of

$$\begin{aligned} \partial_t m_2(\mathbf{x}, t, \mathbf{v}, X) + \mathbf{v} \cdot \nabla m_2(\mathbf{x}, t, \mathbf{v}, X) &= -2v\sigma_t(\mathbf{x}, t, \mathbf{v}, X)m_2(\mathbf{x}, t, \mathbf{v}, X) \\ &\quad + 2u(\mathbf{x}, t, \mathbf{v}, X) \int v\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X)u(\mathbf{x}, t, \mathbf{v}', X)d\mathbf{v}'. \end{aligned} \tag{9.163}$$

The latter equation is nonlinear (see the scattering term). The difficulty to solve (9.163) with an MC method can be compared to the one to solve the quadratic Boltzmann equation [37, 28] for example. In other words, to be solved numerically, it may

- either need an additional linearisation hypothesis. For example, for a Nanbu-like [37] resolution, this implies relying on a time step discretisation and an MC resolution of the **explicated** equation

$$\begin{aligned} \partial_t m_2(\mathbf{x}, t, \mathbf{v}, X) + \mathbf{v} \cdot \nabla m_2(\mathbf{x}, t, \mathbf{v}, X) &= -2v\sigma_t(\mathbf{x}, t, \mathbf{v}, X)m_2(\mathbf{x}, t, \mathbf{v}, X) \\ &\quad + 2u(\mathbf{x}, t^n, \mathbf{v}, X) \int v\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}', X)u(\mathbf{x}, t, \mathbf{v}', X)d\mathbf{v}'. \end{aligned} \tag{9.164}$$

In the above expression, $u(\mathbf{x}, t^n, \mathbf{v}, X)$ is the approximated solution at the beginning of the time step Δt . In other words the convergence depends on N_{MC} but also on the time step Δt .

- Or perform a splitting of operator between the streaming part and the collisional one with an adaptation of Bird's algorithm [28]. This splitting, even if having good mathematical properties (conservations), also introduces a dependence with respect to a time step Δt .
- Or apply an analog MC scheme and keep track of the count rate to take correlations and higher moments into account [52]. This is usually done in a file which must be post-treated (binning and linear fit etc. see [52]). But analog schemes are known to have a slower convergence rate, to be computationally intensive and inadapted to very multiplicative media (in this case the size of the written file is known to explode).

⁵⁵This is easier to understand thanks to the algorithmic representations 3 and 14.

⁵⁶See algorithm 3 for the definition of tallying.

⁵⁷I/O refers to input/output.

⁵⁸The number of coefficients increases with Q .

⁵⁹Multiply (9.147) by $u(\mathbf{x}, t, \mathbf{v}, X)$ to obtain (9.163).

Of course, the above list of alternatives may not be exhaustive. Anyway, the application of gPC does introduce a new parameter (P rather than a time step as in [37] or [28]) but the modifications to an existing solver are minor (compare algorithms 3 and 14) and the approximation with respect to P can even be expected to yield spectral convergence for smooth solutions (in fact, spectral convergence has been proved in this context in [235]). In the example above, a gPC-i-MC approximation of M_2 is simply given by

$$M_2(\mathbf{x}, t, \mathbf{v}) = \sum_{k=0}^{\infty} (u_k^X(\mathbf{x}, t, \mathbf{v}))^2 \approx \sum_{k=0}^P (u_{k, N_{MC}}^X(\mathbf{x}, t, \mathbf{v}))^2,$$

where $\forall k \in \{0, \dots, P\}$ we have

$$u_k^X(\mathbf{x}, t, \mathbf{v}) \approx u_{k, N_{MC}}^X(\mathbf{x}, t, \mathbf{v}) = \sum_{p=1}^{N_{MC}} w_p(t) \delta_{\mathbf{x}}(\mathbf{x}_p(t)) \delta_{\mathbf{v}}(\mathbf{v}_p(t)) \phi_k^X(X_p).$$

In [241], we numerically verify and illustrate the previous points and even consider more elaborated statistical outputs of interest (in particular Sobol indices for sensitivity analysis, see [266, 145, 255]). We also put forward very important gains on the new gPC based MC scheme designed in this section, compared to a non-intrusive gPC application.

To end this section, we revisit the uncertainty quantification problem tackled at the beginning of this section with the new gPC based semi-analog MC scheme. The results are displayed figure 9.4: the new

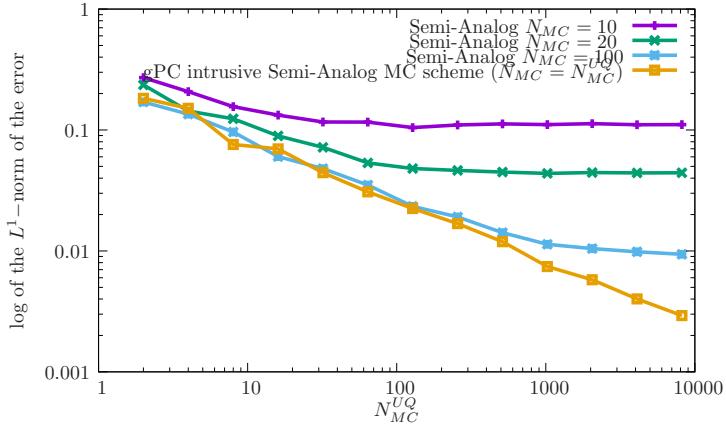


Figure 9.4: Convergence studies with respect to $N = N_{MC}^{UQ}$ and $\Delta = N_{MC}$ as in figure 9.3 (bottom right) together with a curve obtained with the new gPC-i-MC scheme. For the latter, the uncertain parameters are sampled within the N_{MC} MC particles.

method also allows avoiding the kinks in the convergence curves.

9.11.3 Summary

The two previous sections could be (restrictively) summed up as the presentation of two resolution schemes avoiding the kinks in the curves of figures 9.3–9.4. To avoid those kinks, the first method (section 9.11.1) is non-intrusive but implies developing the relevant numerical schemes (the non-analog one here) within the simulation device. In a sense, it is as intrusive as the second one (section 9.11.2), which achieves the same purpose and even has some other interesting features, see [241].

More broadly, in this section, we tackled the resolution of the uncertain linear Boltzmann equation. The first paragraph mainly aimed at putting forward the importance of having a relevant resolution scheme for the uncertain study of interest *prior* to performing an uncertainty analysis. In the second paragraph, we presented a *gPC based intrusive MC scheme*. It takes advantage of the MC resolution and its insensitiveness to an increase in the dimension to treat *on-the-fly* during the MC tracking the uncertain counterpart. It only needs enriching the fields of each MC particles and instrumenting the tracking to compute the statistical observables of interest (high order moments, histograms, Sobol indices

etc.). More details and HPC considerations can be found in [241].

9.12 Application of gPC for MC accelerations for the linear Boltzmann equation

It is difficult talking about MC methods without hinting at *variance reduction* or *acceleration* techniques. MC schemes, in opposition to deterministic ones, are even often chosen for their agility for decreasing the constant K multiplying the convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{N_{MC}}}\right) = \frac{K}{\sqrt{N_{MC}}}$. Under some conditions, K can be assimilated with the standard deviation of the MC resolution. Variance reduction techniques are presented and quite well described in many books, [165, 256, 173], and we do not intend to be exhaustive here. Our aim is to put into perspective the materials of part II and part III. The originality of this section comes from

- the analogy made with Asymptotic Preserving MC schemes (see remark 9.1),
- the introduction of gPC developments to accelerate MC computations (see part II).

Those two points are developed in the following paragraphs but to understand the stakes, at this point of the discussion, it may be more relevant to give a simple example. Let us consider a toy integration problem. Suppose one wants to compute

$$I = \int_0^1 \exp(x) dx, \quad (9.165)$$

with an MC method. Of course, the analytical solution is known, equal to $I = \exp 1 - 1 \approx 1.7183$. The

$I = 1.7183$	Results	$\frac{\text{Std}}{\sqrt{n}}$	Comparison
Classical MC ($n = 1000$)	1.6991	$std_{MC} = 0.0153$	reference
"Good" VR ($n = 1000$)	1.7093	0.0064	$< std_{MC}$
"Bad" VR ($n = 1000$)	1.7500	0.0299	$> std_{MC}$

Table 9.1: The table compares the Classical MC method, a variance reduction (VR) method (control variate) with reduced model $x \rightarrow 1+x$ ("Good"), the same VR method with reduced model $x \rightarrow 1+5x$ ("Bad"). We used $n = 1000$ samples. The abbreviation Std is for the estimated standard deviation, see section 5.2.1.

principle of MC method for simple integration has already been presented in chapter 5 together with its asymptotic properties, see section 5.2.1. Applying an MC scheme consists in introducing $X \sim \mathcal{U}([0, 1])$ and sampling N_{MC} points $(X_i)_{i \in \{1, \dots, N_{MC}\}}$ independent identically distributed (i.i.d.) according to X in order to approximate I by

$$I = \int_0^1 u(x) dx = \mathbb{E}[u(X)] \approx I_{N_{MC}} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} u(X_i). \quad (9.166)$$

The quantity $I_{N_{MC}}$ is called an estimator and its properties will be investigated later on. Table 9.1 sums up the results obtained with the classical MC method, a variance reduction (VR) method (which will be detailed later on) and a badly tuned VR method on toy problem (9.165). Both VR methods rely on the introduction of *a priori* information through a *reduced model* (detailed later on). The results are simple: if the reduced model is not well enough chosen, the results can be worse than in the simple MC case. Figure 9.5 illustrates why the second variance reduction approach fails. The reduced model u^* has to be as close as possible to the integrand so that the variance reduction method resumes to the evaluation of nearly 0. In the second case ($x \rightarrow 1 + 5x$), the area under $x \rightarrow |\exp x - 1 - 5x|$ is even larger than the area under $x \rightarrow \exp x$. And the two MC approximations use the same number of MC samples to explore both areas. In this case, the variance is increased and the estimation is worse, see table 9.1.

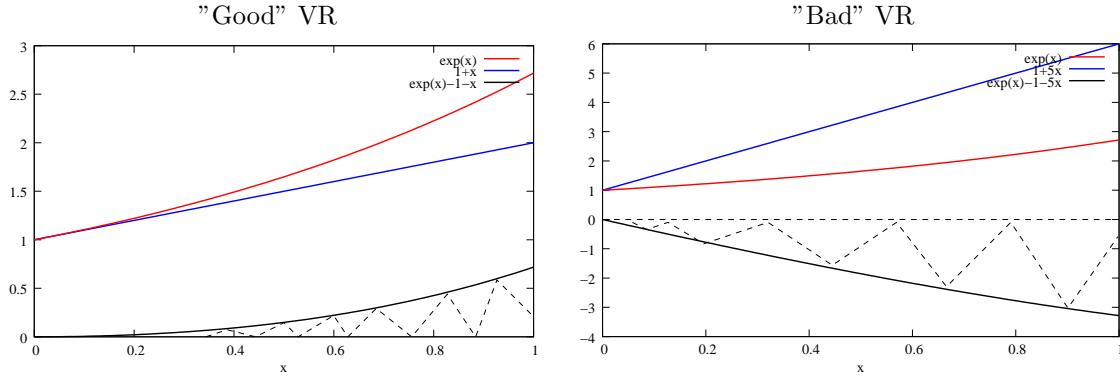


Figure 9.5: On the left, $x \rightarrow \exp x$ (red), $x \rightarrow 1+x$ (blue, good reduced model) and $x \rightarrow \exp x - 1 - x$ (black) are plotted. On the right, $x \rightarrow \exp x$ (red), $x \rightarrow 1 + 5x$ (blue, bad reduced model) and $x \rightarrow \exp x - 1 - 5x$ (black) are plotted. The dashed area is the surface under $x \rightarrow \exp x - 1 - x$ and $x \rightarrow \exp x - 1 - 5x$. In the second case the dashed area is more important than the surface under $x \rightarrow \exp x$ (variance is not reduced), whereas it is smaller in the first case (variance is reduced).

With this example, it is easier having an idea of why being able to find a "good" reduced model is the key step⁶⁰. Finding a good reduced model can reveal to be quite tricky, as illustrated above. One generally relies on *a priori* available information on the integrand. How to practically use it is the next question. There are several ways to introduce a reduced model to accelerate MC computation of (9.166). The most classical ones are *Importance Sampling (IS)* and *Control Variate (CV)* methods. Lots of other methods exist (Stratified Sampling, low discrepancy suites, etc. see [256, 65, 103]) but do not necessarily need the introduction of *a priori* information on the integrand *via* a reduced model u^* . For both methods, u^* has to be the closer possible to u . They only differ from the fact the IS method introduces u^* multiplicatively whereas the CV one introduces u^* additively⁶¹. We briefly go through their principles in the two next paragraphs.

The Control Variate acceleration method

The CV method supposes one has a model (reduced model) $x \rightarrow u^*(x)$ approximating u . For efficiency, few requirements are mandatory: $u^*(x)$ is assumed to be fast to evaluate $\forall x \in \mathbb{R}^Q$ so that $I_0 = \int u^*(x)d\mathcal{P}_X(x) = \mathbb{E}[u^*(X)]$ is known with a good accuracy at a relatively low cost. The reduced model u^* is then plugged in the computation of I by introducing its expression in the expectation

$$I = \mathbb{E}[u(X) - u^*(X) + u^*(X)] = \mathbb{E}[u(X) - u^*(X)] + I_0. \quad (9.167)$$

The problem consequently resumes to the estimation of the expectation of the difference $u(X) - u^*(X)$. The question now is, what does it bring? Let us introduce $(X_i)_{i \in \{1, \dots, N_{MC}\}}$ be N_{MC} i.i.d. realisations of X . Then according to the *law of large numbers* (stated in section 5.2.1), the estimator $I_{N_{MC}}^{CV}$ ensures

$$I_{N_{MC}}^{CV} = I_0 + \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} (u(X_i) - u^*(X_i)) \xrightarrow[N_{MC} \rightarrow \infty]{a.s.} \mathbb{E}[u(X)] = I. \quad (9.168)$$

It is almost surely converging and is unbiased [256]. Being unbiased ensures its variance is an error estimator. It means the constant K multiplying the convergence rate $\frac{1}{\sqrt{N_{MC}}}$ is equal to the variance of the estimator $I_{N_{MC}}^{CV}$, given by

$$K = \mathbb{V}[I_{N_{MC}}^{CV}] = \mathbb{V}[u(X) - u^*(X)]. \quad (9.169)$$

⁶⁰Note that for us, a "good" reduced model does not mean the best: only a reduced model ensuring a variance reduction (and above all not an increase of it!) with relatively low cost.

⁶¹also known as *Difference Formulation* in the literature.

Expression (9.169) consequently ensures the constant K can be reduced if u^* is sufficiently well-suited, i.e. close enough to u . Of course, if $u^* = u$, the error is zero. Note that the Control Variate method is at the basis of Multi-Level MC for uncertainty quantification, see [23, 264], which can be understood as an iterated CV approach. This latter method was not described in part II mainly because we focused on spectral methods.

The Important Sampling acceleration method

The IS method also assumes one has access to a model (reduced model) $x \rightarrow u^*(x)$ approximating u . The main difference comes from the fact u^* is introduced multiplicatively in the expectation $I = \mathbb{E}[u(X)]$ and must satisfy slightly different requirements. Suppose $u^*(X) > 0$ and $\int u^*(x)d\mathcal{P}_X(x) = 1$, then $d\mathcal{P}_{u^*}(x) = u^*(x)d\mathcal{P}_X(x)$ is a probability measure. The IS method is then based on the following change of variable

$$I = \int \frac{u(x)}{u^*(x)} u^*(x)d\mathcal{P}_X(x) = \int \frac{u(x)}{u^*(x)} d\mathcal{P}_{u^*}(x) = \mathbb{E} \left[\frac{u(Y)}{u^*(Y)} \right]. \quad (9.170)$$

In the above expression, Y follows the law defined by the probability measure $d\mathcal{P}_{u^*}(x)$. Implicitly, another condition for the IS method to ensure a gain supposes Y must be sampled easily and quickly. The cdf of the latter probability measure must be fast to inverse, see [33].

To understand how gains can be obtained, let us introduce $(Y_i)_{i \in \{1, \dots, N_{MC}\}}$, N_{MC} i.i.d. realisations of Y . Then, according to the *law of large numbers* (stated in section 5.2.1), the estimator $I_{N_{MC}}^{IS}$ ensures

$$I_{MC}^{IS} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \frac{u(Y_i)}{u^*(Y_i)} \xrightarrow[N_{MC} \rightarrow \infty]{a.s.} \mathbb{E} \left[\frac{u(Y)}{u^*(Y)} \right] = I. \quad (9.171)$$

It converges almost surely and is unbiased, see [256]. The variance of the estimator $I_{N_{MC}}^{IS}$ is given by

$$K = \mathbb{V} [I_{N_{MC}}^{IS}] = \int \left(\frac{u(y)}{u^*(y)} \right)^2 d\mathcal{P}_{u^*}(y) - I^2 = \int \frac{u^2(y)}{u^*(y)} d\mathcal{P}_X(y) - I^2. \quad (9.172)$$

Expression (9.172) ensures the constant K can be reduced if $\frac{u^*}{I}$ is close enough to u . In such case, i.e. if $\frac{u^*}{I} \approx u$, we have

$$\int \frac{u^2(y)}{u^*(y)} d\mathcal{P}_X(y) - I^2 \approx I \int \frac{u^2(y)}{u(y)} d\mathcal{P}_X(y) - I^2 = I \int u(y) d\mathcal{P}_X(y) - I^2 = 0. \quad (9.173)$$

Remark 9.4 Note that the non-analog MC scheme presented in section 9.4, AP in the homogeneous regime, can be reinterpreted as an IS variance reduction with $u^*(\mathbf{x}, t) = u_0(\mathbf{x})e^{-v\sigma_a t}$ (along a characteristic) leading to a zero variance in this particular configuration. The computations of section 9.7.3 are similar to the one performed to obtain (9.173).

Both methods presented above rely on the hypothesis of having a sufficiently well-suited reduced model u^* approximating u . Having such function u^* is not generally straightforward and we have seen in the example in the introductory paragraph that mistaking on u^* can have some dramatic consequences. In the next section 9.12.1, we briefly hint at a parallel between variance reduction techniques and Asymptotic Preserving schemes. Both are equivalent when the regime of interest is identified and used in the design of the resolution scheme. In section 9.12.2, we propose a method based on a gPC decomposition of u in order to reduce the variance in the MC estimations of an integral when *a priori information on the integrand/regime is not available or complex to identify*. In section 9.12.3, we apply the material of section 9.12.2, and use gPC developments as reduced models, to accelerate the MC resolution of the linear Boltzmann equation.

9.12.1 Variance reduction, AP scheme, same problems, different denominations

Some variance reduction techniques have already been presented in this document: the semi-analog and the non-analog schemes (sections 9.3–9.4) both reduced the variance with respect to the analog MC scheme (see section 9.7) in the homogeneous case (see section 9.7). The multiplicity option, on another hand, has been discarded as an interesting MC scheme precisely because it increased the variance. The non-analog scheme was also presented as an Asymptotic Preserving scheme for the homogeneous regime.

We here want to highlight the fact that variance reduction methods and Asymptotic Preserving schemes are closely related if not equivalent. They both aim at decreasing the constant in the $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}})$ convergence rate of the MC method. The main difference between both may come from the fact that variance reduction technique are usually used when the asymptotic regime of interest ($\delta \rightarrow 0$) is complex to identify and characterise, see [165, 181, 17]. When the stiff regime is known and identified, the introduction of the asymptotic regime in the MC computation *via* the construction of an AP scheme may be more efficient. An example for the nonlinear Boltzmann equation coupled to Bateman system is described in chapter 10.

When the stiff regime of interest is hard to determine, we suggest using the spectral property in the L^2 -norm (i.e. for the variance) of the gPC decomposition. The aim is to automatically build a relevant reduced model for the regime/configuration of interest. The methodology is first described for simple integration in the next section 9.12.2. It is then applied to the acceleration of the MC resolution of the linear Boltzmann equation in an intensively studied configuration [165, 181, 17] in section 9.12.3.

9.12.2 Application of gPC to accelerate MC integration

We begin by restating convergence theorem 3.2 with slightly different notations. In its essence, the theorem remains unchanged. This new statement only aims at easing the introduction of gPC for variance reduction.

Corollary 9.1 *Convergence of generalized Polynomial Chaos in new notations: let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let X be an arbitrary random variable of given probability measure $d\mathcal{P}_X$. Let $(\phi_k^X)_{k \in \mathbb{N}}$ be the basis of orthonormal polynomials with respect to $d\mathcal{P}_X$. Let $u(X)$ be an unknown random variable. Suppose that $\mathbb{V}[u(X)] < \infty$. Introduce the polynomial development $u_P^X(X) = \sum_{k=0}^P u_k \phi_k^X(X)$ where the polynomial coefficients $(u_k)_{k \in \mathbb{N}}$ are defined as $u_k = \int u(X) \phi_k^X(X) d\mathcal{P}_X$, the projection of the solution on above polynomial basis associated to $d\mathcal{P}_X$. Then we have*

$$\mathbb{V}[u(X) - u_P^X(X)] \xrightarrow{P \rightarrow \infty} 0. \quad (9.174)$$

Expression (9.174) echoes (9.169), required by u^* to be a relevant ingredient of any CV method. It even ensures the construction of a reduced model up to an arbitrary accuracy. In the next paragraphs, we consequently naturally choose u^* (CV) or $\frac{u^*}{T}$ (IS) as a P -truncated gPC development.

We insist on the fact that *this is not the first time* orthogonal polynomials are used in VR techniques, see⁶² [184, 65, 103, 135, 38, 188, 66, 103]: the aim of this section is to present the properties of gPC in the context of numerical integration (i.e. closely related to initial and source samplings of sections 9.8.1 and 9.9.1) and of the resolution of the transport equation.

Back to the toy problem (9.165) with a gPC_P reduced model for a CV method

Let us illustrate what can be expected from the use of a gPC development as a reduced model on example (9.165) with a CV method: we aim at evaluating

$$I = \int_0^1 \exp x dx = \mathbb{E}[\exp(X)]. \quad (9.175)$$

⁶²The list is not exhaustive but has the particularity of ranging from 1964 to 2003.

Recall $X \sim \mathcal{U}([0, 1])$. The three first components of the gPC basis, here the Legendre one, cf. table 3.1, associated to X are given by

$$\begin{cases} \phi_0(x) = 1, \\ \phi_1(x) = \sqrt{3}(2x - 1), \\ \phi_2(x) = \sqrt{5}(6x^2 - 6x + 1). \end{cases} \quad (9.176)$$

The coefficients of the gPC development in this basis are given by (analytical calculations here)

$$\begin{cases} \exp_0 = -1 + e, \\ \exp_1 = \sqrt{3}(3 - e), \\ \exp_2 = \sqrt{5}(-19 + 7e). \end{cases} \quad (9.177)$$

Figure 9.6 (right) shows the approximation of $x \rightarrow \exp x$ with the first and second order gPC, i.e.

$$\begin{aligned} \exp^1(x) &= \exp_0 \phi_0(x) + \exp_1 \phi_1(x), \\ \exp^2(x) &= \exp_0 \phi_0(x) + \exp_1 \phi_1(x) + \exp_2 \phi_2(x). \end{aligned} \quad (9.178)$$

Note that the second order gPC expansion (black dotted curve) matches the analytical curve (red) on the whole integration domain $[0, 1]$. On the other hand, figure 9.6 also compares the 1st and 2nd Taylor developments⁶³ of $x \rightarrow \exp x$ (right), given by

$$\begin{aligned} \exp_{Taylor}^1(x) &= 1 + x, \\ \exp_{Taylor}^2(x) &= 1 + x + \frac{x^2}{2}. \end{aligned} \quad (9.179)$$

As expected, Taylor expansions are interesting in the context of VR methods but they only match the function locally (vicinity of 0 here). Table 9.2 sums up the results obtained with the classical MC

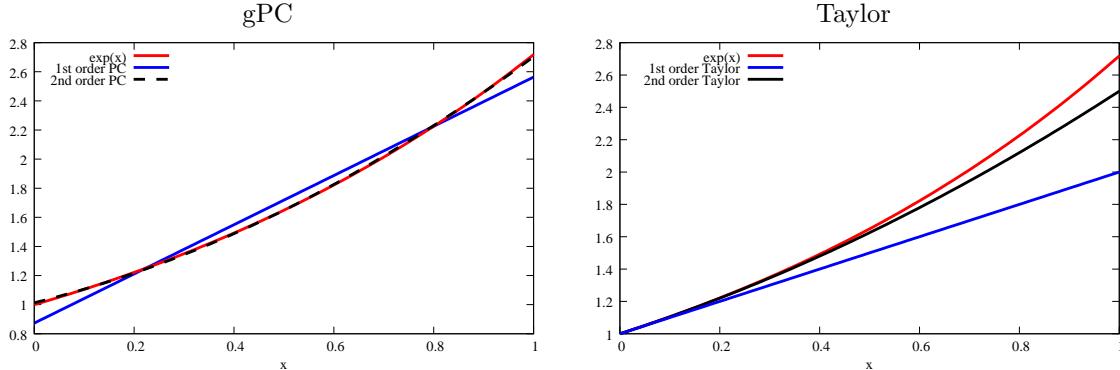


Figure 9.6: On the left, $x \rightarrow \exp x$ (red), $x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x)$ (blue) and $x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x) + \exp_2 \phi_2(x)$ (black) are plotted. On the right, $x \rightarrow \exp x$ (red), $x \rightarrow 1 + x$ (blue) and $x \rightarrow 1 + x + \frac{1}{2}x^2$ (black) are plotted

method and the CV method for four different reduced models:

- 1st and 2nd order Taylor expansion of $x \rightarrow \exp x$ in the vicinity of 0, see (9.179),
- and the 1st and 2nd order gPC expansions of the same function, see (9.178).

The Taylor-CV method reduces variance considerably especially with the 2nd order approximation. The gPC-CV method reaches the same accuracy as the 2nd order Taylor-CV method with order 1. The 2nd order gPC-CV method increases the gain of a factor $96.75/8.18 = 11.82$ in comparison to the 1st order gPC.

These results are encouraging but needs to be interpreted carefully: in all those calculations, we supposed the gPC coefficients were known exactly. This may not be true in real computations as they

⁶³Corresponding to the "good" reduced models of section 9.12.2.

$I = 1.7183$	Results	$\frac{\text{Std}}{\sqrt{n}}$	Gain
Classical MC ($n = 3000$)	1.7151	9.0×10^{-3}	reference
1 st order Taylor-CV ($n = 3000$)	1.7169	3.8×10^{-3}	2.36
2 nd order Taylor-CV ($n = 3000$)	1.7179	1.1×10^{-3}	8.18
1 st order gPC CV ($n = 3000$)	1.7181	1.1×10^{-3}	8.18
2 nd order gPC CV ($n = 3000$)	1.7184	9.6×10^{-5}	96.75

Table 9.2: The table compares the Classical MC method to a 1st ($x \rightarrow 1+x$) and 2nd ($x \rightarrow 1+x+\frac{1}{2}x^2$) order Taylor-CV method and to a 1st ($x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x)$) and 2nd ($x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x) + \exp_2 \phi_2(x)$) order gPC-CV method. The gPC coefficients are given by (9.177) and the basis by (9.176). The gain (last column) is the ratio of the std obtain by the Classical MC method and the std obtained with the considered VR method.

have to be estimated. Nevertheless, it gives a good idea of what can be asymptotically achieved with such reduced model. In the following sections, we suggest several ways to estimate the gPC coefficients.

The last section tends to show that provided an accurate estimation of the coefficients of the gPC development of the integrand $u(X)$, gPC combined to VR methods (CV or IS) can lead to a considerable and automatic increase in the accuracy of the integral evaluation. We here suggest simple ways for estimating the gPC coefficients. Let us come back to our general problem of integrating

$$I = \int u(x)d\mathcal{P}_X(x) = \mathbb{E}[u(X)]. \quad (9.180)$$

Recall u is unknown in the sense it can only be evaluated. Let us introduce $(\phi_k^X)_{k \in \mathbb{N}}$ the gPC basis associated to X ⁶⁴. We want to estimate the gPC expansions of $u(X)$ on this basis, i.e. we are looking for the coefficients

$$u_k = \int u(x)\phi_k^X(x)d\mathcal{P}_X(x), \forall k \in \{0, \dots, P\}.$$

Several solutions are possible, we suggest to develop three of them.

MC method for the computation of the $(u_k)_{k \in \{0, \dots, P\}}$

A first possibility is to compute the $(u_k)_{k \in \{0, \dots, P\}}$ with an MC method. By definition, we have

$$u_k = \int u(x)\phi_k^X(x)d\mathcal{P}_X(x) = \mathbb{E}[u(X)\phi_k^X(X)], \forall k \in \{0, \dots, P\},$$

where X is a random vector of dimension Q of probability measure $d\mathcal{P}_X$. Let us consider n_{MC} i.i.d. realisations of X denoted by $(x_i)_{i \in \{1, \dots, n_{MC}\}}$. according to theorem 5.1,

$$u_k^{n_{MC}} = \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} u(x_i)\phi_k^X(x_i), \quad (9.181)$$

is a convergent unbiased estimator of u_k , $\forall k \in \{0, \dots, P\}$. Now introduce N_{MC} i.i.d. realisations of X denoted by $(X_i)_{i \in \{1, \dots, N_{MC}\}}$. Then the following estimator

$$I_{n_{MC}, N_{MC}} = u_0^{n_{MC}} + \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \left(u(X_i) - \sum_{k=0}^P u_k^{n_{MC}} \phi_k(X_i) \right) \xrightarrow[N_{MC} \rightarrow \infty]{a.s.} \mathbb{E}[u(X)] = I, \quad (9.182)$$

⁶⁴cf. section 3.4.

converges $\forall n_{MC}$ (according to theorem 5.1) and is unbiased. Indeed, we have

$$\begin{aligned} I_{n_{MC}, N_{MC}} &= \frac{1}{n_{MC}} \sum_{j=1}^{n_{MC}} u(x_j) + \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} u(X_i) - \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \left(\sum_{k=0}^P \frac{1}{n_{MC}} \left(\sum_{j=1}^{n_{MC}} u(x_j) \phi_k(x_j) \right) \phi_k(X_i) \right), \\ &= \frac{1}{n_{MC}} \sum_{j=1}^{n_{MC}} u(x_j) + \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} u(X_i) - \left(\sum_{k=0}^P \frac{1}{n_{MC}} \left(\sum_{j=1}^{n_{MC}} u(x_j) \phi_k(x_j) \right) \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \phi_k(X_i) \right). \end{aligned} \quad (9.183)$$

If we now take the expectation of $I_{n_{MC}, N_{MC}}$, we get

$$\begin{aligned} \mathbb{E}[I_{n_{MC}, N_{MC}}] &= \underbrace{\frac{1}{n_{MC}} n_{MC} \mathbb{E}[u(X)] + \frac{1}{N_{MC}} N_{MC} \mathbb{E}[u(X)]}_{\text{as the } x_j \text{ and } X_i \text{ s are i.i.d.}} \\ &\quad - \left(\sum_{k=0}^P \frac{1}{n_{MC}} \underbrace{\mathbb{E} \left[\sum_{j=1}^{n_{MC}} u(x_j) \phi_k(x_j) \right]}_{\text{as } x_j \text{ and } X_i \text{ are independent}} \underbrace{\frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \mathbb{E}[\phi_k(X_i)]}_{\text{ }} \right), \\ \mathbb{E}[I_{n_{MC}, N_{MC}}] &= \mathbb{E}[u(X)] + \mathbb{E}[u(X)] \\ &\quad - \left(\sum_{k=0}^P \underbrace{\mathbb{E}[u(X) \phi_k(X)]}_{\text{as } x_j \text{ are i.i.d.}} \underbrace{\frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \mathbb{E}[\phi_k(X_i)]}_{=\delta_{k,0}(\text{orthonormality})} \right), \\ &= \mathbb{E}[u(X)]. \end{aligned} \quad (9.184)$$

The last line is characteristic of an unbiased estimator, see [256].

The main drawback of this first possibility is that the VR techniques now implies $n_{MC} + N_{MC}$ evaluations of the unknown function u rather than N_{MC} . Besides, n_{MC} may need to be important for efficiency. The next method suggests estimating the coefficients $(u_k)_{k \in \{0, \dots, P\}}$ without the use of an MC method.

Quadrature rules for the computation of the $(u_k)_{k \in \{0, \dots, P\}}$

A second possibility developed in [103, 38], which is also often used in uncertainty quantification, see part II, consists in the use of quadrature rules (see [265, 180, 34]) for the evaluation of the coefficients $(u_k)_{k \in \{0, \dots, P\}}$. Let us introduce the deterministic points $(x_l, w_l)_{l \in \{1, \dots, n_q\}}$, where $(x_l)_{l \in \{1, \dots, n_q\}}$ are the points and $(w_l)_{l \in \{1, \dots, n_q\}}$ are their associated weights. Both are deterministic and ensure a discretisation of $(X, d\mathcal{P}_X)$. In other words, we have $\forall k \in \{0, \dots, P\}$

$$u_k \approx u_k^{n_q} = \sum_{l=1}^{n_q} w_l u(x_l) \phi_k^X(x_l). \quad (9.185)$$

The following estimator

$$I_{n_q, N_{MC}} = u_0^{n_q} + \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \left(u(X_i) - \sum_{k=0}^P u_k^{n_q} \phi_k^X(X_i) \right) \xrightarrow[N_{MC} \rightarrow \infty]{a.s.} \mathbb{E}[u(X)] = I, \quad (9.186)$$

converges $\forall n_q$ (according to theorem 5.1) and is unbiased (as the gPC coefficients are deterministically evaluated).

Once again, the gPC acceleration of the CV method needs $n_q + N_{MC}$ evaluations of the unknown function u but in general $n_q \ll N_{MC}$. Of course, with such strategy, the quality of the reduced model depends on the smoothness of the integrand and of the dimensionality of X , see section 5.2.3. In the

next section, we suggest a last way of estimating the same coefficients implying only N_{MC} evaluations of the unknown function u , exactly as with the Classical MC approach.

A biased MC estimator for the computation of the $(u_k)_{k \in \{0, \dots, P\}}$

We here detail the implications of having a biased estimator for evaluating the gPC coefficients $(u_k)_{k \in \{0, \dots, P\}}$. Introduce N_{MC} i.i.d. MC points $(X_i)_{i \in \{1, \dots, N_{MC}\}}$. According to theorem 5.1,

$$u_k^{N_{MC}} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} u(X_i) \phi_k^X(X_i), \quad (9.187)$$

is a convergent unbiased estimator of u_k , $\forall k \in \{0, \dots, P\}$. We here suggest to reuse the evaluations $(u(X_i))_{i \in \{1, \dots, N_{MC}\}}$ of u at the previous MC points. The approach does not need more sampling or evaluations than the N_{MC} ones for the $(u_k)_{k \in \{0, \dots, P\}}$. Integral I is approximated by the estimator

$$\bar{I}_{N_{MC}} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \left[u(X_i) - \sum_{k=0}^P u_k^{N_{MC}} \phi_k^X(X_i) \right]. \quad (9.188)$$

Table 9.3 presents the results on the same toy problem (9.165) with this low cost gPC acceleration (9.188). The variance has been reduced if we compare the low cost gPC acceleration's results to the ones

$I = 1.7183$	Results	$\frac{\text{Std}}{\sqrt{n}}$	Gain
Classical MC ($n = 3000$)	1.7079	9.1×10^{-3}	reference
1 st order Taylor-CV ($n = 3000$)	1.7145	3.9×10^{-3}	2.33
2 nd order Taylor-CV ($n = 3000$)	1.7188	1.1×10^{-3}	8.27
1 st order gPC CV ($n = 3000$)	1.7194	1.2×10^{-3}	7.58
2 nd order gPC CV ($n = 3000$)	1.7175	7.9×10^{-4}	11.51

Table 9.3: The table compares the Classical MC method to a 1st ($x \rightarrow 1+x$) and 2nd ($x \rightarrow 1+x+\frac{1}{2}x^2$) order Taylor-CV method and to a 1st ($x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x)$) and 2nd ($x \rightarrow \exp_0 \phi_0(x) + \exp_1 \phi_1(x) + \exp_2 \phi_2(x)$) order gPC-CV method. The conditions are the same as in table 9.2 except from the fact that the gPC coefficients are now estimated during the VR computation through the procedure described in this section. The gain (last column) is the ratio of the std obtained by the Classical MC method and the std obtained with the considered VR method.

from the Classical MC approach. Of course, the gain with respect to the Taylor based VR method is less important than the one obtained in table 9.2 as the gPC coefficients are now estimated (rather than known exactly).

We insist this method does not ensure estimator $\bar{I}_{N_{MC}}$ defined in (9.188) is unbaised as the coefficients

$(u_k^{N_{MC}})_{k \in \{0, \dots, P\}}$ depends on $(X_i)_{i \in \{1, \dots, N_{MC}\}}$. The bias of the estimator is indeed given by

$$\begin{aligned}
\beta[\bar{I}_{N_{MC}}] &= \mathbb{E}[\bar{I}_{N_{MC}} - u(X)], \\
\beta[\bar{I}_{N_{MC}}] &= \mathbb{E}[u(X)] - \frac{1}{N_{MC}^2} \sum_{k=0}^P \sum_{l,i=1}^{N_{MC}} \mathbb{E} [u(X_l) \phi_k^X(X_l) \phi_k^X(X_i)], \\
\beta[\bar{I}_{N_{MC}}] &= \mathbb{E}[u(X)] - \frac{1}{N_{MC}^2} \sum_{k=0}^P \sum_{l=1}^{N_{MC}} \mathbb{E} [u(X_l) \phi_k^X(X_l) \phi_k^X(X_l)] \\
&\quad - \frac{1}{N_{MC}^2} \sum_{k=0}^P \sum_{\substack{l,i=0 \\ i \neq l}}^{N_{MC}} \underbrace{\mathbb{E} [u(X_l) \phi_k^X(X_l) \phi_k^X(X_i)]}_{=\mathbb{E}[u(X_l) \phi_k^X(X_l)] \underbrace{\mathbb{E} [\phi_k^X(X_i)]}_{=\delta_{k,0}}} , \\
\beta[\bar{I}_{N_{MC}}] &= \mathbb{E}[u(X)] - \frac{1}{N_{MC}} \mathbb{E}[u(X)] - \frac{1}{N_{MC}} \sum_{k=1}^P \mathbb{E} [u(X) \phi_k^2(X)] \\
&\quad - (1 - \frac{1}{N_{MC}}) \mathbb{E}[u(X)], \\
\beta[\bar{I}_{N_{MC}}] &= -\frac{1}{N_{MC}} \sum_{k=1}^P \mathbb{E} [u(X) \phi_k^2(X)] = \mathcal{O}\left(\frac{1}{N_{MC}}\right).
\end{aligned} \tag{9.189}$$

The bias tends to zero as N_{MC} tends to infinity ($\forall P \in \mathbb{N}$). The estimator is said *consistent* but not unbiased, see [256]. Nevertheless, the bias $\beta[\bar{I}_{N_{MC}}]$ is $\mathcal{O}(\frac{1}{N_{MC}})$ so that for important N_{MC} , $\beta[\bar{I}_{N_{MC}}]$ is negligible in comparison to the MC error. Indeed, we have $\mathcal{O}(\frac{1}{\sqrt{N_{MC}}}) \underset{N_{MC} \gg 1}{\gg} \mathcal{O}(\frac{1}{N_{MC}})$. This will be confirmed numerically in the following examples. Practically, having a *biased* (but consistent) MC estimator implies the estimator converges but the variance is not anymore an error estimator, see [256].

Application of the gPC_P reduced model for variance reduction on few test-problems

We suggest to test gPC reduced models for MC accelerations on several test-functions. We integrate them on $[0, 1]$ with respect to the Lebesgue measure and on \mathbb{R} with respect to the gaussian measure:

$$\begin{aligned}
u^1(x) &= \exp x, \\
u^2(x) &= \cos x, \\
u^3(x) &= \mathbf{1}_{[\frac{1}{2}, \infty[}(x), \\
u^4(x) &= \exp(-x^2), \\
u^5(x) &= \exp(5x), \\
u^6(x) &= \mathbf{1}_{[0.4, 0.6]}(x).
\end{aligned} \tag{9.190}$$

These functions are emphasizing different aspects and difficulties for integration with MC methods (also relative to resolution of transport equations) and VR methods.

Remark 9.5 (Error estimations) As tackled in the previous paragraph, in the case of unbiased estimators, the standard deviation is an error estimator and can be evaluated a posteriori. Otherwise, it is not, see [256]. In the following calculations, when the estimator is unbiased, the error is computed by estimating the standard deviation of the MC results. When it is not, i.e. when estimator (9.188) is applied, it is evaluated by computing the L^2 -norm of the error with respect to the analytical results (no a posteriori error information is available or relevant/reliable in practice).

In the following examples, we apply the gPC formalism for variance reduction. When the integration is carried on $[0, 1]$ with respect to the Lebesgue measure, the normalized Legendre polynomials are chosen as the gPC basis⁶⁵. When the integration is carried on \mathbb{R} with respect to the gaussian measure, the normalized Hermite polynomials are chosen as the gPC basis⁶⁶. This is in agreement with the Askey

⁶⁵Orthonormal with respect to the measure $x \mapsto \mathbf{1}_{[0,1]}$.

⁶⁶Orthonormal with respect to the measure $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

scheme, see table 3.1 and section 3.3. Note that more elaborated choices could have been done concerning the gPC basis, cf. i-gPC in chapter 6. This will not be investigated in this section but has been tested and gives quite satisfactory results.

Figure 9.7 compares the logarithm of the variance ($\ln(\frac{std}{\sqrt{N_{MC}}})$) with respect to the number of samples (N_{MC}) for the evaluation of

$$\int_{[0,1]} u^k(x) dx, \text{ for } k \in \{1, \dots, 6\}, \quad (9.191)$$

with

- the Classical MC method (blue curve),
- a CV method with reduced model given by (9.192),
- and the gPC accelerated CV method we propose in this section.

The reduced models used in order to accelerate the convergence of (9.191) are given by (9.192). They are Taylor development of the integrand $(u^k)_{k \in \{1, \dots, 6\}}$ except for the indicatrix functions u^3 and u^6 . They are given by

$$\begin{aligned} u_{Taylor}^1(x) &= 1 + x, \\ u_{Taylor}^2(x) &= 1 - \frac{x^2}{2}, \\ u_{Guess}^3(x) &= 1 + x, \\ u_{Taylor}^4(x) &= 1 - x^2, \\ u_{Taylor}^5(x) &= 1 + 5x, \\ u_{Guess}^6(x) &= 1 + x. \end{aligned} \quad (9.192)$$

For the gPC accelerated CV method, we used the low cost (biased estimator (9.188)) for estimating the gPC coefficients $(u^k)_{k \in \{0, \dots, P\}}$.

Figure 9.7 shows that for every function, every gPC acceleration reduces the variance without more estimations of the integrand than the Classical MC approach. For figure 9.7 (u^2), the 1st order gPC acceleration is less efficient than the Taylor-CV method. This is due to the fact that the cosine function is even. Higher truncation orders give satisfactory results. The $2P + 1$ orders are superposed with the $2P$ order ones. For figure 9.7 (u^3), 1st and 2nd orders are superposed as well as 3rd and 4th orders. Every orders are more efficient than the Taylor-CV method. The gPC acceleration shows great improvements for u^4 and u^5 with respect to the Taylor-CV method. Note that for u^5 , the Taylor-CV method gives results close to the Classical MC approach. For u^6 , the improvements of the gPC accelerations are less important but still reduce the variance (note that once again, the Taylor-CV method gives the same results as the Classical MC approach).

Figure 9.8 compares the logarithm of the variance ($\ln(\frac{std}{\sqrt{N_{MC}}})$) with respect to the number of samples (N_{MC}) for the evaluation of

$$\int_{\mathbb{R}} u^k(x) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \text{ for } k \in \{1, \dots, 6\}, \quad (9.193)$$

with

- the Classical MC method (blue curve),
- a CV method with reduced model given by (9.192),
- and the gPC accelerated CV method we propose in section 9.12.2.

Once again, for the CV method the reduced model of (9.190) are Taylor developments of the integrand given by (9.192). For the gPC accelerated CV method, we used the low cost biased estimator (9.188).

For functions u^1, u^4 of figure 9.8, the results are as expected: the gPC acceleration always gives better results than the Taylor-CV method. For u^4 the Taylor-CV method gives the same results as the

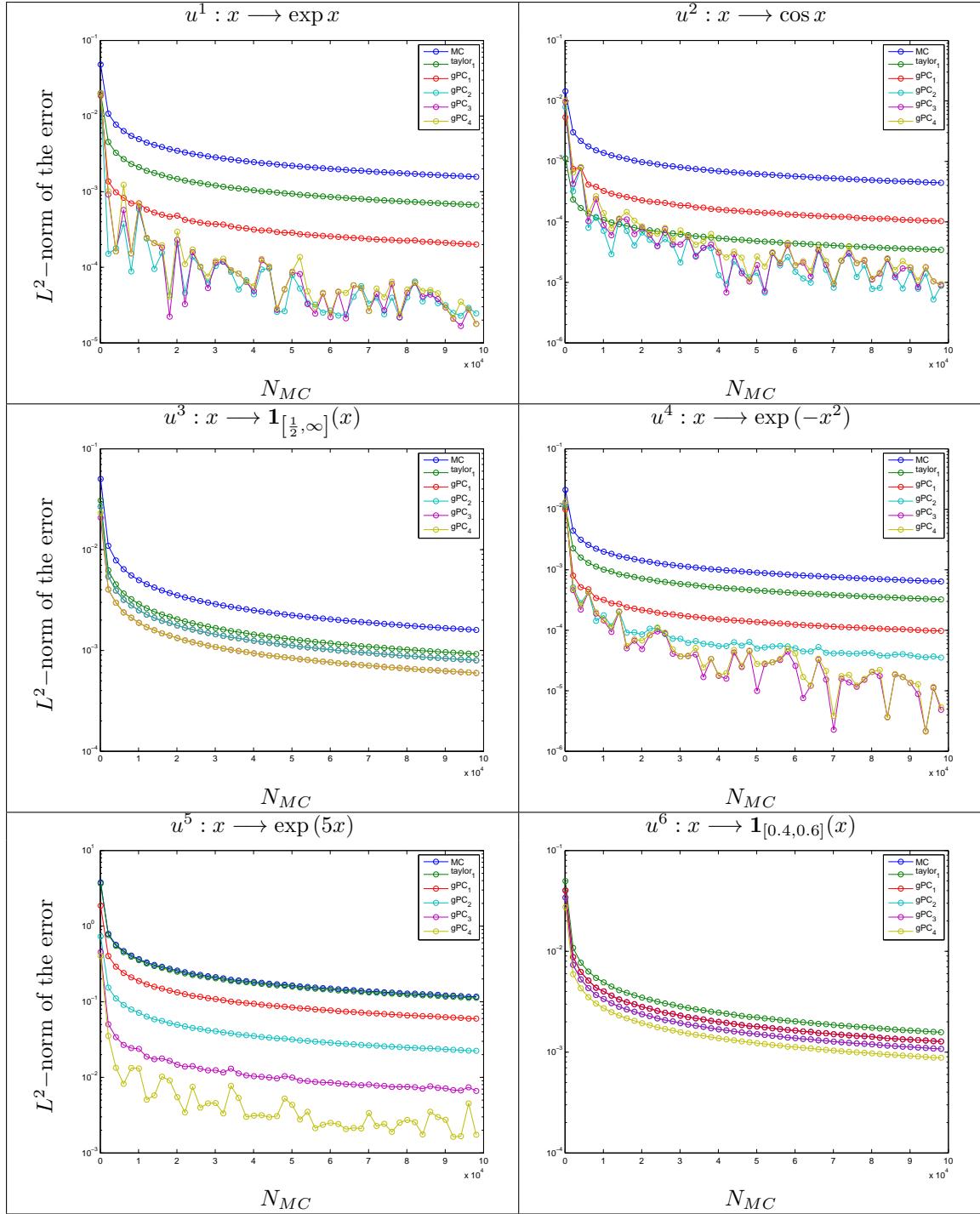


Figure 9.7: Convergence tests for the MC, Taylor-CV and gPC accelerated CV methods: integration over $[0, 1]$ with respect to the Lebesgue measure. The figure shows the logarithm of the estimation of the std with respect to the number of samples. Note that the gPC accelerations always reduce the variance of the MC method without *a priori* knowledge on the integrand.

Classical MC method. For function u^2 , the 1st order gPC accelerated CV method gives the same results as the Classical MC method. This is once again due to the fact that the cosine is even ($u_1^{N_{MC}} \approx 0$). Note that this test (u^2) emphasizes the fact that the gPC acceleration do not give worse results than the Classical MC method. This is not the case for the Taylor-CV method, as illustrated on figure 9.8 (u^3) where the Taylor curve is above the MC curve. Function u^5 is particular, cf. [165]: indeed, the exact

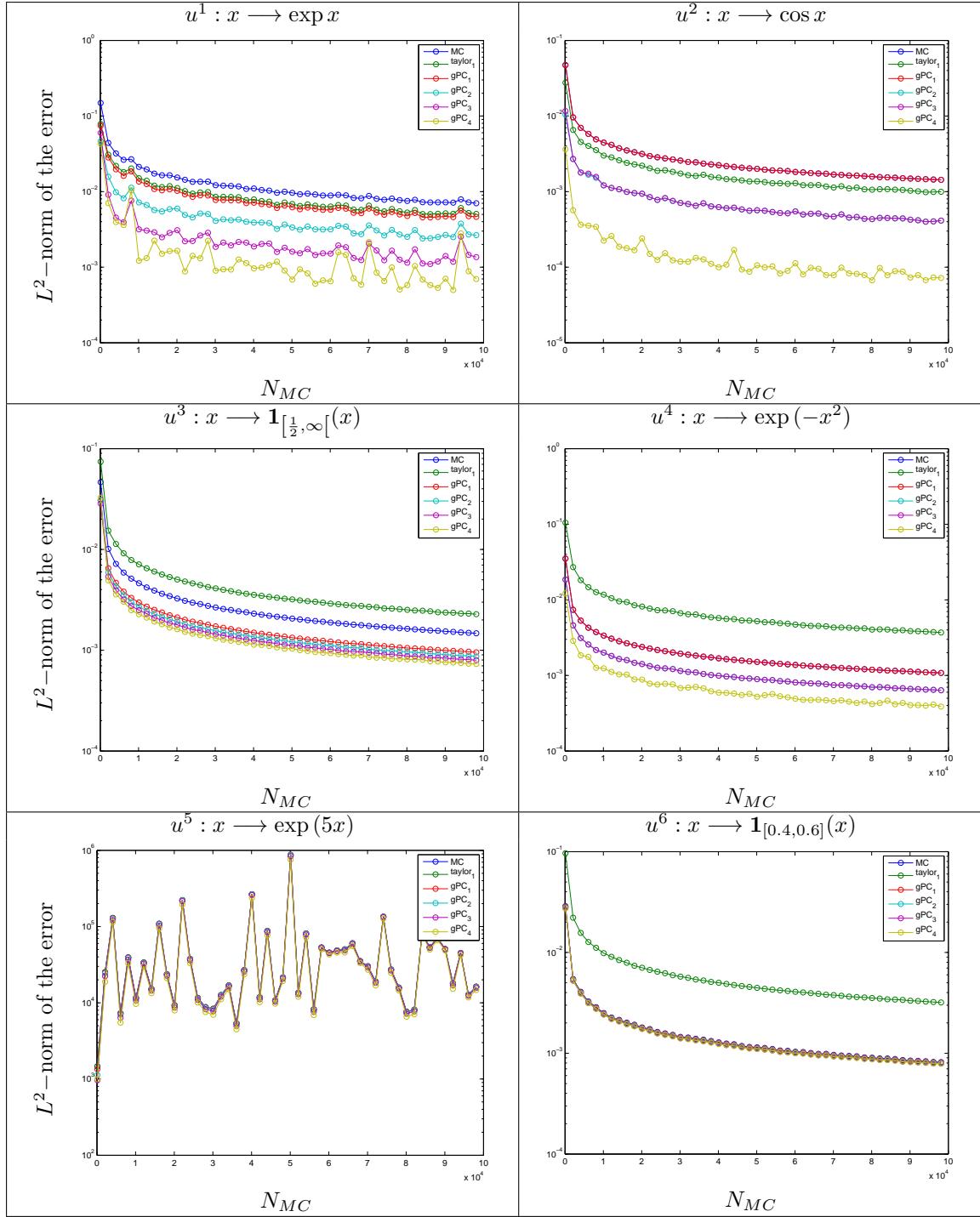


Figure 9.8: Convergence tests for the MC, Taylor-CV and gPC accelerated CV methods: integration over \mathbb{R} with respect to the gaussian measure. The figure shows the logarithm of the estimation of the std with respect to the number of samples.

variance of the estimator for this function $\exp(5X)$ where X is a normalized centered gaussian random variable can be evaluated and is given by $e^{50} - e^{25} \approx 5.1847 \cdot 10^{21}$. The variance is very important and even once reduced, it is still very high, see [165]. On this difficult case, every methods show the same poor behaviour due to the intrinsic important variance of the solution. Function u^6 is also particularly difficult to integrate. It is a thin heavyside function integrated on \mathbb{R} . Only few realisations of the RV X

contribute to the estimation of the expectation. Once again, the gPC acceleration gives the same results as the Classical MC approach whereas the Taylor-CV method gives worse results (loss of more than a decade in accuracy!). Note that the computational time for the gPC-CV method is not displayed in this section as it is the same as the one for the Classical MC approach: indeed, reusing the former MC points (cf. section 9.12.2) only implies few more operations which are not time consuming with respect to the sampling.

Until now, we have mainly considered gPC acceleration of the CV method. The gPC acceleration is also compatible with the IS one. Figure 9.9 revisits the same problems as previously but considering a gPC acceleration of the IS method. For this method, the development is only carried up to order 1 mainly because the IS method implies an inversion of the cdf of the evaluated gPC expansion⁶⁷ which can reveal to be quite tricky for important P . Alternative methods exist in order to sample random variables defined through normalized gPC expansions, from histograms, from approximated discrete pdfs or by rejection sampling [173] for example. They aim at avoiding an inversion. Evaluating these methods with gPC acceleration is beyond the scope of this section.

		Accuracy	time $\times 10^{-3}$	ratio of times
u^1	Classical MC ($N_{MC} = 220000$)	10^{-3}	9.089 s.	1 (reference)
	gPC_1 ($N_{MC} = 4000$)	10^{-3}	0.840 s.	10.82
u^2	Classical MC ($N_{MC} = 20000$)	10^{-3}	1.203 s.	1 (reference)
	gPC_1 ($N_{MC} = 1200$)	10^{-3}	0.418 s.	2.87
u^3	Classical MC ($N_{MC} = 100000$)	$1.6 \cdot 10^{-3}$	$6.70 \cdot 10^3$ s.	1 (reference)
	gPC_1 ($N_{MC} = 21000$)	$1.6 \cdot 10^{-3}$	$0.609 \cdot 10^3$ s.	11.00
u^4	Classical MC ($N_{MC} = 40000$)	10^{-3}	2.245 s.	1 (reference)
	gPC_1 ($N_{MC} = 1000$)	10^{-3}	0.389 s.	5.77
u^5	Classical MC ($N_{MC} = 15000000$)	10^{-2}	$0.763 \cdot 10^3$ s.	1 (reference)
	gPC_1 ($N_{MC} = 1500000$)	10^{-2}	$0.130 \cdot 10^3$ s.	5.86

Table 9.4: Comparison of computational times times to attain the same accuracies with the Classical MC method and the gPC accelerated IS for the $(u^k)_{k \in \{1, \dots, 5\}}$ problems (9.190).

Figure 9.9 compares the Classical MC method to the Taylor IS and the gPC accelerated IS method with integration carried on $[0, 1]$ for $(u^k)_{k \in \{1, \dots, 6\}}$. The 1st order gPC accelerated IS gives better results than the Classical MC and the Taylor IS on $(u^k)_{k \in \{1, \dots, 5\}}$. Even for u^2 , for which the Taylor IS is second order (see (9.192)). The u^6 case is particularly difficult. It consists of an indicatrix with a thin support. For this test-case, every methods are equivalent. Table 9.4 presents quantitative results on the gains in computational time for these last problems. The gPC acceleration enables gains from ≈ 2 to 11 with only first orders expansions.

Remark 9.6 (Possible Optimizations) Note that for the later examples, the same number of point is always used in the two steps of the algorithm. Very simple and obvious optimizations can increase the efficiency of the approach:

- [(i)] indeed, to estimate the gPC coefficients, we used $N_{(i)} = N_{MC}$ points,
- [(ii)] and we used the same number of points to evaluate estimator (9.188) thanks to the reduced model of step (i), $N_{(ii)} = N_{MC} = N_{(i)}$ points .

The gPC-IS accelerated method can also be used with two different numbers of points in steps (i) and (ii), i.e. with $N_{(i)} \neq N_{(ii)}$, to increase the gain (CPU time vs. accuracy). This is emphasized in the problem of section 9.12.3, see results of table 9.9.

⁶⁷Inversion of the pdf of $u_P^X(x) = \sum_{k=0}^P u_k \phi_k^X(x)$, implying the computation of the roots of the polynomial cf. [287].

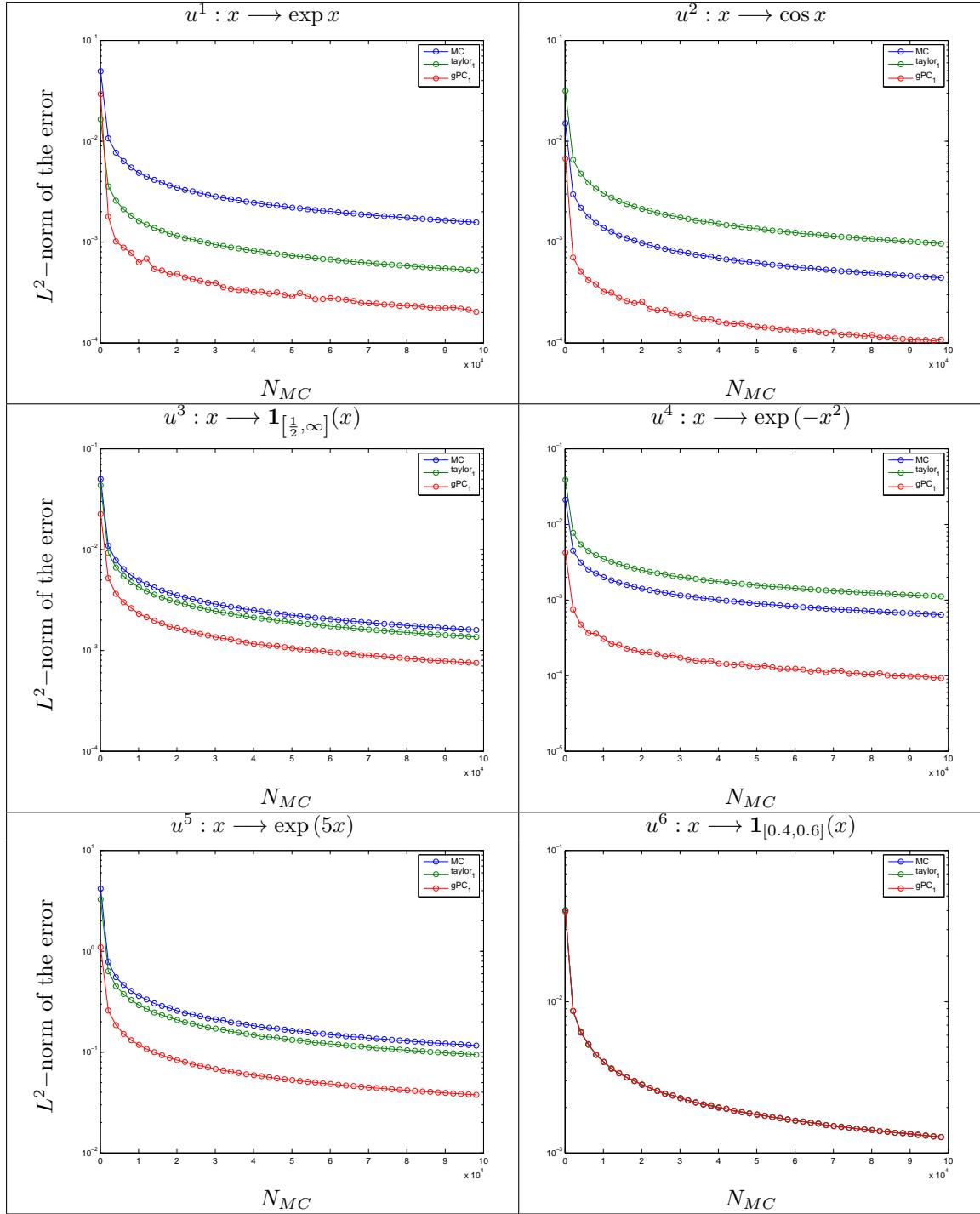


Figure 9.9: Convergence tests for the MC, Taylor-IS and gPC accelerated IS methods: integration over $[0, 1]$ with respect to the Lebesgue measure. The figure shows the logarithm of the estimation of the std with respect to the number of samples.

Application of the gPC_P reduced model for variance reduction on multi-dimensional test-problems

From now on, we have only dealt with one dimensional integration problems. Here, we illustrate the gPC acceleration on some two dimensional test-cases, presented in figure 9.10. These problems are relevant for several reasons:

- The treatment of higher stochastic dimension is important as the gPC expansions are subject to the "Curse of dimensionality", see chapter 3 and [213]. With dimension and truncation order, the number of gPC coefficient to evaluate grows exponentially fast⁶⁸.
- These problems are often encountered for computation of presence fraction/mass fraction in the context of material interfaces crossing cells (Finite Volume methods), see chapter 7.
- The considered problems are more and more anisotropic leading to difficulties for classical VR technics when the direction of interest is not known *a priori*.

Problems of figure 9.10 consists of computing thanks to MC methods (hit and miss MC or rejection sampling) some surfaces within a cell of size $[0, 1] \times [0, 1]$. We suggest to evaluate the surface of a circle

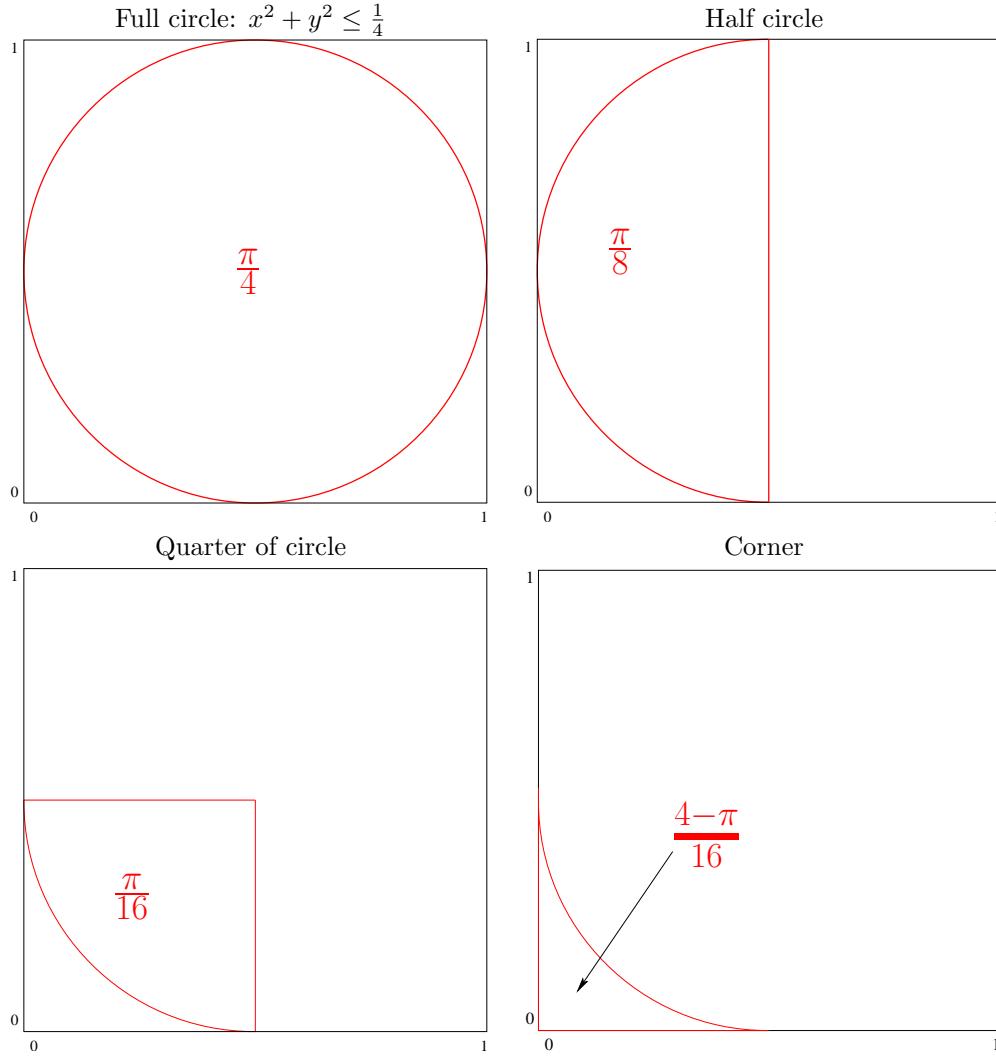


Figure 9.10: Presentation of the 2-D test-cases: it consists of evaluating the (red) surfaces: a full circle of radius 0.5 ($\frac{\pi}{4}$), half a circle ($\frac{\pi}{8}$), a quarter of circle ($\frac{\pi}{16}$) and the outside corner of a full circle ($\frac{4-\pi}{16}$) of the same radius. The first problem has uniform directions and every other problems are direction dependent: the gPC-VR (both CV and IS) automatically build a reduced model detecting the direction of interest (see figures 9.11 and 9.12) and reduces variance.

of radius $\frac{1}{2}$ (isotropic problem), the surface of half this circle, the surface of a quarter of this circle and

⁶⁸The number of coefficient is $M = \frac{D+P!}{D!P!}$ where P is the truncation order in every stochastic dimensions, D denote the dimension.

the surface of one corner⁶⁹ (the three last corresponding to anisotropic problems).

The results comparing the Classical MC, a Taylor VR and some gPC accelerated VR for problems of figure 9.10 are given in figures 9.11 (CV method) and 9.12 (IS method). For the gPC accelerated CV method, the 2-dimensional gPC basis is built by tensorization of the normalized one dimensional Legendre basis. In the context of the gPC accelerated IS method, the 2-dimensional basis is 1st order, as in section 9.12.2, and separable (no correlation terms⁷⁰). Once again, this simplifies the inversion of the cdf in the IS procedure. Let us now describe the results of figures 9.11 and 9.12. The full circle problem is isotropic: this explains why the odd gPC orders are not contributing to an increase of the accuracy. The gPC₁ reduced model gives the same results as the simple MC approach on figures 9.11 and 9.12. This is due to the fact the first order gPC coefficients are zero. Note that the results are not better but also not worse. In this case (gPC₁), the gPC acceleration is less efficient than MC in the sense it gives the *same accuracy* but computes the gPC coefficients leading to a loss of time, see tables 9.5 and 9.6. Nevertheless, at least, the accuracy is not deteriorated.

Remark 9.7 *Note that this additional computational times can be avoided by merely checking the values of the first coefficient gPC development. In the case it does not contribute to accuracy (i.e. if $gPC_1 \approx 0$), we do not perform the second step of the algorithm. But this implies the introduction of an additional numerical parameter.*

For the anisotropic problems, the 1st order gPC reduced models are efficient, see tables 9.5 and 9.6. For the gPC accelerated CV method, see figure 9.11 and table 9.5, taking polynomial orders $P > 6$ leads to a gain of about one decade in computational time with respect to MC for a given accuracy. Let us consider the case of the gPC accelerated IS method (figure 9.12 and table 9.6) for the anisotropic problems. For these problems, at fixed truncation order $P = 1$, the more the test-case is anisotropic, the more the gain is important. This enables understanding the main difference between gPC accelerated IS and gPC accelerated CV. The IS method suppose the sampling is done according to the pdf corresponding to the gPC₁ expansion. This pdf samples mainly in the region of interest. This explains the important gain obtained in the case of the "corner" problem (ratio ≈ 80). On the other hand, the gPC accelerated CV method only reuses the uniform sampling from the uniform pdfs on $[0, 1] \times [0, 1]$, leading to many misses and a less important gain, see table 9.5. Note that this latter has the advantage of being easier to implement than the gPC accelerated IS for orders $P > 1$ and ensuring a gain in accuracy even for isotropic problems. Finally, we would like to emphasize the fact that in the context of VR method, polynomials are not always the best reduced models. The latter 2-dimensional problems together with the step functions u^3 and u^6 of section 9.12.2 could be treated more efficiently by considering discontinuous reduced models. We want to emphasize here the possibility to combine gPC with piecewise polynomial functions see ME-gPC [294, 201, 278], or wavelets [185], Multi-Resolution Analysis [168] or i-gPC (see section 6 and [238, 242]). The same automated methodology presented in this section could be used with these reduced models. Of course, many of these approaches imply having *a priori* information on the smoothness of the integrand as it is for example the case in the example of this section 9.12.2.

Remark 9.8 (CV vs. IS) *The IS accelerated MC computations present important gains in the case of anisotropic problems and poor ones in the case of isotropic ones. This can be explained by the fact that IS approaches rely on a new sampling, done according to a new pdf allowing exploring the most important parts of the stochastic domain. In the case of isotropic problems, the sampling is the same as initially and the second step of the IS computation does not contribute to an increase of the accuracy while needing more computational time (two identical samplings).*

To finish this section, we briefly highlight the similarity of the above *simple integration* problems and the *initial or source sampling* phases in MC resolutions of the linear Boltzmann equation, see sections 9.8.1–9.9.1.

⁶⁹See figure 9.10 bottom right.

⁷⁰i.e. $u(x_1, x_2) = (u_0^1 + u_1^1 \phi_1(x_1))(u_0^2 + u_1^2 \phi_1(x_2))$ where u_j^i for $i = 1, 2$ and $j = 0, 1$ are computed independently in each directions x_1 and x_2 .

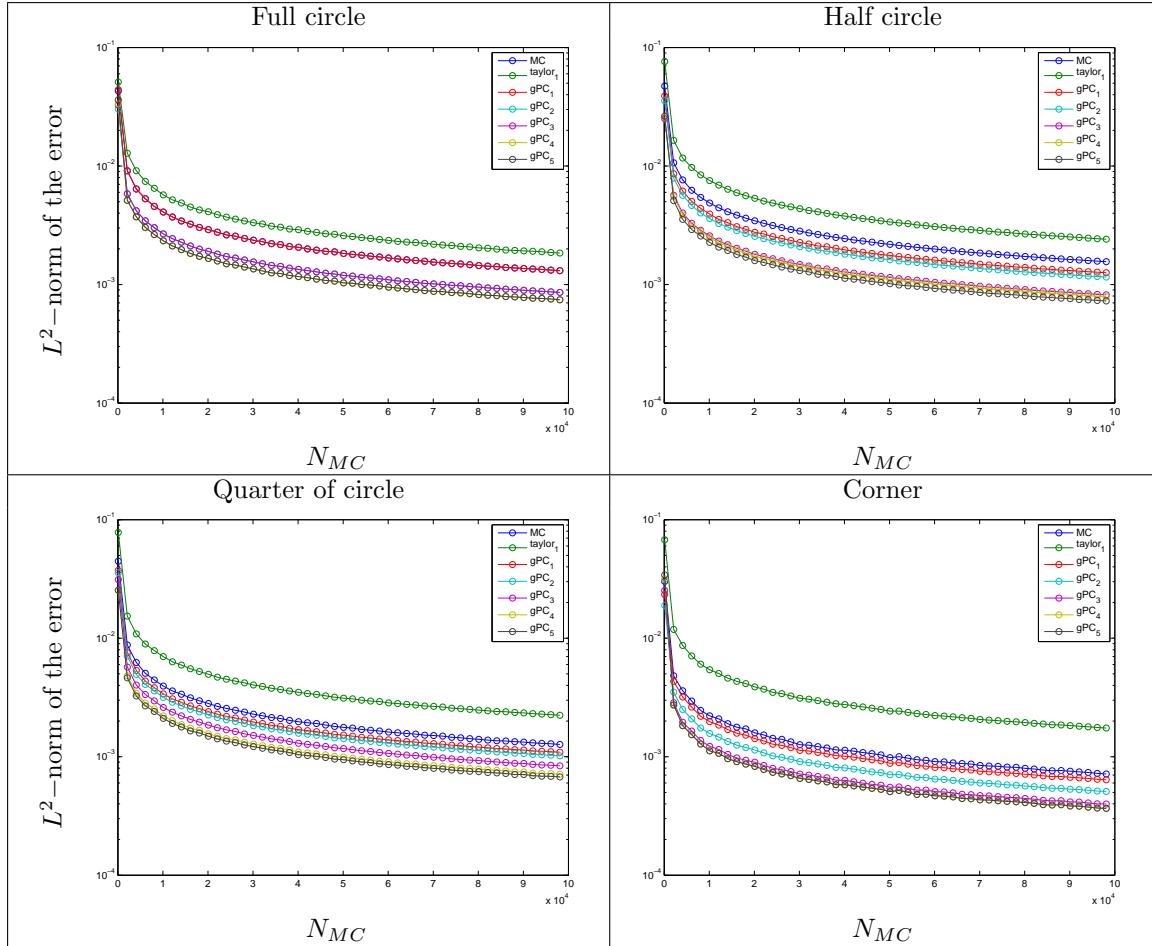


Figure 9.11: Convergence tests for the MC, Taylor-CV and gPC accelerated CV methods: integration over $[0, 1] \times [0, 1]$ with respect to the Lebesgue measure. The figure shows the logarithm of the estimation of the std with respect to the number of samples. Note that the gPC accelerations always reduce the variance of the MC method without *a priori* knowledge on the integrand.

9.12.3 Acceleration by gPC of the MC resolution of the linear Boltzmann equation

In this section, we suggest applying the latter materials, gPC acceleration of MC computations, to the resolution of the linear Boltzmann equation. Several authors [181, 17, 33] used VR techniques based on solutions of the transport equation obtained with a *deterministic* first step, for example to solve a reduced model built from the solution of the adjoint equation in a particular configuration [181, 17, 33]. In the presented algorithm, both steps, construction of a (gPC based) reduced model and resolution using it, are *stochastic*. It presents the advantage of having two unconditionally stable steps. We consider the test-case studied in [165, 181, 17, 33]. It consists in solving the stationary linear Boltzmann equation

$$\omega \cdot \nabla_{\mathbf{x}} u(\mathbf{x}, \omega) + \sigma_t u(\mathbf{x}, \omega) = \sigma_s \int u(\mathbf{x}, \omega') d\omega', \quad (9.194)$$

where $\mathbf{x} = (x_1, x_2) \in \mathcal{D} = [0, 17] \times [0, 15]$ and $\omega = \mathbb{S}^2$. The problem is homogeneous

$$\begin{cases} \sigma_t = 1, \\ \sigma_s = 0.9, \end{cases} \quad (9.195)$$

with particles incoming in the simulation domain with direction $\omega = (0, 1)$ from boundary $\{x_2 = 0, x_1 \in [0, 17]\}$ (the bottom one, see figure 9.13). The observable of interest is the flux through the surface

Full Circle	Accuracy	time	ratio of times
Classical MC ($N_{MC} = 100000$)	1.3×10^{-3}	6.623 s.	1 (reference)
gPC_1 ($N_{MC} = 100000$)	1.3×10^{-3}	13.453 s.	0.49 or 1
gPC_2 ($N_{MC} = 40000$)	1.3×10^{-3}	2.170 s.	3.05
gPC_3 ($N_{MC} = 40000$)	1.3×10^{-3}	2.216 s.	2.98
gPC_4 ($N_{MC} = 30000$)	1.3×10^{-3}	1.300 s.	5.09
gPC_5 ($N_{MC} = 30000$)	1.3×10^{-3}	1.340 s.	4.94
gPC_6 ($N_{MC} = 22000$)	1.3×10^{-3}	0.806 s.	8.21
gPC_8 ($N_{MC} = 19000$)	1.3×10^{-3}	0.827 s.	8.00
Half of Circle	Accuracy	time	ratio of times
Classical MC ($N_{MC} = 100000$)	1.5×10^{-3}	6.631 s.	1 (reference)
gPC_1 ($N_{MC} = 65000$)	1.5×10^{-3}	5.690 s.	1.16
gPC_2 ($N_{MC} = 55000$)	1.5×10^{-3}	4.085 s.	1.62
gPC_3 ($N_{MC} = 28000$)	1.5×10^{-3}	1.133 s.	5.85
gPC_4 ($N_{MC} = 28000$)	1.5×10^{-3}	1.167 s.	5.68
gPC_5 ($N_{MC} = 23000$)	1.5×10^{-3}	0.817 s.	8.11
gPC_7 ($N_{MC} = 17000$)	1.5×10^{-3}	0.502 s.	13.80
gPC_8 ($N_{MC} = 15000$)	1.5×10^{-3}	0.548 s.	12.10
Quarter of Circle	Accuracy	time	ratio of times
Classical MC ($N_{MC} = 100000$)	1.3×10^{-3}	6.681 s.	1 (reference)
gPC_1 ($N_{MC} = 65000$)	1.3×10^{-3}	5.660 s.	1.18
gPC_2 ($N_{MC} = 60000$)	1.3×10^{-3}	4.850 s.	1.37
gPC_3 ($N_{MC} = 38000$)	1.3×10^{-3}	1.996 s.	3.34
gPC_4 ($N_{MC} = 30000$)	1.3×10^{-3}	1.292 s.	5.17
gPC_5 ($N_{MC} = 25000$)	1.3×10^{-3}	0.948 s.	7.04
gPC_7 ($N_{MC} = 17000$)	1.3×10^{-3}	0.571 s.	11.70
gPC_8 ($N_{MC} = 15000$)	1.3×10^{-3}	0.543 s.	12.30
Corner	Accuracy	time	ratio of times
Classical MC ($N_{MC} = 56000$)	10^{-3}	2.108 s.	1 (reference)
gPC_1 ($N_{MC} = 40000$)	10^{-3}	2.170 s.	0.97
gPC_2 ($N_{MC} = 25000$)	10^{-3}	0.867 s.	2.43
gPC_3 ($N_{MC} = 15000$)	10^{-3}	0.324 s.	6.50
gPC_4 ($N_{MC} = 13000$)	10^{-3}	0.257 s.	8.20
gPC_5 ($N_{MC} = 13000$)	10^{-3}	0.266 s.	7.92
gPC_7 ($N_{MC} = 8700$)	10^{-3}	0.153 s.	13.77
gPC_8 ($N_{MC} = 7800$)	10^{-3}	0.150 s.	14.05

Table 9.5: Comparison between times for reaching the same accuracy with the Classical MC method and the gPC accelerated CV for the problems of figure 9.10. The gPC acceleration gives satisfactory results (ratio > 1) for every problems. Note that the ratio of the full circle problem has been corrected according to remark 9.7.

$\{x_2 = 15, x_1 \in [0, 17]\}$ (top boundary, see figure 9.13).

Let us now apply an IS variance reduction technique for this problem. It consists of introducing $u^*(\mathbf{x}, \omega) > 0$ such that $\int u^*(\mathbf{x}, \omega) d\omega = 1, \forall \mathbf{x} \in \mathcal{D}$ as a reduced model and perform the change of variable

$$u(\mathbf{x}, \omega) = f(\mathbf{x}, \omega)u^*(\mathbf{x}, \omega). \quad (9.196)$$

Introducing (9.196) into (9.194) leads to the following transport equation of unknown f :

$$\omega \cdot \nabla_{\mathbf{x}} f(\mathbf{x}, \omega) + \left(\sigma_t + \frac{\nabla_{\mathbf{x}} u^*(\mathbf{x}, \omega)}{u^*(\mathbf{x}, \omega)} \right) f(\mathbf{x}, \omega) = \frac{\sigma_s}{u^*(\mathbf{x}, \omega)} \int f(\mathbf{x}, \omega') u^*(\mathbf{x}, \omega') d\omega'. \quad (9.197)$$

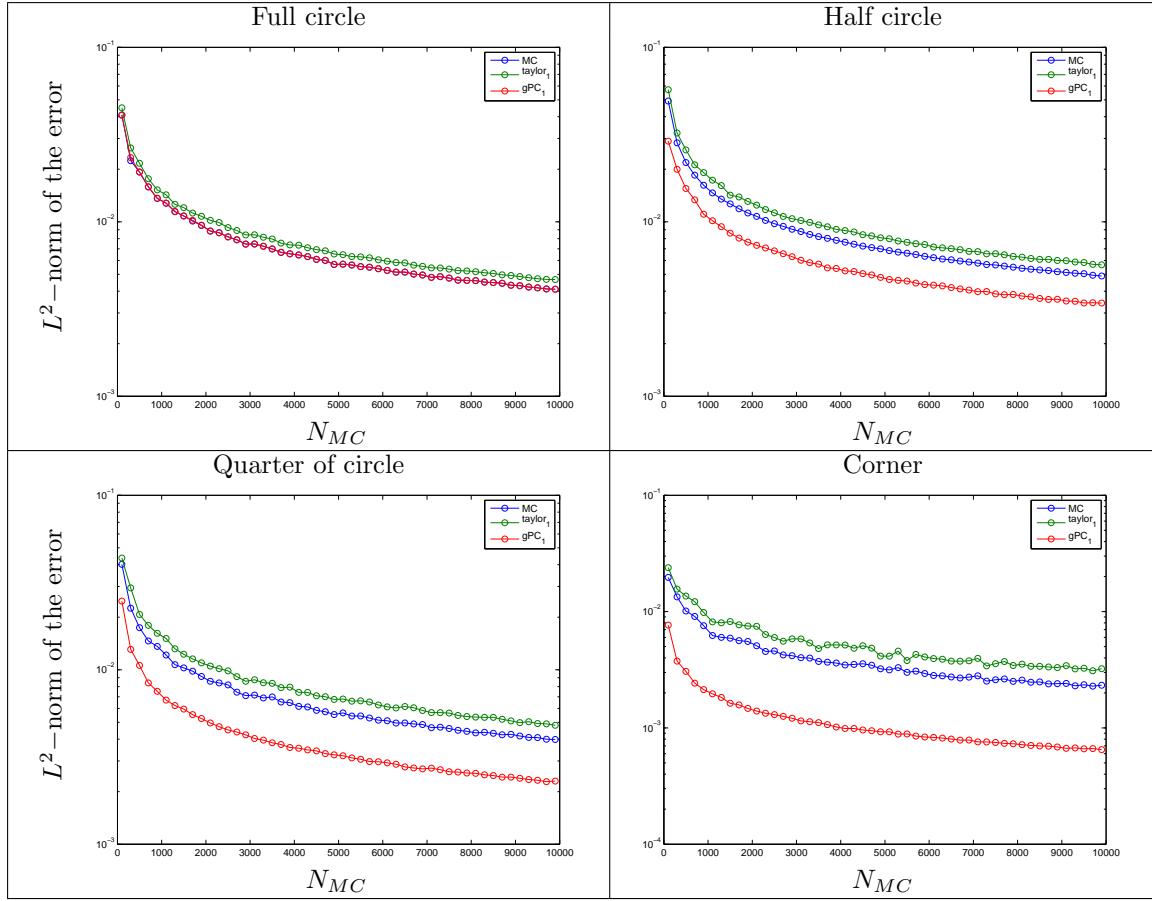


Figure 9.12: Convergence tests for the MC, Taylor-IS and gPC accelerated IS methods: integration over $[0, 1] \times [0, 1]$ with respect to the Lebesgue measure. The figure shows the logarithm of the estimation of the std with respect to the number of samples.

Full Circle		Accuracy	time	ratio of times
Classical MC ($n = 100000$)		1.3×10^{-3}	6.623 s.	1 (reference)
gPC_1 ($n = 100000$)		1.3×10^{-3}	13.263 s.	0.50 or 1
Half of Circle		Accuracy	time	ratio of times
Classical MC ($n = 100000$)		1.5×10^{-3}	6.687 s.	1 (reference)
gPC_1 ($n = 50000$)		1.5×10^{-3}	3.347 s.	2.00
Quarter of Circle		Accuracy	time	ratio of times
Classical MC ($n = 100000$)		1.3×10^{-3}	6.652 s.	1 (reference)
gPC_1 ($n = 29000$)		1.3×10^{-3}	1.150 s.	5.78
Corner		Accuracy	time	ratio of times
Classical MC ($n = 55000$)		10^{-3}	2.020 s.	1 (reference)
gPC_1 ($n = 3900$)		10^{-3}	0.025 s.	80.79

Table 9.6: Comparison between times for reaching the same accuracy with the Classical MC method and the gPC accelerated IS for the problems of figure 9.10.

For simplicity, in the following paragraph, we consider $u^*(\mathbf{x}, \omega) = u^1(\mathbf{x})u^2(\omega)$, relaxing the conditions⁷¹ on u^* to $\int u^2(\omega)d\omega = 1$ and $u^1 > 0$, $u^2 > 0$. Besides, the expressions of u^1 and u^2 are chosen as follows:

⁷¹Note that care will be taken to satisfy $\sigma_t + \frac{\nabla_{\mathbf{x}}u^*}{u^*} > 0$ for convenience.

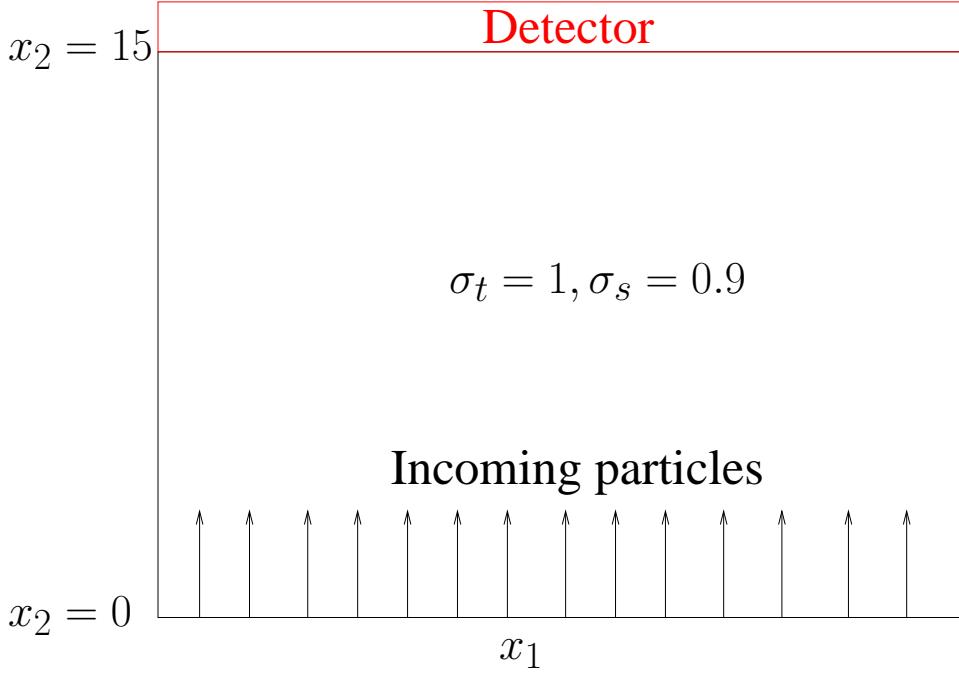


Figure 9.13: Description of the test-case: the particles are incoming from the surface $\{x_2 = 0, x_1 \in [0, 17]\}$ with direction $\omega = (0, 1)$. The problem is homogeneous and the detector is the surface $\{x_2 = 15, x_1 \in [0, 17]\}$ located at 15 mean free paths of the particles as $\sigma_t = 1$.

- $u^1(\mathbf{x}) = \exp \sigma_t \vec{F} \cdot \mathbf{x}$ consists of spatial exponential transform of parameter⁷² \vec{F} , see [17, 280].
- $u^2(\omega) = 1 + F\omega \cdot \omega_0$ where $F = |\vec{F}| \in [0, 1]$ can be compared to an angular biasing toward direction ω_0 through a change in the distribution of the directions ω after collisions⁷³, see [17, 181, 280].

With these choices and adding the fact that the direction of interest is taken as $\omega_0 = (0, 1)$, equation (9.197) simplifies to

$$\omega \cdot \nabla_{\mathbf{x}} v(\mathbf{x}, \omega_1) + \sigma_t (1 + K\omega_1) f(\mathbf{x}, \omega_1) = \frac{\sigma_s}{1 + K\omega_1} \int_{-1}^1 f(\mathbf{x}, \omega'_1) \frac{1 + K\omega'_1}{\pi \sqrt{1 - (\omega'_1)^2}} d\omega'_1. \quad (9.198)$$

In the above expression, we have

$$\begin{cases} w(\omega_1) = \frac{1 + F\omega_1}{\pi \sqrt{1 - \omega_1^2}} > 0, \text{ for } F \in [0, 1], \omega_1 \in [-1, 1], \\ \int_{-1}^1 w(\omega_1) d\omega_1 = 1. \end{cases}$$

In other words, $w(\omega_1)d\omega_1$ is a probability measure. We introduce the gPC basis associated to the pdf w , denoted by $(\phi_k(\omega_1))_{k \in \mathbb{N}}$. Here, this basis corresponds to the normalized Chebyshev polynomials $\phi_0(x) = 1, \phi_1(x) = \sqrt{2}x, \phi_2(x) = 2\sqrt{2}x^2 - \sqrt{2}, \dots$. Applying the gPC-IS method described previously consists in running one first Classical MC computation in order to evaluate the gPC coefficients u_0, u_1 of the solution $u(x, \omega_1) \approx u_0(x)\phi_0(\omega_1) + u_1(x)\phi_1(\omega_1)$. They are then introduced in a second 'biased'

⁷²This parameter \vec{F} will be linked afterward with gPC coefficients.

⁷³This parameter F will be linked afterward with gPC coefficients and u^2 will be identified as a normalized gPC development.

computation *via* the coefficient⁷⁴

$$F = \frac{u_1}{u_0} \sqrt{2}, \quad (9.199)$$

at the basis of the exponential transform. Relating K and the gPC coefficients u_0, u_1 *via* (9.199) aims at emphasizing the two approaches (exponential transform and gPC) are complementary. The second (gPC) helps identifying a relevant reduced model, the first one (exponential transform) is only a way to use the gPC reduced model into the MC computations.

Remark 9.9 (Comparison with results obtained with Deterministic Models in [17]) *Figure 9.14 compares the coefficient $F(x_2)$ obtained in [17] with a deterministic reduced model (left picture) to the one obtained with the gPC-IS variance reduction method. Both approaches recover the relevant coefficient of the exponential transform allowing a variance reduction in the configuration of interest.*

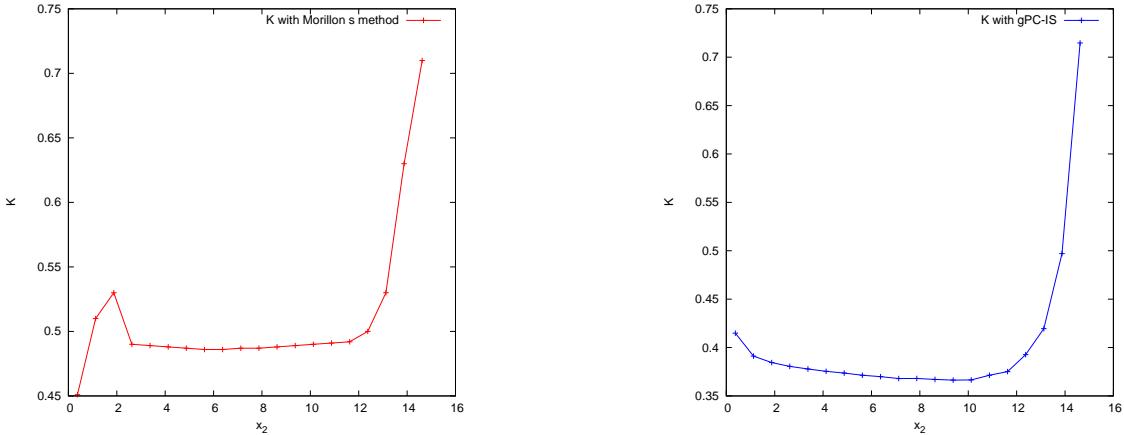


Figure 9.14: Comparison of coefficients of the values of coefficient F with respect to x_2 , 20 cells, obtained with the gPC-IS, on the right hand side of the figure and the method suggested in [17] (figure VIII.1, p. 137) on the left hand side of the figure. The figures show a good agreement despite the use of different methods.

Let us now present several numerical results obtained with the above procedure. Figure 9.15 shows the flux through the surface $\{x_2 = 15, x_1 \in [0, 17]\}$ for several computations. On figure 9.15 (left), three computations are displayed: there are two MC computations with 200000 and 5000000 MC particles and one gPC-IS computation with 200000 particles. With 200000 particles, the gPC-IS approach gives results comparable with the Classical MC approach with 5000000 particles. Note that the computational cost is diminished as the gPC-IS (200000 particles) calculation took 12 s. whereas the Classical MC approach (5000000), for approximately the same accuracy, took 2 min. 15 s. The quantitative results for this problem are displayed in tables 9.7 and 9.8. Figure 9.15 (right) shows two simulations. The first one is a classical MC calculation with 10000000 particles (CPU time 3 min. 11 s.). The second one is a gPC-IS calculations with 200000 particles in the first step (step (i) of remark 9.6) of the algorithm (estimation of the gPC coefficients $(u_k)_{k \in \{0,1\}}$) and 1000000 particles for the second step (CPU time 36s.) (step (ii) of remark 9.6). This example illustrates the fact that different combinations are possible for the number of MC points used in the different steps of the algorithm to increase the gain. The quantitative results for this problem are displayed in table 9.9.

Tables 9.7, 9.8 and 9.9 present the CPU times, the estimated standard deviation (Std) and the quality (or figure of merit, FOM), defined by $Q = \frac{1}{\text{Std}^2 \times \text{CPU time}}$ as in [17], with respect to the number of MC particles of the computations for solving problem (9.194). The computations are performed with a classical MC method (table 9.7), a gPC-IS method (table 9.8) and a gPC-IS tuned method (table 9.9). The tuning of the last approach consists of taking less particles (10 times less in this case) to estimate the

⁷⁴Due to the fact that $\phi_1(x) = \sqrt{2}x$ and by normalization.

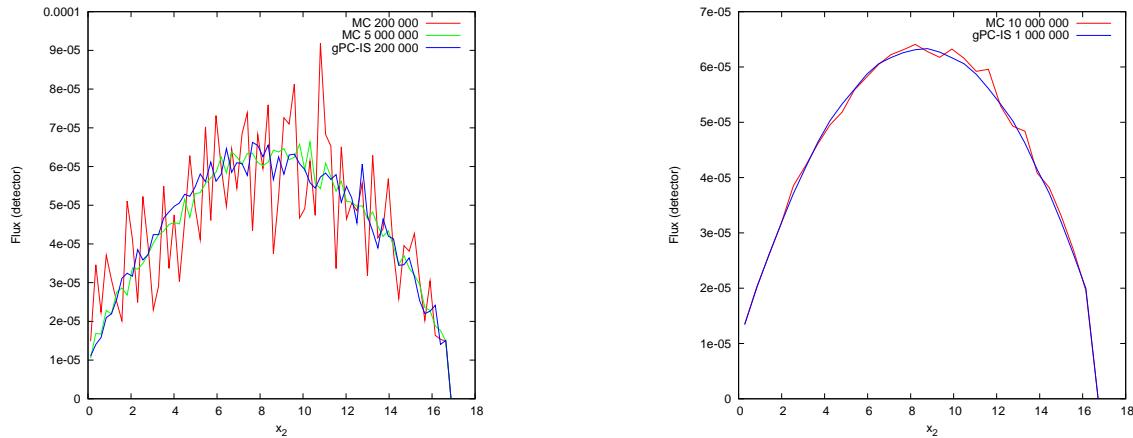


Figure 9.15: Left: Three computations are displayed: two classical MC calculations with 200000 (CPU time 7 s.) and 5000000 (CPU time 2 min. 15 s.) particles and one gPC-IS accelerated calculation with 200000 (CPU time 12 s.) particles. The accelerated computation (only 200000 particles) reaches the same accuracy than the MC calculation (with 5000000 particles). Right: the first computation (classical MC) has 10000000 (CPU time 3 min. 11 s.) particles whereas the gPC-IS computation uses 200000 particles to evaluate the coefficient F , see (9.199), and 1000000 particles in the second step of the algorithm (total CPU time for this gPC-IS computation: 36 s.). The gPC-IS computation is much less noisy than the converged MC one.

gPC coefficients u_0, u_1 (step (i) of remark 9.6) than for the biased computation following the estimation (step (ii) of remark 9.6). Note that this tuning could have already been used in the computations of sections 9.12.2–9.12.2.

MC points	CPU time (s)	Std	Q_{MC}	$\frac{Q_{gPC}}{Q_{MC}}$
10000	0.21	0.01607	18439.49	1
100000	1.93	0.01432	2526.71	1
1000000	18.80	0.01374	281.75	1
10000000	190.01	0.01372	27.95	1

Table 9.7: Classical MC computations: we display the CPU times, the estimated standard deviation (Std) and the quality $Q_{MC} = \frac{1}{\text{Std}^2 \times \text{CPU time}}$ with respect to the number of MC particles in the simulations.

Tables 9.7 and 9.8 allow comparing classical MC calculations to the gPC-IS ones. The gain with gPC-IS is important, see table 9.8, and tends to increase with the number of MC particles N_{MC} . This is related to the fact that the gPC coefficients are more accurately computed as N_{MC} increases. The gains represent more than one decade.

gPC points	CPU time (s)	Std	Q_{gPC}	$\frac{Q_{gPC}}{Q_{MC}}$
10000	0.49	0.009049	24923.13	1.35
100000	4.70	0.002770	27729.53	10.97
1000000	46.00	0.002205	4471.20	15.86
10000000	461.05	0.002132	475.88	17.02

Table 9.8: gPC-IS accelerated computations: we display the CPU times, the estimated standard deviation (Std) and the quality $Q_{gPC} = \frac{1}{\text{Std}^2 \times \text{CPU time}}$ with respect to the number of MC particles in the simulations. For these computations, the same number of particles is used to estimate the gPC coefficients u_0, u_1 . The gain with respect to MC computations are also given.

Tables 9.8 and 9.9 allow comparing two gPC-IS approaches with different numbers of particles used in step (i) of remark 9.6. For the computations of table 9.8, the same number of particles is used for the

two steps, i.e. $N_{(i)} = N_{(ii)} = N_{MC}$. For the computations of table 9.9, different numbers of particles are used for the two steps and we choose $N_{(i)} < N_{(ii)}$. In practice, we took $N_{(i)} = \frac{1}{10} \times N_{(ii)} = \frac{1}{10} \times N_{MC}$. The comparison of tables 9.8 and 9.9, allows putting forward some possible optimizations: it is possible

gPC* points	CPU time (s)	Std	Q_{gPC^*}	$\frac{Q_{gPC^*}}{Q_{MC}}$
10000	0.30	0.009145	39857.61	2.16
100000	3.01	0.006505	7851.25	3.10
1000000	29.8	0.002678	4679.10	16.60
10000000	313.12	0.002132	702.61	25.13

Table 9.9: gPC*-IS accelerated computations: we display the CPU times, the estimated standard deviation (Std) and the quality $Q_{gPC^*} = \frac{1}{\text{Std}^2 \times \text{CPU time}}$ with respect to the number of MC particles in the simulations. For these computations*, we used a lower number of particles ($N_{MC}/10$ each time) so as to estimate the gPC coefficients u_0, u_1 . The gain with respect to MC computations are also given.

using less particles in step (i) without diminishing the accuracy of the full calculation, see table 9.9 (last line). With the precedent choice (i.e. $N_{(i)} = \frac{1}{10} \times N_{(ii)}$), it is possible to gain more than two decades in quality (see the rate of 25.13 in table 9.9) where the gPC-IS ($N_{(i)} = N_{(ii)}$) allowed a gain of 17.02 (see table 9.8). The accuracy is maintained with less particles in step (i), i.e. a faster restitution time for the simulation.

Remark 9.10 (Unstationary Problems) *We here illustrated the gPC based VR techniques on a stationary problem. Its application to unstationary problems is straightforward: it consists of repeating the same algorithm as precedently but for each time steps of the simulation.*

Note that we do not apply the gPC-CV variance reduction technique to the resolution of the linear Boltzmann equation. The application of the CV formalism is straightforward thanks to the material of the previous sections: the CV approach implies a transformation of the initially considered linear Boltzmann equation into a new one with a modified source term (see also [45] for some detailed explanations) which can be handled thanks to the material of section 9.9.

9.12.4 Summary

To sum up this section, we first put forward what we consider an important analogy between AP schemes and Variance Reduction techniques. Both aim at reducing the constant multiplying the convergence rate of the MC method. Two situations occur:

- if the stiff regime of interest is clearly identified, we think it is more efficient relying on an AP scheme formalism to tackle the problem. The identified stiff regime of interest can then be plugged in the MC computations *via* a change of variable closely related to the IS variance reduction method. Examples will be given in the next chapter 10 in a coupled context.
- If the stiff regime/configuration of interest is complex to identify, we suggest the introduction of a gPC based reduced model to accelerate automatically (i.e. thanks to computations rather than analysis) the MC resolution. Note that this is not the first time UQ based methods are applied in order to reduce variance, see [40, 41] for example. The purpose here is not to be exhaustive but to emphasize close relations between the different parts of the document (part II and III) and hint at a continuity in my contributions.

Finally, section 9.12.2 (dealing with simple integration problems without further hints at the linear Boltzmann equation) may seem far from the considerations of the other sections of this chapter. We insist this is not the case. The problematic is comparable to the one of sections 9.8.1–9.9.1, dealing with initial and source samplings. Those are crucial steps in a direct MC resolution of the linear Boltzmann equation.

Chapter 10

Monte-Carlo methods for the nonlinear Boltzmann equation

Asymptotic-Preserving Monte-Carlo schemes for two different physics, two different regimes

Contents

10.1 Boltzmann equation coupled to Bateman system (neutronics)	254
10.1.1 Classical MC schemes for neutron transport	255
10.1.2 An Asymptotic Preserving MC scheme for neutron transport	260
10.1.3 Summary	265
10.2 Boltzmann equation coupled to Stefan's law (photronics)	266
10.2.1 Classical Monte-Carlo schemes for photon transport	269
10.2.2 Two Asymptotic Preserving MC schemes for photon transport	274
10.2.3 Summary	283

In this chapter, we present our contribution to the MC resolution of the nonlinear Boltzmann equation. The nonlinearity comes from a coupling with another equation or system of equations (depending on the physics). The first section 10.1 deals with a neutronic application and the second one, section 10.2, is related to photonics. Independently of the application, the resolution strategies are mainly based on a splitting between the linearized Boltzmann counterpart, solved with an MC method, and the other set of equations, usually solved with a deterministic scheme. In plasma physics, such split resolution scheme implying both an MC resolution and a deterministic one is commonly called a Particle In Cell (PIC) method. In neutronics or photonics, such denomination does not really hold and the solvers are usually called MC schemes. But we insist both terms describe the same resolution strategy.

The coupling with some additional equation or set of equations introduces nonlinearity. Linearity is mandatory for an MC resolution. The latter intensively uses the fact that if every MC particle u_p is a particular solution of the transport equation (see expression (9.15)), $\sum_{p=1}^{N_{MC}} u_p$ is also a (converging in law, see [165]) solution of the same equation. Consequently, to apply an MC discretisation, a relevant linearisation of the coupled system has to be introduced. By relevant, we mean the linearisation choice must be driven by the asymptotic regime one aims at capturing. Care will be taken to highlight this in the two following sections. Depending on the physics, the regime of interest may differ and the relevant linearisation bearing good asymptotical properties too. For the two physical applications, we first briefly present the system of equations. We then describe the classical MC resolution schemes. We identify and illustrate their main drawbacks with respect to the asymptotic regime of interest. We finally suggest some Asymptotic Preserving (AP) MC schemes.

10.1 Boltzmann equation coupled to Bateman system (neutronics)

In this section, we are interested in the resolution of the time-dependent problem of particle transport in a media whose composition evolves with time due to particle interactions (reactions). We suppose transport to be driven by the linear Boltzmann equation (10.1a) for particles having position $\mathbf{x} \in \mathcal{D} \subset \mathbb{R}^3$, velocity¹ $\mathbf{v} \in \mathbb{R}^3$, at time $t \in [0, T] \subset \mathbb{R}^+$. Quantity $u(\mathbf{x}, t, \mathbf{v})$ is the density of presence of the particles at $(\mathbf{x}, t, \mathbf{v})$. We assume the time variation of the media composition (vector $\boldsymbol{\eta}$) can be accurately modeled by Bateman equations (10.1b) (see [126]). Quantity $\boldsymbol{\sigma}_r(\mathbf{x}, t, \mathbf{v}) = (\sigma_r^1(\mathbf{x}, t, \mathbf{v}), \dots, \sigma_r^M(\mathbf{x}, t, \mathbf{v}))^t$ is the vector of reaction rates (depending on velocity/energy). We consider the vector of reaction rates is stiff. By stiff, we mean the characteristic time for the reactions is much smaller than the transport one, at least for some media components, in some subsets of the computation domain \mathcal{D} . Care will be taken to identify this regime in the following section. As a result, problem (10.1) is stiff, nonlinear and strongly coupled:

$$\begin{cases} \partial_t u(\mathbf{x}, t, \mathbf{v}) + \mathbf{v} \nabla_{\mathbf{x}} u(\mathbf{x}, t, \mathbf{v}) + \sigma_t(\boldsymbol{\eta}(\mathbf{x}, t), v) v u(\mathbf{x}, t, \mathbf{v}) = \int \sigma_s(\boldsymbol{\eta}(\mathbf{x}, t), \mathbf{v}', \mathbf{v}) v u(\mathbf{x}, t, \mathbf{v}') d\mathbf{v}', \\ \partial_t \boldsymbol{\eta}(\mathbf{x}, t) = \int \boldsymbol{\sigma}_r(\boldsymbol{\eta}(\mathbf{x}, t), \mathbf{v}) v u(\mathbf{x}, t, \mathbf{v}) d\mathbf{v}. \end{cases} \quad (10.1)$$

The interaction of particles with matter is described *via* the total interaction probability of particles with media $\sigma_t(\mathbf{x}, t, \mathbf{v})$ and a scattering term $\sigma_s(\mathbf{x}, t, \mathbf{v}, \mathbf{v}')$. Macroscopic interaction properties depend on both microscopic ones designated by $(\sigma_{\alpha, m})_{\alpha \in \{t, s\}}$ and the media composition vector $\boldsymbol{\eta}(\mathbf{x}, t) = (\eta_1(\mathbf{x}, t), \dots, \eta_M(\mathbf{x}, t))^t$:

$$\sigma_t(\boldsymbol{\eta}(\mathbf{x}, t), v) = \sum_{m=1}^M \sigma_{t,m}(v) \eta_m(\mathbf{x}, t), \text{ and } \sigma_s(\boldsymbol{\eta}(\mathbf{x}, t), \mathbf{v}', \mathbf{v}) = \sum_{m=1}^M \sigma_{s,m}(\mathbf{v}', \mathbf{v}) \eta_m(\mathbf{x}, t). \quad (10.2)$$

Under this general form, model (10.1) can be relevant in many fields of applications. The Bateman counterpart (10.1b) may be considered a particular case of the Lotka-Volterra system (see [223]) in which we only kept the strong coupling term. Amongst the applications (non exhaustive list), one can quote biology [223] with population dynamics, or physics with burn-up computations in neutronics [95, 147, 148, 98]. In the latter case, the particles (u) are neutrons, the media ($\boldsymbol{\eta}$) is composed of nuclides. Of course, the numerical methodology we develop in this paper is general and can be broadened to a larger scope.

Our aim is to pedagogically put forward the limitations of solvers involving a splitting between the transport equation (solved using MC method) and the Bateman system when the latter is stiff. For this, it is enough working on a simplified problem. Let us first assume a monokinetic particle transport equation, i.e. $u(\mathbf{x}, t, \mathbf{v}) = u(\mathbf{x}, t, \omega)$. Besides, considering the scalar Bateman equation ($\eta(\mathbf{x}, t) = \eta_1(\mathbf{x}, t)$) where the reactions are modeled only with a scalar reaction rate $\sigma_r(\mathbf{x}, t) = \sigma_r \eta(\mathbf{x}, t)$ does not alter the nature of the coupling. Of course, the material of this section can be extended to the general case. Its complete description has been published in [3] together with numerical examples highlighting the drawbacks of the classical scheme we describe here. Still, we insist the material of this section is not redundant with [3], it is complementary. Under the previous hypothesis, the collisional counterpart becomes

$$\forall \alpha \in \{t, r\}, \sigma_\alpha(\eta(\mathbf{x}, t)) = \sigma_\alpha(\mathbf{x}, t) = \sigma_\alpha \eta(\mathbf{x}, t), \quad \sigma_s(\eta(\mathbf{x}, t), \omega', \omega) = \sigma_s(\mathbf{x}, t, \omega', \omega) = \sigma_s \eta(\mathbf{x}, t) P_s(\omega', \omega).$$

The term $P_s(\omega', \omega)$ corresponds to the probability for a particle having direction ω' to get out of an interaction with direction ω (i.e. $\forall \omega, \int P_s(\omega', \omega) d\omega' = 1$). The simplified system, still strongly coupled,

¹or energy $v = |\mathbf{v}| \in \mathbb{R}^+$, direction $\omega = \frac{\mathbf{v}}{|\mathbf{v}|} \in \mathbb{S}^2$.

is thus a 2-equations system:

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + v\omega \nabla_{\mathbf{x}} u(\mathbf{x}, t, \omega) + \sigma_t \eta(\mathbf{x}, t) v u(\mathbf{x}, t, \omega) = \sigma_s \eta(\mathbf{x}, t) \int P_s(\omega', \omega) v u(\mathbf{x}, t, \omega') d\omega', \\ \partial_t \eta(\mathbf{x}, t) = \eta(\mathbf{x}, t) \sigma_r v \int u(\mathbf{x}, t, \omega) d\omega. \end{cases} \quad (10.3)$$

We also introduce $\sigma_a = \sigma_t - \sigma_s$ as in the previous chapter. The above system is general with respect to the coupling and can still² describe very different regimes (absorbing, multiplicative, reactive etc.) by changing the values of $(\sigma_\alpha)_{\alpha \in \{s, t, a, r\}}$. Let us now describe the classical methodology applied to solve (10.3) in section 10.1.1 and the new Asymptotic Preserving MC scheme we suggest in section 10.1.2.

10.1.1 Classical MC schemes for neutron transport

We first describe the most common methodology to solve system (10.3). It consists in a splitting between the transport phase (10.3a) and the Bateman phase (10.3b). Such splitting is very convenient in practice. For example, one can solve system (10.3) by relying on two different simulation codes: one solving the linear Boltzmann equation and the other the Bateman system. The idea is to use the output of the first as inputs of the second and iterate. Our aim here is to highlight the main drawback of such methodology when encountering a stiff reaction regime (neutronics for example, see [97, 98, 148, 95, 96], [3]).

The classical MC scheme description for the resolution of (10.3)

We here present the (classical) split solver applied for the resolution of system (10.3). We analyse the solver (linear Boltzmann phase + Bateman phase) on one time step $t \in [0, \Delta t]$ in the limit of an infinitely accurate MC resolution (as in [77] in photonics). In order to apply an MC scheme, one needs a linearisation hypothesis. The latter can be summed-up as follows: the cross-sections are assumed constant (explicit) with respect to time on $[0, \Delta t]$, i.e.

$$\sigma_\alpha(\eta(\mathbf{x}, t)) \approx \sigma_\alpha(\eta(\mathbf{x}, 0)) = \sigma_\alpha \eta_0(\mathbf{x}), \forall \alpha \in \{s, t, a, r\}.$$

The transport phase in $[0, t]$ consequently becomes

$$\partial_t u(\mathbf{x}, t, \omega) + v\omega \nabla_{\mathbf{x}} u(\mathbf{x}, t, \omega) + v\sigma_t \eta_0(\mathbf{x}) u(\mathbf{x}, t, \omega) = \int v\sigma_s \eta_0(\mathbf{x}) P_s(\mathbf{x}, \omega', \omega) u(\mathbf{x}, t, \omega') d\omega'. \quad (10.4)$$

It is solved with an MC scheme (see chapter 9). The MC resolution is here performed with the non-analog scheme of section 9.4 but the conclusions of this section does not depend on this choice³. The MC resolution is followed by a Bateman phase during which the nuclide density η must be consistently updated.

Regarding the transport phase, by construction of the non-analog MC resolution scheme, theorem 3.2.1 of [165] ensures the convergence of the MC solver toward the solution of (10.4) in the limit $N_{MC} \rightarrow \infty$ for the considered time step $[0, \Delta t = t]$. In order to ensure the convergence of the scheme for (10.3) in the limit $\Delta t \rightarrow 0$, it remains to perform consistent tallies to update the nuclide concentrations at the end of the time step. This contribution, or tally, must be in agreement with both the structure of equation (10.3b), and the hypothesis made in the previous transport phase (i.e. $\eta(\mathbf{x}, t) = \eta(\mathbf{x}, 0) = \eta_0(\mathbf{x})$). We here rely on an explicit Euler scheme for the time discretisation of the Bateman counterpart⁴. The application of the split MC/explicit Euler solver consists in integrating (10.3b) on time step $[0, t]$ so that the equation becomes

$$\eta(\mathbf{x}, t) = \eta_0(\mathbf{x}) + v\sigma_r \int_0^t \eta(\mathbf{x}, s) \int u(\mathbf{x}, s, \omega) d\omega ds. \quad (10.5)$$

²Some considerations about the generalization are tackled in section 10.1.3.

³Asymptotically, every other MC scheme converges and consequently allows recovering (10.4).

⁴Others are detailed in [3], once again, the conclusions of this section do not depend on this particular choice.

With the explicit hypothesis applied during the transport phase, the time integrated part rewrites

$$\int_0^t \eta(\mathbf{x}, s) \int u(\mathbf{x}, s, \omega) d\omega ds \approx \Delta t \eta_0(\mathbf{x}) \frac{1}{\Delta t} \int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds. \quad (10.6)$$

The scheme is explicit for the nuclide concentration η but not for the density of particles u . The integral $\frac{1}{\Delta t} \int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds$ is evaluated thanks to a *tally* during the MC phase. Time integrated observables are commonly called *track length estimator* in an MC resolution context (see [268, 165, 173, 52]). The MC discretisation of the latter expression is obtained plugging the MC discretisation $\sum_p u_p$ of u , obtained from (9.15), into (10.6). This leads to

$$\frac{1}{\Delta t} \int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds \stackrel{N_{MC}}{\approx} \frac{1}{\Delta t} \int_0^t \int \sum_{p=1}^{N_{MC}} u_p(\mathbf{x}, s, \omega) d\omega ds = \sum_{p=1}^{N_{MC}} \frac{1}{\Delta t} \int_0^t w_p(s) \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds. \quad (10.7)$$

The contribution of the MC particle p is non-zero only if $\int_0^t \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds$, the *local time at x* , is non zero. Let us introduce the N_p local times $(t_p^l)_{l \in \{1, \dots, N_p\}}$ spent at position \mathbf{x} for particle p . Then (10.7) can be rewritten

$$\frac{1}{\Delta t} \int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds \approx \sum_{p=1}^{N_{MC}} \sum_{i=1}^{N_p} \frac{1}{\Delta t} \int \delta_{t_p^i}(s) w_p(s) ds = \frac{1}{\Delta t} \sum_{p=1}^{N_{MC}} \sum_{l=1}^{N_p} w_p(t_p^l) = \sum_{p=1}^{N_{MC}} \sum_{l=1}^{N_p} \Delta u_p^l(\mathbf{x}). \quad (10.8)$$

The term $\Delta u_p^l(\mathbf{x})$ is the contribution, track length estimator, of the particle p for the time t_p^l spent at \mathbf{x} . Suppose the tracking of the MC particles is now instrumented to compute the sums over the flights of each particle in each cell, this means we will have access to $\sum_{p=1}^{N_{MC}} \sum_{l=1}^{N_p} \Delta u_p^l(\mathbf{x})$. Then system is closed by updating η at each position \mathbf{x} by plugging the previous contribution into (10.5). The time step is then over.

The domain is usually discretised into $N_{\mathbf{x}}$ cells as in section 9.6 with constant nuclide densities per cell. In other words, we have $\eta_0(\mathbf{x}) = \sum_{i=1}^{N_{\mathbf{x}}} \eta_0^i \mathbf{1}_{\mathcal{D}_i}(\mathbf{x})$ (equivalent to constant cross-sections per cell). Integrating the local time $\int \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds$ in one cell leads to a local interval of time $\int_{\mathcal{D}_i} \int \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds d\mathbf{x} = \int \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}_p(s)) ds$. If we introduce the N_p local *intervals of time* ($\Delta t_p^l = t_p^{l+1} - t_p^l$) $_{l \in \{1, \dots, N_p\}}$ spent in cell i for particle p during time step $[0, \Delta t = t]$, the expression of the contribution of particle p to cell i becomes

$$\frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} \Delta u_p^l(\mathbf{x}) d\mathbf{x} = \Delta u_p^{l,i} = \frac{1}{\Delta t} \int_{t_p^l}^{t_p^{l+1}} w_p(s) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}_p(s)) ds = w_p(t_p^l) \frac{1 - e^{-v\sigma_a^i \eta_0^i \Delta t_p^l}}{v\sigma_a^i \eta_0^i \Delta t}. \quad (10.9)$$

It is consistent and it only remains to use the above expression in (10.8) to update the nuclide concentration and end the time step.

Algorithm 15 presents how the direct resolution of section 9.5.2 must be instrumented to perform the previous tallies, mandatory to update the nuclide density η consistently. The tracking phase only needs the addition of one line, function `compute_track_length`, which is detailed in algorithm 16.

Algorithm 15: The general canvas for the different MC schemes described in term of algorithmic operations in order to compute (direct) $U(\mathbf{x}, t) = \int u(\mathbf{x}, t, \mathbf{v}) d\omega dv$ with instrumentations to compute track length estimators.

```

1 #SAMPLING described in algorithm 7 or algorithm 8
2 call sampling( $N_{MC}$ )
3 set  $t = \Delta t$ 
4 #Time step loop
5 while  $t < T$  do
6     #Initialize to zero the array of the quantity of interest on the whole simulation domain  $\mathcal{D}$ 
7     set  $U(\mathbf{x}, t) = 0 \forall \mathbf{x} \in \mathcal{D}$ 
8     set  $\Delta U(\mathbf{x}, t) = 0 \forall \mathbf{x} \in \mathcal{D}$ 
9     #TRACKING: make sure each  $u_p$  is an MC particles
10    for  $p \in \{1, \dots, N_{MC}\}$  do
11        set  $s_p = t - \Delta t$  #this will be the current time of particle p
12        while  $s_p < t$  and  $w_p > 0$  do
13            if  $x_p \notin \mathcal{D}$  then
14                #here a general function for the application of arbitrary boundary conditions
15                apply_boundary_conditions( $\mathbf{x}_p, s_p, \mathbf{v}_p$ )
16            end
17            sample  $\tau_{inter} = \text{sample\_interaction\_time}(\mathbf{v}_p, i_p)$ 
18            compute  $\tau_{exit} = \text{compute\_cell\_exit\_time}(\mathbf{x}_p, \mathbf{v}_p, i_p)$ 
19            compute  $\tau_{census} = \max(t - \tau, 0)$ 
20            set  $\tau = \min(\tau_{exit}, \tau_{census}, \tau_{inter})$ 
21            #move the particle p
22             $\mathbf{x}_p \leftarrow \mathbf{x}_p - \mathbf{v}_p \tau$ ,
23            #change its weight
24             $(K, r) = \text{compute\_weight\_modif}(\mathbf{v}_p, \tau, \tau_{census}, \tau_{exit}, \tau_{inter}, i_p)$ 
25             $\Delta U(\mathbf{x}_p, t) += \text{compute\_track\_length}(w_p, \mathbf{v}_p, \tau, \tau_{census}, \tau_{exit}, \tau_{inter}, i_p)$ 
26             $w_p \leftarrow K \times w_p$ 
27            if  $\tau == \tau_{census}$  then
28                #set the life time of particle p to zero:
29                 $s_p \leftarrow t$ 
30                #tally the contribution of particle p
31                 $U(\mathbf{x}_p, t) += w_p$ 
32            end
33            if  $\tau == \tau_{exit}$  then
34                #The particle p changes of cell: find its new cell number
35                 $i_p = \text{find_neighbouring\_cell}(i_p, \mathbf{v}_p)$ 
36                #set the life time of particle p to:
37                 $s_p \leftarrow s_p + \tau < t$ 
38            end
39            if  $\tau == \tau_{inter}$  then
40                #Sample the angle and velocity of particle p
41                 $\mathbf{V}' = \text{sample\_velocity}(\mathbf{v}_p, r, i_p)$ 
42                set  $\mathbf{v}_p = \mathbf{V}'$ 
43                #set the life time of particle p to:
44                 $s_p \leftarrow s_p + \tau < t$ 
45            end
46        end
47    end
48     $t \leftarrow t + \Delta t$ 
49 end

```

The instrumentation to compute the track length estimator is in blue in algorithm 15. Expression (10.9) is estimated along the flight path of any MC particle (independently of the event). In algorithm 16, we even present the track length estimator for the semi-analog or the analog MC schemes.

Algorithm 16: The track length estimator depending on the MC scheme

```

1 Function compute_track_length(real wp0, real v, real τmin, real τcensus, real τexit, real τinter, real i)
2   set Δ = 1
3   if MC_scheme == non-analog then
4     Δ = wp0  $\frac{1 - e^{-v\sigma_a^i \eta_0^i \tau_{\min}}}{v\sigma_a^i \eta_0^i \Delta t}$ 
5   end
6   else
7     #It corresponds to the limit σai → 0 of the above estimator
8     Δ = wp0  $\frac{\tau_{\min}}{\Delta t}$ 
9   end
10  return Δ

```

Numerical Analysis of the classical MC scheme for Boltzmann/Bateman

The (split) MC scheme sketched in the previous paragraph is simple but lacks accuracy and can lead to unaffordable constraints on the time step Δt for a stable and accurate resolution. This is illustrated in [3] on various configurations and test-problems. In this paragraph, we perform the numerical analysis of the MC scheme for the coupled system (10.3) to identify the constraining term (in term of accuracy mainly). Of course, we obtain the same conclusion as in [3] but with an analysis-driven discussion rather than a numerical-examples based one. Let us introduce the couple (U^e, η) solution of (10.3) integrated with respect to ω . It is solution of

$$\begin{cases} \partial_t U^e(\mathbf{x}, t) + F^e(\mathbf{x}, t) = -v\sigma_a \eta(\mathbf{x}, t) U^e(\mathbf{x}, t), \\ \partial_t \eta(\mathbf{x}, t) = v\sigma_r \eta(\mathbf{x}, t) U^e(\mathbf{x}, t). \end{cases} \quad (10.10a)$$

$$(10.10b)$$

In the above expression, we defined $F^e(\mathbf{x}, t) = \int \partial_{\mathbf{x}} \mathbf{v} u(\mathbf{x}, t, \omega) d\omega$. The above system is not closed as F^e depends on u . But the analytical quantities U^e, F^e are only auxiliary variables at this stage of the talk, aiming at easing future calculations. The analytical solution of (10.3) integrated over angles can then be formally rewritten

$$U^e(\mathbf{x}, t) = U(\mathbf{x}, t) - \int_0^t F^e(\mathbf{x}, s) e^{-\int_s^t v\sigma_a \eta(\mathbf{x}, \alpha) d\alpha}.$$

In the above expression, $U(\mathbf{x}, t)$ is solution of the homogeneous counterpart of (10.10), i.e. with source term $F^e = 0$. In other words, $(U(\mathbf{x}, t), q(\mathbf{x}, t))$ is solution of

$$\begin{cases} \partial_t U(\mathbf{x}, t) = -v\sigma_a q(\mathbf{x}, t) U(\mathbf{x}, t), \\ \partial_t q(\mathbf{x}, t) = +v\sigma_r q(\mathbf{x}, t) U(\mathbf{x}, t). \end{cases} \quad (10.11a)$$

$$(10.11b)$$

The latter can be solved analytically in this particular case (monokinetic, scalar Bateman) and has solution (see [3]):

$$\begin{aligned} U(\mathbf{x}, t) &= \frac{(\sigma_r U_0(\mathbf{x}) + q_0(\mathbf{x}) \sigma_a) U_0(\mathbf{x})}{\sigma_r U_0(\mathbf{x}) + q_0(\mathbf{x}) \sigma_a \exp(v(\sigma_r U_0(\mathbf{x}) + q_0(\mathbf{x}) \sigma_a) t)}, \\ q(\mathbf{x}, t) &= \frac{(\sigma_r U_0(\mathbf{x}) + q_0(\mathbf{x}) \sigma_a) q_0(\mathbf{x})}{\sigma_r U_0(\mathbf{x}) \exp(-v(\sigma_r U_0(\mathbf{x}) + q_0(\mathbf{x}) \sigma_a) t) + q_0(\mathbf{x}) \sigma_a}. \end{aligned} \quad (10.12)$$

We furthermore introduce one last expression $U^{N.A.}(\mathbf{x}, t) = U_0(\mathbf{x}) e^{-v\sigma_a \eta_0(\mathbf{x}) t}$. The superscript $N.A.$ recalls that this is what is solved along the characteristics for every MC particles (see the definition of the non-analog MC scheme in section 9.4). Quantities $U^{N.A.}(\mathbf{x}, t)$ and $U(\mathbf{x}, t)$ can be numerically

compared: $U^{N.A.}$ is a second order ($\mathcal{O}(\Delta t^2)$) approximation of U :

$$U(\mathbf{x}, t = \Delta t) - U^{N.A.}(\mathbf{x}, t = \Delta t) \underset{\Delta t \sim 0}{=} \frac{1}{2} \sigma_r v^2 q_0(\mathbf{x}) \sigma_a U_0^2(\mathbf{x}) \Delta t^2 + \mathcal{O}(\Delta t^3). \quad (10.13)$$

In other words, we have

$$\begin{aligned} U^e(\mathbf{x}, t) &= U(\mathbf{x}, t) & - \int_0^t F^e(\mathbf{x}, s) e^{- \int_s^t v \sigma_a \eta(\mathbf{x}, \alpha) d\alpha}, \\ &= U^{N.A.}(\mathbf{x}, t) + \mathcal{O}(\Delta t^2) & + U^F(\mathbf{x}, t). \end{aligned} \quad (10.14)$$

Now, introduce ϕ , solution of the linear (with $\eta(\mathbf{x}, t) = \eta_0(\mathbf{x})$) transport equation on time step $[0, t]$. It satisfies

$$\partial_t \phi(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + v \sigma_t \eta_0(\mathbf{x}) \phi(\mathbf{x}, t, \omega) = v \sigma_s \eta_0(\mathbf{x}) \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega'.$$

It is solved with the non-analog MC scheme. As we focus on the time discretisation, we can assume, without loss of generality, an infinitely accurate MC resolution of the above equation just as in [77]. On another hand, the transport equation of solution u can be rewritten as

$$\begin{aligned} \partial_t u(\mathbf{x}, t, \omega) + \mathbf{v} \nabla_{\mathbf{x}} u(\mathbf{x}, t, \omega) + \sigma_t \eta_0(\mathbf{x}) e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} v u(\mathbf{x}, t, \omega) = \\ \sigma_s \eta_0(\mathbf{x}) e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} \int P_s(\omega', \omega) v u(\mathbf{x}, t, \omega') d\omega'. \end{aligned}$$

In the above formulae, we introduced the expression of the analytical nuclide density expressed thanks to our auxiliary variable U^e . Finally introduce $e(\mathbf{x}, t, \omega) = u(\mathbf{x}, t, \omega) - \phi(\mathbf{x}, t, \omega)$. Quantity e is the discrepancy during time step $[0, t]$ between the solution ϕ of the linearized problem and the analytical solution u . The equation satisfied by e is given by

$$\begin{aligned} \partial_t e(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} e(\mathbf{x}, t, \omega) + v \sigma_t \eta_0(\mathbf{x}) (e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} u(\mathbf{x}, t, \omega) - \phi(\mathbf{x}, t, \omega)) = \\ v \sigma_s \eta_0(\mathbf{x}) \left(e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} \int P_s(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega' - \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right). \end{aligned}$$

By rearranging the different terms, we obtain

$$\begin{aligned} \partial_t e(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} e(\mathbf{x}, t, \omega) + v \sigma_t \eta(\mathbf{x}, t) \left(e(\mathbf{x}, t, \omega) + (1 - e^{-v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha}) \phi(\mathbf{x}, t, \omega) \right) = \\ v \sigma_s \eta(\mathbf{x}, t) \left(\int P_s(\omega', \omega) e(\mathbf{x}, t, \omega') d\omega' + (1 - e^{-v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha}) \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right). \end{aligned}$$

By introducing T , the operator implicitly defined by $T(u(\mathbf{x}, t, \omega)) = 0$, we can rewrite the equation satisfied by e as

$$\begin{aligned} T(e(\mathbf{x}, t, \omega)) = \\ (1 - e^{-v \sigma_r (\int_0^t [U^{N.A.}(\mathbf{x}, \alpha) + \mathcal{O}(\Delta t^2) + U^F(\mathbf{x}, \alpha)] d\alpha)}) v \eta(\mathbf{x}, t) \left(-\sigma_t \phi(\mathbf{x}, t, \omega) + \sigma_s \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right). \end{aligned}$$

The above expression allows first recovering some already known results:

- if $\sigma_r = 0$, the MC scheme for the linear Boltzmann equation is unconditionally accurate with respect to the time discretisation. Indeed, in this case $T(e) = 0$ and $u = \phi$ (uniqueness of the solution u for operator T , see [127]) in the limit of an infinity of MC particles.
- We have the same result if $\eta = 0$, i.e. for particles evolving in a vacuum.
- It also holds if $\sigma_t = \sigma_s$ and if the problem is isotropic (i.e. if $\phi = \int \phi$). In this case, we even have an unconditionally accurate approximation with respect to Δt for $\sigma_r \neq 0$.

Of course, the latter remarks are case dependent. More generally, we have

$$T(e(\mathbf{x}, t, \omega)) = (1 - e^{-v \sigma_r (\int_0^t [U^{N.A.}(\mathbf{x}, \alpha) + \mathcal{O}(\Delta t^2) + U^F(\mathbf{x}, \alpha)] d\alpha)}) \Gamma(\mathbf{x}, t, \omega) \text{ with } \Gamma \neq 0.$$

Quantity Γ is given by

$$\Gamma(\mathbf{x}, t, \omega) = v\eta(\mathbf{x}, t) \left(-\sigma_a \phi(\mathbf{x}, t, \omega) - \sigma_s \phi(\mathbf{x}, t, \omega) + \sigma_s \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right). \quad (10.15)$$

Let us now study the effects of the time discretisation scheme. Let us introduce the mean of the flux on time step $[0, t]$

$$\overline{U}^F(\mathbf{x}) = \frac{1}{\Delta t} \int_0^t U^F(\mathbf{x}, \alpha) d\alpha. \quad (10.16)$$

Let us use expression (10.13) to write

$$T(e(\mathbf{x}, t, \omega)) = \left(1 - e^{v\sigma_r U_0(\mathbf{x}) \Delta t \frac{1-e^{-v\sigma_a \eta_0(\mathbf{x}) \Delta t}}{v\sigma_a \eta_0(\mathbf{x}) \Delta t} - \frac{1}{2} v^3 q_0(\mathbf{x}) \sigma_a U_0^2(\mathbf{x}) \sigma_r^2 \Delta t^3 + \mathcal{O}(\Delta t^4) - v\sigma_r \Delta t \overline{U}^F(\mathbf{x})} \right) \Gamma(\mathbf{x}, t, \omega).$$

Thanks to the previous simplifications, we isolated the term where $(\sigma_\alpha)_{\alpha \in \{s, t, a, r\}}$ and Δt compete.

Now assume we are in a regime such that $\sigma_r \sim \frac{1}{\delta^2}$, $\sigma_a \sim \frac{1}{\delta^2}$, $\sigma_s \sim \frac{1}{\delta}$, $v \sim \delta$ and $\overline{U}^F \sim \delta^2$ with $\delta \rightarrow 0$. We will see in the next section it characterises a stiff reactive regime. We then have the two following expressions

$$\begin{aligned} \Gamma_\delta(\mathbf{x}, t, \omega) &\underset{\delta \sim 0}{=} -\frac{1}{\delta} \eta(\mathbf{x}, t) \phi(\mathbf{x}, t, \omega) - \eta(\mathbf{x}, t) \phi(\mathbf{x}, t, \omega) + \eta(\mathbf{x}, t) \int \phi(\mathbf{x}, t, \omega') \omega', \\ \text{and } T(e) &= - \left(1 - e^{\frac{\Delta t}{\delta} U_0 \frac{1-e^{-\eta_0 \frac{\Delta t}{\delta}}}{\eta_0 \frac{\Delta t}{\delta}} - \frac{1}{2} q_0 U_0^2 \frac{\Delta t^3}{\delta^2} + \mathcal{O}(\Delta t^4) - \Delta t \delta} \right) \left(\frac{1}{\delta} \eta \phi + \eta \phi - \eta \int \phi \right). \end{aligned}$$

Assume we have $\frac{\Delta t}{\delta} \ll 1$ (*via* the choice of Δt), we can write

$$T(e(\mathbf{x}, t, \omega)) \underset{\frac{\Delta t}{\delta} \sim 0}{=} -\eta(\mathbf{x}, t) \phi(\mathbf{x}, t, \omega) \left[U_0(\mathbf{x}) \frac{\Delta t}{\delta} + \frac{1}{2} v^3 q_0(\mathbf{x}) U_0^2(\mathbf{x}) \frac{\Delta t^3}{\delta^3} - \Delta t \right]. \quad (10.17)$$

Suppose we want an approximation ϕ of u ensuring $T(e) = \mathcal{O}(\Delta t)$, it demands $\Delta t = \delta^2$. Such condition is numerically intensive in many physical applications, see in neutronics for example [97, 98, 148, 95, 96], [3]. Finally, let us rewrite (10.17) as $T(e(\mathbf{x}, t, \omega)) = -\phi(\mathbf{x}, t, \omega) K_\delta(\mathbf{x}, t, \omega, \Delta t) \Delta t$, then

$$K_\delta(\mathbf{x}, t, \omega, \Delta t) = \left[\frac{1}{\delta} U_0(\mathbf{x}) - 1 + \frac{1}{2} q_0(\mathbf{x}) U_0^2(\mathbf{x}) \frac{\Delta t^2}{\delta^3} \right]. \quad (10.18)$$

The coefficient K_δ does not satisfy the conditions of definition 9.1 regarding AP schemes as $K_\delta \xrightarrow[\delta \rightarrow 0]{} \infty$. Consequently, the above MC scheme is not AP in the reactive regime.

10.1.2 An Asymptotic Preserving MC scheme for neutron transport

The Asymptotic Preserving MC scheme we present in this paper has already been presented in [3] in a more general context (non-monokinetic, non-scalar Bateman case). The material of this section remains complementary to [3] in the sense the new MC scheme has been described *via* an illustration-driven analysis in [3] whereas we here perform the numerical analysis (in a simple case). We here compare the convergence rates of the classical and the AP schemes.

Asymptotic regime of interest for system (10.3)

We suggest identifying more precisely the stiff regime of (10.3) for which the previous MC scheme fails to produce accurate (and even stable see [3]) solutions, see [97, 98, 148, 95, 96], [3]. It is helpful to

non-dimensionalize the coupled system (10.3). Let us introduce

$$\begin{cases} \mathbf{x} = \mathbf{x}^* \mathcal{X}, v = v^* \mathcal{V}, t = t^* \mathcal{T}, \\ \sigma_\alpha = \sigma_\alpha^* \frac{1}{\lambda_\alpha}, \forall \alpha \in \{s, t, a, r\}. \end{cases} \quad (10.19)$$

The upperscript * denotes a nondimensional quantity. Let us introduce $u^*(\mathbf{x}^*, t^*, \omega) = u(\mathbf{x}, t, \omega)$, then

$$\begin{aligned} \frac{1}{T} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) &= \partial_t u(\mathbf{x}, t, \omega), & \frac{1}{\mathcal{X}} \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) &= \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega), \\ \frac{1}{T} \partial_{t^*} \eta^*(\mathbf{x}^*, t^*) &= \partial_t \eta(\mathbf{x}, t). \end{aligned}$$

Using the above expressions in the nonlinear Boltzmann equation coupled to Bateman one (10.3) yields

$$\begin{cases} \frac{\mathcal{X}}{T} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) + \mathcal{V} v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_a^* \eta^*(\mathbf{x}^*, t^*) \frac{\mathcal{V} \mathcal{X}}{\lambda_a} u^*(\mathbf{x}^*, t^*, \omega) = \\ v^* \sigma_s^* \eta^*(\mathbf{x}^*, t^*) \frac{\mathcal{V} \mathcal{X}}{\lambda_s} \int u^*(\mathbf{x}^*, t^*, \omega) d\omega. \\ \frac{1}{T} \partial_{t^*} \eta^*(\mathbf{x}^*, t^*) = v^* \frac{\mathcal{V} \sigma_r^*}{\lambda_r} \eta^*(\mathbf{x}^*, t^*) \int u^*(\mathbf{x}^*, t^*, \omega) d\omega. \end{cases}$$

Let us decompose $\sigma_t = \sigma_a + \sigma_s$ to obtain

$$\begin{cases} \frac{\mathcal{X}}{T} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) + \mathcal{V} v^* \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + v^* \sigma_a^* \eta^*(\mathbf{x}^*, t^*) \frac{\mathcal{V} \mathcal{X}}{\lambda_a} u^*(\mathbf{x}^*, t^*, \omega) \\ + v^* \sigma_s^* \eta^*(\mathbf{x}^*, t^*) \frac{\mathcal{V} \mathcal{X}}{\lambda_s} u^*(\mathbf{x}^*, t^*, \omega) = v^* \sigma_s^* \eta^*(\mathbf{x}^*, t^*) \frac{\mathcal{V} \mathcal{X}}{\lambda_s} \int u^*(\mathbf{x}^*, t^*, \omega) d\omega, \\ \frac{\mathcal{X}}{T} \partial_{t^*} \eta^*(\mathbf{x}^*, t^*) = v^* \sigma_r^* \frac{\mathcal{V} \mathcal{X}}{\lambda_r} \eta^*(\mathbf{x}^*, t^*) \int u^*(\mathbf{x}^*, t^*, \omega) d\omega. \end{cases}$$

Now suppose $\frac{\mathcal{X}}{T} = \mathcal{O}(\frac{1}{\delta}) = \frac{\mathcal{V} \mathcal{X}}{\lambda_a} = \frac{\mathcal{V} \mathcal{X}}{\lambda_r}$, $\mathcal{V} = \mathcal{O}(\delta)$ and $\frac{\mathcal{V} \mathcal{X}}{\lambda_s} = \mathcal{O}(\delta)$ ensuring $\mathcal{O}(\frac{1}{\delta^2}) = \frac{\mathcal{X}}{\lambda_a} = \frac{\mathcal{X}}{\lambda_r}$ and $\frac{\mathcal{X}}{\lambda_s} = \mathcal{O}(1)$, we have (we drop the upperscript for convenience)

$$\begin{cases} \frac{1}{\delta} \partial_t u(\mathbf{x}, t, \omega) + \delta \mathbf{v} \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) \\ + v \sigma_a \eta(\mathbf{x}, t) \frac{1}{\delta} u(\mathbf{x}, t, \omega) \\ + \delta v \sigma_s \eta(\mathbf{x}, t) u(\mathbf{x}, t, \omega) = \delta v \sigma_s \eta(\mathbf{x}, t) \int u(\mathbf{x}, t, \omega) d\omega. \\ \frac{1}{\delta} \partial_t \eta(\mathbf{x}, t) = v \frac{\sigma_r}{\delta} \eta(\mathbf{x}, t) \int u(\mathbf{x}, t, \omega) d\omega. \end{cases}$$

Performing a Hilbert development, i.e. $u = u^0 + u^1 \delta + u^2 \delta^2 + \mathcal{O}(\delta^3)$ see [143], and considering only the first order (i.e. u^0) leads to

$$\begin{cases} \partial_t u^0(\mathbf{x}, t) = -v \sigma_a \eta^0(\mathbf{x}, t) u^0(\mathbf{x}, t), \\ \partial_t \eta^0(\mathbf{x}, t) = +v \sigma_r \eta^0(\mathbf{x}, t) u^0(\mathbf{x}, t). \end{cases} \quad (10.20)$$

It corresponds to the monokinetic homogeneous regime as $\delta \rightarrow 0$. The second order yields

$$\mathbf{v} \partial_{\mathbf{x}} u^0(\mathbf{x}, t, \omega) + v \sigma_s \eta^0(\mathbf{x}, t) u^0(\mathbf{x}, t, \omega) = v \sigma_s \eta^0(\mathbf{x}, t) \int u^0(\mathbf{x}, t, \omega') d\omega'.$$

It implies $F^e(\mathbf{x}, t) = \partial_{\mathbf{x}} \int \mathbf{v} u^0(\mathbf{x}, t, \omega) d\omega = 0 = \mathcal{O}(\delta^2)$. As a consequence, in the stiff regime of interest, $\bar{U}^F(\mathbf{x})$ related to F^e through (10.14) and defined in (10.16) is also $\bar{U}^F(\mathbf{x}) = \mathcal{O}(\delta^2)$. In the next section, we present a general methodology, already detailed in the general case in [3], to build an Asymptotic Preserving MC scheme for regime (10.20).

Construction of an Asymptotic Preserving MC scheme for regime (10.20)

We recall the system we aim at solving is given by

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + \mathbf{v} \nabla_{\mathbf{x}} u(\mathbf{x}, t, \omega) + \sigma_t \eta(\mathbf{x}, t) v u(\mathbf{x}, t, \omega) = \sigma_s \eta(\mathbf{x}, t) \int P_s(\omega', \omega) v u(\mathbf{x}, t, \omega') d\omega' \\ \partial_t \eta(\mathbf{x}, t) = \eta(\mathbf{x}, t) \sigma_r v \int u(\mathbf{x}, t, \omega) d\omega. \end{cases} \quad (10.21)$$

We need an MC scheme able to capture efficiently the asymptotic regime defined by:

$$\begin{cases} \partial_t U(\mathbf{x}, t) = -v \sigma_a q(\mathbf{x}, t) U(\mathbf{x}, t), \\ \partial_t q(\mathbf{x}, t) = +v \sigma_r q(\mathbf{x}, t) U(\mathbf{x}, t). \end{cases} \quad (10.22)$$

The system of ODE (10.22) can be solved analytically in this configuration (monokinetic, scalar Bateman) and its solution is given by (10.12), see [3]. To construct the Asymptotic Preserving MC scheme, we suggest linearizing (10.21) on time step $[0, t]$ substituting q , solution of (10.22), to η in (10.21). This leads to

$$\begin{cases} \partial_t \phi(\mathbf{x}, t, \omega) + \mathbf{v} \nabla_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + \sigma_t q(\mathbf{x}, t) v \phi(\mathbf{x}, t, \omega) = \sigma_s q(\mathbf{x}, t) \int P_s(\omega', \omega) v \phi(\mathbf{x}, t, \omega') d\omega' \\ \partial_t \eta(\mathbf{x}, t) = q(\mathbf{x}, t) \sigma_r v \int \phi(\mathbf{x}, t, \omega) d\omega. \end{cases} \quad (10.23)$$

In (10.23), the (now linear) transport equation (10.23a) is *self consistent* and has a form which has already been intensively encountered (time dependent cross-sections) all along the document. The time evolution of the nuclide density η only depends on ϕ , solution of the linearized transport equation, q being supposedly known. Consistent tallies during the MC resolution will ensure its update. The linearized transport equation can then be rewritten as a expectation

$$\phi(\mathbf{x}, t, \omega) = \mathbb{E} \left[\begin{array}{ccc} +\mathbf{1}_{[t, \infty]}(\tau) & e^{-v\sigma_a \int_0^t q(\mathbf{x} - \mathbf{v}\alpha, \alpha) d\alpha} & \phi_0(\mathbf{x} - \mathbf{v}t, \omega) \\ +\mathbf{1}_{[0, t]}(\tau) & e^{-v\sigma_a \int_{t-\tau}^t q(\mathbf{x} - \mathbf{v}\alpha, \alpha) d\alpha} & \phi(\mathbf{x} - \mathbf{v}\tau, t - \tau, W') & P_s(W', \omega) \end{array} \right],$$

over the following set of random variables

$$\begin{cases} \tau \sim \mathcal{E}_t(v\sigma_s), \\ W' \sim P_s(W', \omega). \end{cases}$$

The exponential law in the above expression depends on time. The sampling of τ is made according to

$$f_\tau(\mathbf{x}, t, \mathbf{v}, s) ds = \mathbf{1}_{[0, \infty]}(s) v \sigma_s q(\mathbf{x} - \mathbf{v}s, t - s) e^{-\int_0^s v \sigma_s q(\mathbf{x} - \mathbf{v}\alpha, t - \alpha, \mathbf{v}) d\alpha} ds.$$

It is obtained from plugging the time evolution of q in expression (9.27). The weight modification⁵ of any MC particles is done according to

$$w_p(t) = w_p(0) e^{-v\sigma_a \int_0^t q(\mathbf{x} - \mathbf{v}\alpha, \alpha) d\alpha} = w_p(0) \frac{U(\mathbf{x}, t)}{U_0(\mathbf{x})}.$$

The consistent tally to update the nuclide density can be obtained in a same way as before. First by integrating (10.23b) on time step $[0, t]$. Second by plugging the expression of q in (10.23b) together with

⁵To obtain the previous expressions for the probability measure of the time interaction or the weight modification, we performed similar calculations as the ones presented in the previous chapter 9. They are only particular cases. In order to avoid redundancies, they are not repeated here.

the MC discretisation. It leads to

$$\begin{aligned}
\eta(\mathbf{x}, t) - \eta_0(\mathbf{x}) &= \int_0^t \iint v \sigma_r q(\mathbf{x}, s) u(\mathbf{x}, s, \omega) d\omega ds, \\
&= \sum_{p=1}^{N_{MC}} \sum_{i=1}^{N_p} \int_{t_i}^{t_{i+1}} v \sigma_r q(\mathbf{x}, s) w_p(s) \delta_{\mathbf{x}}(\mathbf{x}_p(s)) \delta_{\omega}(\omega_p(s)) ds, \\
&= \sum_{p=1}^{N_{MC}} \sum_{i=1}^{N_p} \int_{t_i}^{t_{i+1}} v \sigma_r q(\mathbf{x}, s) w_p(t_i) \frac{U(\mathbf{x}_p(s), s)}{U_0(\mathbf{x}_p(t_i))} \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds, \\
&= \sum_{p=1}^{N_{MC}} \sum_{i=1}^{N_p} v \sigma_r w_p(t_i) \frac{1}{U_0(\mathbf{x}_p(t_i))} \int_{t_i}^{t_{i+1}} q(\mathbf{x}, s) U(\mathbf{x}_p(s), s) \delta_{\mathbf{x}}(\mathbf{x}_p(s)) ds, \\
&= \sum_{p=1}^{N_{MC}} \sum_{i=1}^{N_p} \frac{\sigma_r}{\sigma_a} w_p(t_i) \left(1 - \frac{U(\mathbf{x}_p(t_{i+1}), t_{i+1})}{U(\mathbf{x}_p(t_i), t_i)} \right).
\end{aligned}$$

It only remains to instrument the new tracking with the consistent track length estimator. It is given by

$$\Delta u_p^i = \frac{\sigma_r}{\sigma_a} w_p(t_i) \left(1 - \frac{U(\mathbf{x}_p(t_{i+1}), t_{i+1})}{U(\mathbf{x}_p(t_i), t_i)} \right). \quad (10.24)$$

It allows closing the time step.

Note that in this section, we recovered the same MC scheme as in [3] in a slightly different way. In [3], we perform a change of variable $u(\mathbf{x}, t, \omega) = U(\mathbf{x}, t)f(\mathbf{x}, t, \omega)$ and identify the transport equation satisfied by f as in [3] or in section 9.9.2 (source term). Here, we invoked the plugging of $q \approx \eta$ on time step $[0, t]$. The advantage of presenting the scheme as in [3] comes from the fact it is easier emphasizing the Asymptotic Preserving MC scheme can be resumed to a balanced gain-loss transport equation on f in the transport phase. In other words, we have along each characteristics, see [3],

$$\partial_s \int f(\mathbf{x} + v\omega s, s, \omega) d\omega = 0, \quad \forall s \in [0, t].$$

It also eases its comparison with Quasi-Static (QS) methods, see sections 9.9.2 or 9.12.1. We refer to [3] for more details on this point and insist that on the simplified system (10.3), the two derivations lead to the same MC scheme. In the general case (non-monokinetic non-scalar Bateman), the two different ways to introduce the linearisation lead to two slightly different MC discretisation schemes. Still, both bear the asymptotic preserving property presented in the next section.

The previous expressions, for the sampling of the time interaction, the weight modification and the tally were general. To end the description of the Asymptotic Preserving MC scheme, let us introduce a cell discretisation as described in section 9.6. Let us assume a constant per cell discretisation for the particle and nuclide densities of the reduced model. We have $\forall \mathbf{x} \in \mathcal{D}, s \in [0, t]$:

$$\begin{aligned}
U(\mathbf{x}, s) &= \sum_{i=1}^{N_{\mathbf{x}}} U_i(s) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}), \\
q(\mathbf{x}, s) &= \sum_{i=1}^{N_{\mathbf{x}}} q_i(s) \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}).
\end{aligned}$$

With this previous hypothesis, the inversion of the cdf of the probability measure for the sampling of the interaction time can be done analytically. We can explicitly exhibit the sampling of the time interaction

τ for an MC particle p in cell i :

$$\tau = t - \frac{1}{v(\sigma_a^i q_0^i + U_0^i \sigma_r^i)} \ln \left(\frac{(q_0^i \sigma_a^i \exp(vt(\sigma_a^i q_0^i + U_0^i \sigma_r^i)) + \sigma_r^i U_0^i) \exp\left(\frac{\sigma_a^i}{\sigma_s^i} \ln(\mathcal{U}_\tau)\right) - \sigma_r^i U_0^i}{q_0^i \sigma_a^i} \right). \quad (10.25)$$

Its expression takes into account the nuclide density variation along the flight path of the treated MC particle. We can verify that at the first order with respect to $\sigma_r^i \sim 0$ we have $\tau = -\frac{1}{v\sigma_s^i q_0^i} \ln(\mathcal{U}_\tau)$. It corresponds to the sampling of the interaction time for the classical scheme (non stiff regime), see section 9.6. We can even put forward the second order development with respect to σ_r^i :

$$\tau_{\sigma_r^i \sim 0} = -\frac{\ln(\mathcal{U}_\tau)}{v\sigma_s^i q_0^i} + \left(\frac{U_0^i t}{\sigma_a^i q_0^i} + \frac{U_0^i (1 - \mathcal{U}_\tau)^{\frac{\sigma_a^i}{\sigma_s^i}}}{v(\sigma_a^i q_0^i)^2 \exp(v\sigma_a^i q_0^i t)} - \frac{U_0^i \left(v\sigma_a^i q_0^i t - \frac{\ln(\mathcal{U}_\tau) \sigma_a^i}{\sigma_s^i} \right)}{v(\sigma_a^i q_0^i)^2} \right) \sigma_r^i + \mathcal{O}((\sigma_r^i)^2). \quad (10.26)$$

Higher order approximations can be obtained in the same manner. The explicit expression for the weight modification for an MC particle p remaining in cell i between times 0 and t is given by

$$w_p(t) = w_p(0) \frac{U^i(t)}{U_0^i} = \frac{(\sigma_r^i U_0^i + q_0^i \sigma_a^i)}{\sigma_r^i U_0^i + q_0^i \sigma_a^i \exp(v(\sigma_r^i U_0^i + q_0^i \sigma_a^i)t)}. \quad (10.27)$$

Quantity $U^i(t)$ is solution of (10.22) in cell i . We can once again verify that asymptotically with $\sigma_r^i \sim 0$, we recover the classical weight modification along the flight path of each MC particles

$$U^i(t)_{\sigma_r^i \sim 0} = U_0^i e^{-v\sigma_a^i q_0^i t} - \sigma_r^i (U_0^i)^2 e^{-v\sigma_a^i q_0^i t} \frac{-1 + e^{-v\sigma_a^i q_0^i t} + v\sigma_a^i q_0^i t}{\sigma_a^i q_0^i} + \mathcal{O}(\sigma_r^2). \quad (10.28)$$

Of course, the same applies to (10.24), the tally to update the nuclide densities.

In practice, the described Asymptotic Preserving MC scheme can be developed in the same general canvas as in section 9.8 provided an additional line to compute the track length estimator as in algorithm 15. For this, it is enough modifying

- function sample_interaction_time using expression (10.25),
- function compute_weight_modif using expression (10.27),
- function compute_track_length using expression (10.24).

The other functions, in a more general case (non-monokinetic for example), must be consistently be developed and we refer to [3] for their detailed descriptions. In this document, we now focus on the numerical analysis of the newly built MC scheme.

Numerical Analysis of the Asymptotic Preserving MC scheme for Boltzmann/Bateman

Let us now perform the numerical analysis of the newly built Asymptotic Preserving MC scheme and compare the results with the one of the previous classical one (mainly with expression (10.17)). The analysis is here pretty fast thanks to the already introduced material. On time step $[0, t]$, the previous process supposes solving the following linearized transport equation

$$\begin{aligned} \partial_t \phi(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + v\sigma_t & \underbrace{q(\mathbf{x}, t)}_{\eta_0(\mathbf{x}) e^{v\sigma_r \int_0^t U(\mathbf{x}, \alpha) d\alpha}} \phi(\mathbf{x}, t, \omega) = \\ & v\sigma_s \underbrace{q(\mathbf{x}, t)}_{\eta_0(\mathbf{x}) e^{v\sigma_r \int_0^t U(\mathbf{x}, \alpha) d\alpha}} \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega'. \end{aligned}$$

In the above expression, η has been substituted by q . It is expressed with respect to U , solution of the reduced model (10.22) preserving the stiff regime of interest. Once again, let us introduce $e = u - \phi$ and subtract the two transport equations to get

$$\partial_t e(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} e(\mathbf{x}, t, \omega) + v \sigma_t \eta_0(\mathbf{x}) (e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} u(\mathbf{x}, t, \omega) - e^{v \sigma_r \int_0^t U(\mathbf{x}, \alpha) d\alpha} \phi(\mathbf{x}, t, \omega)) = \\ v \sigma_s \eta_0(\mathbf{x}) \left(e^{v \sigma_r \int_0^t U^e(\mathbf{x}, \alpha) d\alpha} \int P_s(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega' - e^{v \sigma_r \int_0^t U(\mathbf{x}, \alpha) d\alpha} \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right).$$

Let us use the fact that $U = U^e - U^F$ in the above expression to rewrite

$$\partial_t e(\mathbf{x}, t, \omega) + \mathbf{v} \partial_{\mathbf{x}} e(\mathbf{x}, t, \omega) + v \sigma_t \eta(\mathbf{x}, t) \left(u(\mathbf{x}, t, \omega) - e^{-v \sigma_r \int_0^t U^F(\mathbf{x}, \alpha) d\alpha} \phi(\mathbf{x}, t, \omega) \right) = \\ v \sigma_s \eta(\mathbf{x}, t) \left(\int P_s(\omega', \omega) u(\mathbf{x}, t, \omega') d\omega' - e^{-v \sigma_r \int_0^t U^F(\mathbf{x}, \alpha) d\alpha} \int P_s(\omega', \omega) \phi(\mathbf{x}, t, \omega') d\omega' \right).$$

We obtain

$$\begin{aligned} T(e(\mathbf{x}, t, \omega)) &= \Gamma(\mathbf{x}, t, \omega) (1 - e^{-v \sigma_r \Delta t \bar{U}^F(\mathbf{x})}), \\ &\stackrel{\substack{v \sigma_r \sim \frac{1}{\delta} \\ \bar{U}^F \sim \delta^2}}{=} \left[-\frac{1}{\delta} \eta(\mathbf{x}, t) \phi(\mathbf{x}, t, \omega) \right] (1 - e^{-\frac{1}{\delta} \Delta t \delta^2}), \\ &\stackrel{\Delta t \rightarrow 0}{=} -\eta(\mathbf{x}, t) \phi(\mathbf{x}, t, \omega) \Delta t = \mathcal{O}(\Delta t) = K_\delta \Delta t = K \Delta t. \end{aligned} \quad (10.29)$$

It is first order in Δt independently of $\delta \rightarrow 0$ (as $K_\delta = K$). Expression (10.29) must be compared to (10.17). With the new Asymptotic Preserving MC scheme, the dependence of $K_\delta(\mathbf{x}, t, \omega) = -\eta(\mathbf{x}, t) \phi(\mathbf{x}, t)$ with respect to δ is considerably weakened and the new MC scheme satisfies the conditions of definition 9.1. The new MC scheme is AP in the reactive regime, bigger time steps can be used and computational gains are at stakes. We rely on [3] for numerical examples.

10.1.3 Summary

To end this section, we would like to come back on some aspects of the presented Asymptotic Preserving MC scheme:

1. first, the new MC scheme is more *intrusive* than the classical one which can be applied calling successively two simulation codes. Nevertheless, we rely on the comparison of (10.17) and (10.29) to motivate and convince the reader willing to solve (10.1) in the stiff regime (10.22).
2. The generalization of the Asymptotic Preserving MC scheme to system (10.1) (instead of (10.3)) has been presented in [3]. In particular in [3] we identify the stiff regime of interest in the non-monokinetic non-scalar Bateman case. The numerical analysis performed in this document still holds for such more general reduced model. The computations are only more complex. In fact, the real challenge in this generalized context consists in the efficient resolution of the reduced model solving the stiffness along the flight path of each MC particles, see [3].
3. Regarding the forementioned real challenge, for the simplified configuration considered in the previous paragraphs, an analytical solution for the stiff regime (10.22) is available. One cannot do better in term of efficiency (see [3]). In fact, with this simplified configuration, we presented what can asymptotically be expected with a cheap and efficient resolution of the reduced model. In [3], less ideal cases were also tackled and for them, we introduced a numerical solver to estimate *on-the-fly* the reduced model. In some particular configurations, it is not obvious whether the Asymptotic Preserving scheme is still more efficient or not⁶.

Now, with expressions (10.26) and (10.28) and the development they introduce with respect to σ_r , we wanted to highlight it is possible to define *high order MC scheme*. For MC schemes, the relevant quantities on which one must increase the order are typically the time interaction, the weight modification etc. (see for example (10.26) and (10.28)). Relying on such high-order developments may lead to even less costly approximations.

⁶For some configurations, for the same accuracy, we have similar restitution times between the classical and the AP scheme, see [3]. In fact, those cases are closer to $\delta = \mathcal{O}(1)$ than $\delta \ll 1$ as assumed in this section.

4. To finish, we insist on the fact the methodology to build an Asymptotic Preserving MC scheme is general and can be summed up in few phases:

- first, identify the stiff regime of interest.
- Build a relevant reduced model and approach it (efficiently).
- Plug the solution of the reduced model in the transport equation to linearize it. This can be done *via* a change of variable as in [3] or by plugging it directly in the cross-sections' expressions as in this document.
- Solve a *linearized Boltzmann equation with time dependent cross-sections* with an MC scheme (with similar linearisation hypothesis as for Quasi-Static methods, see sections 9.9.2–10.1 and [3]).
- Consistently update the quantities which are not discretised with MC particles and end the time step.

Section 9.9.2 is typically an example of application of such methodology for problems with stiff sources. Another example is given in remark 10.3 of the next section.

In the following section, we study a different nonlinear Boltzmann equation. It implies a different stiff regime and different relevant linearisations for the construction an Asymptotic Preserving MC scheme. Care will be taken to hint at point 4.) above to emphasize the methodology applied in this section also does for different physical applications.

10.2 Boltzmann equation coupled to Stefan's law (photonics)

In this section, we are interested in the resolution of the time-dependent problem of particle transport in a media whose density of internal energy evolves with time due to particle interactions (Stefan's law). We suppose transport to be driven by the linear Boltzmann equation (10.30) for particles having position $\mathbf{x} \in \mathcal{D} \subset \mathbb{R}^3$, velocity c (speed of light), frequency $\nu \in \mathbb{R}^+$, direction $\omega \in \mathbb{S}^2 = [0, 2\pi] \times [0, \pi]$, at time $t \in [0, T] \subset \mathbb{R}^+$. Quantity $u(\mathbf{x}, t, \nu, \omega)$ is the density of energy of the particles at $(\mathbf{x}, t, \nu, \omega)$. We assume the time variation of the media's density of internal energy $E(\mathbf{x}, t)$ can be accurately modeled by Stefan's law (10.30b), see [245, 203, 59]. As a result, problem (10.30) is nonlinear and strongly coupled:

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \nu, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \nu, \omega) + c\sigma_t(\mathbf{x}, t, \nu)u(\mathbf{x}, t, \nu, \omega) \\ \quad = \iint c\sigma_s(\mathbf{x}, t, \nu', \nu, \omega', \omega)u(\mathbf{x}, t, \nu', \omega')d\omega'd\nu' + c\sigma_a(\mathbf{x}, t, \nu)B(\mathbf{x}, t, \nu), \end{array} \right. \quad (10.30a)$$

$$\partial_t E(\mathbf{x}, t) = \iint c\sigma_a(\mathbf{x}, t, \nu)(u(\mathbf{x}, t, \nu, \omega) - B(\mathbf{x}, t, \nu))d\omega d\nu. \quad (10.30b)$$

To close system (10.30), we need to define the dependences of $\sigma_s, \sigma_t, \sigma_a, B$ with respect to the unknowns u and E . First, the source term B corresponds to the Planckian distribution defined as

$$B(\mathbf{x}, t, \nu) = B(T(\mathbf{x}, t), \nu) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{hc}{kT(\mathbf{x}, t)}} - 1}. \quad (10.31)$$

It describes the spectral density of electromagnetic radiation emitted by a black body in thermal equilibrium at a given temperature $T(\mathbf{x}, t)$. The constant k and h are respectively the Boltzmann and the Planck constants. Let us integrate the above expression over frequency and angles, we have

$$\int B(\mathbf{x}, t, \nu)d\nu = aT^4(\mathbf{x}, t) = \frac{8k^4\pi^5}{15c^3h^3}T^4(\mathbf{x}, t). \quad (10.32)$$

In the above expression, a is the radiative constant. Second, as detailed in chapter 9, we introduce

$$\sigma_s(\mathbf{x}, t, \nu) = \iint \sigma_s(\mathbf{x}, t, \nu', \nu, \omega', \omega)d\omega'd\nu'.$$

The opacities σ_s , σ_t and σ_a are related *via* the simple relation $\sigma_a(\mathbf{x}, t, \nu) = \sigma_t(\mathbf{x}, t, \nu) - \sigma_s(\mathbf{x}, t, \nu)$. For the same reason as in chapter 9, the dependence with respect to ω of $\sigma_s(\mathbf{x}, t, \nu)$ is not recalled. The cross-sections are commonly called *opacities* in such physical context and depend on \mathbf{x}, t *via* the temperature $T(\mathbf{x}, t)$, i.e.

$$\sigma_\alpha(\mathbf{x}, t, \cdot) = \sigma_\alpha(T(\mathbf{x}, t), \cdot), \forall \alpha \in \{s, t, a\}.$$

The density of internal energy $E(\mathbf{x}, t)$ also depends on $T(\mathbf{x}, t)$ *via* an equation of state. For a perfect gas, this is the well-known relation $E(\mathbf{x}, t) = C_v T(\mathbf{x}, t)$ where C_v is the adiabatic coefficient of the considered material. System (10.30) can consequently be rewritten in a closed form *via* the introduction of the temperature

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \nu, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \nu, \omega) + c\sigma_t(T(\mathbf{x}, t), \nu)u(\mathbf{x}, t, \nu, \omega) \\ = \iint c\sigma_s(T(\mathbf{x}, t), \nu', \nu, \omega', \omega)u(\mathbf{x}, t, \nu', \omega') d\omega' d\nu' + c\sigma_a(T(\mathbf{x}, t), \nu)B(T(\mathbf{x}, t), \nu), \end{array} \right. \quad (10.33a)$$

$$\left. \begin{array}{l} \partial_t E(T(\mathbf{x}, t)) = \int c\sigma_a(T(\mathbf{x}, t), \nu) \left(\int u(\mathbf{x}, t, \nu, \omega) d\omega - B(T(\mathbf{x}, t), \nu) \right) d\nu. \end{array} \right. \quad (10.33b)$$

The next step consists in simplifying (10.33) to focus on the difficult numerical aspects. Of course, care will be taken to make sure the Asymptotic Preserving MC schemes we suggest in the following sections can be extended to the complete problem (10.33).

To simplify problem (10.33), we first assume $\sigma_t = \sigma_a = \sigma$. In other words, there is no (physical) scattering. Besides, we aim at focusing on the nonlinearity induced by the source term B rather than by the opacities or the equation of state. We consequently assume that $\sigma(T(\mathbf{x}, t), \nu) = \sigma(\mathbf{x}, \nu)$ and $E(T) = C_v T$. Furthermore, we consider the *grey approximation*, i.e. the opacities do not depend on ν . It ensures we can solve problem (10.33) with respect to $u(\mathbf{x}, t, \omega) = \int u(\mathbf{x}, t, \nu, \omega) d\nu$, solution of (10.33) integrated over every frequencies:

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})aT^4(\mathbf{x}, t), \end{array} \right. \quad (10.34a)$$

$$\left. \begin{array}{l} \partial_t E(T(\mathbf{x}, t)) = c\sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - aT^4(\mathbf{x}, t) \right). \end{array} \right. \quad (10.34b)$$

To simplify the notations and focus on the nonlinearity introduced by the source term, we even go one step beyond and set $a = C_v = 1$ so that (10.34) can be rewritten in a simplified closed form

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})E^4(\mathbf{x}, t), \end{array} \right. \quad (10.35a)$$

$$\left. \begin{array}{l} \partial_t E(\mathbf{x}, t) = c\sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - E^4(\mathbf{x}, t) \right). \end{array} \right. \quad (10.35b)$$

The asymptotic regime we aim at capturing with (10.35) has already been intensively studied in the literature, see [245, 203, 59]. To characterise it, let us introduce

$$\left\{ \begin{array}{l} \mathbf{x} = \mathbf{x}^* \mathcal{X}, t = t^* \mathcal{T}, \\ \sigma = \sigma^* \frac{1}{\lambda}. \end{array} \right. \quad (10.36)$$

The superscript * denotes a nondimensional quantity. Let us denote by $u^*(\mathbf{x}^*, t^*, \omega) = u(\mathbf{x}, t, \omega)$ and $E^*(\mathbf{x}^*, t^*) = E(\mathbf{x}, t)$, then

$$\begin{aligned} \frac{1}{\mathcal{T}} \partial_{t^*} u^*(\mathbf{x}^*, t^*, \omega) &= \partial_t u(\mathbf{x}, t, \omega), & \frac{1}{\mathcal{X}} \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) &= \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega), \\ \frac{1}{\mathcal{T}} \partial_{t^*} E^*(\mathbf{x}^*, t^*) &= \partial_t E(\mathbf{x}, t). \end{aligned}$$

This leads to

$$\begin{cases} \partial_{t^*} u^*(\mathbf{x}^*, t, \omega) + c \frac{\mathcal{T}}{\chi} \omega \partial_{\mathbf{x}^*} u^*(\mathbf{x}^*, t^*, \omega) + \frac{\mathcal{T}}{\lambda} c \sigma^*(\mathbf{x}^*) u^*(\mathbf{x}^*, t^*, \omega) = c \sigma^*(\mathbf{x}^*) (E^*)^4(\mathbf{x}^*, t^*), \\ \partial_{t^*} E^*(\mathbf{x}^*, t^*) = \frac{\mathcal{T}}{\lambda} c \sigma^*(\mathbf{x}^*) \left(\int u^*(\mathbf{x}^*, t^*, \omega) d\omega - (E^*)^4(\mathbf{x}^*, t^*) \right). \end{cases} \quad (10.37a)$$

Now suppose $\frac{c\mathcal{T}}{\chi} = \mathcal{O}(\frac{1}{\delta})$ and $\frac{\mathcal{T}}{\lambda} c \sigma^* = \frac{1}{\delta^2} \sigma^*$, we have (we drop the superscript * for convenience)

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + \frac{1}{\delta} \omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) = \frac{1}{\delta^2} \sigma(\mathbf{x}) (E^4(\mathbf{x}, t) - u(\mathbf{x}, t, \omega)), \\ \partial_t E(\mathbf{x}, t) = \frac{1}{\delta^2} \sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - E^4(\mathbf{x}, t) \right). \end{cases} \quad (10.38)$$

The above non-dimensionalization (10.38) from (10.34) can be found in many books and publications [245, 203, 59, 77, 301, 48, 178]. All along this section, we work on the closed forms (10.35) and (10.38) of unknowns (u, E) . System (10.35), independently of the value of δ , conserves the total density of energy of the system 'particles+media' as

$$\partial_t \left(\int u(\mathbf{x}, t, \omega) d\omega + E(\mathbf{x}, t) \right) + \partial_{\mathbf{x}} \int c \omega u(\mathbf{x}, t, \omega) d\omega = 0.$$

This invariant is important in many applications and one may want the MC scheme to ensure it numerically, independently of the discretisation parameters ($\Delta \mathbf{x}$, Δt or N_{MC}). Let us study the limit regime $\delta \rightarrow 0$. For this, we perform a Hilbert development [143, 48, 178], i.e. $u = u^0 + u^1 \delta + u^2 \delta^2 + \mathcal{O}(\delta^3)$ and $E = E^0 + E^1 \delta + E^2 \delta^2 + \mathcal{O}(\delta^3)$, and identify the stiff asymptotic regime of interest for (10.35). Plugging the previous development in (10.38), and more precisely in the transport counterpart (10.35a), leads to

$$\partial_t \begin{pmatrix} 0 \\ 0 \\ u_0 \delta^2 \\ \sum_{i=1}^{\infty} u_i \delta^{i+2} \end{pmatrix} + \omega \partial_{\mathbf{x}} \begin{pmatrix} 0 \\ u_0 \delta \\ u_1 \delta^2 \\ \sum_{i=2}^{\infty} u_i \delta^{i+1} \end{pmatrix} = \sigma \left[\begin{pmatrix} E_0^4 \\ (E^4)_1 \delta \\ (E^4)_2 \delta^2 \\ \sum_{i>2} (E^4)_i \delta^i \end{pmatrix} - \begin{pmatrix} u_0 \\ u_1 \delta \\ u_2 \delta^2 \\ \sum_{i=3}^{\infty} u_i \delta^i \end{pmatrix} \right].$$

For the media counterpart (10.35b), we have

$$\partial_t \begin{pmatrix} 0 \\ 0 \\ E_0 \delta^2 \\ \sum_{i=1}^{\infty} E_i \delta^{i+2} \end{pmatrix} = - \int \sigma \left[\begin{pmatrix} E_0^4 \\ (E^4)_1 \delta \\ (E^4)_2 \delta^2 \\ \sum_{i>2} (E^4)_i \delta^i \end{pmatrix} - \begin{pmatrix} u_0 \\ u_1 \delta \\ u_2 \delta^2 \\ \sum_{i=3}^{\infty} u_i \delta^i \end{pmatrix} \right].$$

By identifying the coefficients of $1, \delta, \delta^2$ we finally obtain:

$$\begin{cases} E_0^4 = u_0, \\ \omega \partial_{\mathbf{x}} u_0 = \sigma((E^4)_1 - u_1), \\ \partial_t u_0 + \omega \partial_{\mathbf{x}} u_1 = \sigma((E^4)_2 - u_2), \\ (E^4)_1 = \int u_1, \\ \partial_t E_0 = - \int \sigma((E^4)_2 - u_2). \end{cases} \quad (10.39)$$

The first line of the previous expression ensures $u_0(\mathbf{x}, t, \omega) = u_0(\mathbf{x}, t)$. Together with the second line of (10.39), it implies $\frac{1}{3} \partial_{\mathbf{x}} u_0(\mathbf{x}, t) = -\sigma(\mathbf{x}) \int \omega u_1(\mathbf{x}, t, \omega) d\omega$. The latter expression is commonly called Fick's law. Now, plugging the last expression in the third line of (10.39) and integrating it over the angular distribution yields

$$\partial_t u_0 - \partial_{\mathbf{x}} \frac{1}{3\sigma} \partial_{\mathbf{x}} u_0 = \int \sigma((E^4)_2 - u_2).$$

Finally, adding the last line of (10.39) to the above equation leads to

$$\begin{cases} E_0^4(\mathbf{x}, t) = u_0(\mathbf{x}, t), \\ \partial_t(E_0(\mathbf{x}, t) + E_0^4(\mathbf{x}, t)) - \partial_{\mathbf{x}} \frac{1}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} E_0^4(\mathbf{x}, t) = 0. \end{cases} \quad (10.40)$$

System (10.40) is commonly called the *equilibrium* (relative to $E_0^4(\mathbf{x}, t) = u_0(\mathbf{x}, t)$) *diffusion* (relative to the second order operator in (10.40)) limit for system (10.35). In the next sections, we aim at building MC schemes to solve (10.35) with good asymptotical properties in the limit (10.40).

10.2.1 Classical Monte-Carlo schemes for photon transport

We first describe one of the most common methodology to solve system (10.35). It consists in a splitting between the transport phase (10.35a) and the implicit Stefan law (10.35b). It is commonly denoted by IMC for Implicit Monte-Carlo and has been introduced by Fleck and Cummings in [110]. The implicitation, presented in the next paragraph, introduces an artificial scattering term. This numerical trick is important and efficient in practice. This is mainly for this aspect we choose to describe this scheme in this section rather than others (such as the Carter-Forrest [57] or the N'Kaoua [211] ones, see [77] in which their asymptotical properties are studied).

In the next section, we briefly describe the IMC scheme for the resolution of (10.35). The beginning of the description may recall [77] but we go further by analysing the effect of a (spatial) discrepancy on the source term. It is crucial for the complete study of the asymptotic regime we aim at capturing (teleportation error, see [301]).

The classical MC scheme for the resolution of (10.35)

Once again, we suggest describing the MC scheme on one time step $[0, t]$. The IMC linearisation of system (10.35) can be summed up as follow: first, consider the new variable $\Theta(\mathbf{x}, t) = E^4(\mathbf{x}, t)$ such that $\partial_t \Theta(\mathbf{x}, t) = 4E^3(\mathbf{x}, t)\partial_t E(\mathbf{x}, t) = 4\Theta^{\frac{3}{4}}(\mathbf{x}, t)\partial_t E(\mathbf{x}, t) = \beta(\Theta(\mathbf{x}, t))\partial_t E(\mathbf{x}, t)$. The IMC scheme relies on rewriting system (10.35) with respect to the variables (u, Θ)

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\Theta(\mathbf{x}, t), \\ \partial_t \Theta(\mathbf{x}, t) = \beta(\Theta(\mathbf{x}, t))c\sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - \Theta(\mathbf{x}, t) \right). \end{cases} \quad (10.41a)$$

$$(10.41b)$$

Let us consider an explicitation⁷ of $\beta(\Theta(\mathbf{x}, t)) \approx \beta(\Theta(\mathbf{x})) = \beta(\mathbf{x})$ together with an implicated source term $\Theta(\mathbf{x}, t) \approx \Theta_t(\mathbf{x})$ in (10.41). Quantity $\Theta_t(\mathbf{x})$ denotes a constant with respect to time evaluation of $\Theta(\mathbf{x}, t)$ at the end of the times step $[0, t]$. With the above hypothesis, system (10.41) can be rewritten on time step $[0, t]$:

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\Theta_t(\mathbf{x}), \\ \partial_t \Theta(\mathbf{x}, t) = \beta(\mathbf{x})c\sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - \Theta_t(\mathbf{x}) \right). \end{cases} \quad (10.42a)$$

$$(10.42b)$$

Let us integrate (10.42b) on time step $[0, t]$ to obtain

$$\Theta_t(\mathbf{x}) - \Theta(\mathbf{x}) = \beta(\mathbf{x})c\sigma(\mathbf{x}) \left(\int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds - \Delta t \Theta_t(\mathbf{x}) \right).$$

We can evaluate the implicitation as follows

$$\begin{aligned} \Theta_t(\mathbf{x}) &= \Theta(\mathbf{x}) \frac{1}{1 + c\sigma(\mathbf{x})\beta(\mathbf{x})\Delta t} + \frac{\beta(\mathbf{x})c\sigma(\mathbf{x})}{1 + c\sigma(\mathbf{x})\beta(\mathbf{x})\Delta t} \int_0^t \int u(\mathbf{x}, s, \omega) d\omega ds, \\ &\approx \Theta(\mathbf{x}) \frac{1}{1 + c\sigma(\mathbf{x})\beta(\mathbf{x})\Delta t} + \frac{\beta(\mathbf{x})c\sigma(\mathbf{x})\Delta t}{1 + c\sigma(\mathbf{x})\beta(\mathbf{x})\Delta t} \int u(\mathbf{x}, t, \omega) d\omega, \\ &\approx \Theta(\mathbf{x})f(\mathbf{x}, \Delta t) + (1 - f(\mathbf{x}, \Delta t)) \int u(\mathbf{x}, t, \omega) d\omega. \end{aligned} \quad (10.43)$$

⁷i.e. $\Theta(\mathbf{x}, 0) = \Theta(\mathbf{x})$.

In the above expression, f is commonly called the Fleck factor [110].

The above expression of $\Theta_t(\mathbf{x})$ is then plugged in (10.42a) and a second equation is introduced to ensure energy conservation on time step $[0, t]$. We finally obtain

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = \\ c\sigma(\mathbf{x})f(\mathbf{x}, \Delta t)\Theta(\mathbf{x}) + c\sigma(\mathbf{x})(1 - f(\mathbf{x}, \Delta t)) \int u(\mathbf{x}, t, \omega) d\omega, \end{array} \right. \quad (10.44a)$$

$$\left. \begin{array}{l} \partial_t E(\mathbf{x}, t) = c\sigma(\mathbf{x})f(\mathbf{x}, \Delta t) \left(\int u(\mathbf{x}, t, \omega) d\omega - \Theta(\mathbf{x}) \right). \end{array} \right. \quad (10.44b)$$

System (10.44) is now only weakly coupled in the sense (10.44a) is self consistent and (10.44b) only depends on u solution of (10.44a) and other known fields (β, Θ, f). Equation (10.44a) is now linear and has a form that has already been intensively encountered in this document. The linearisation introduces an artificial scattering term, defined *via* the opacity $\sigma_s = \sigma(1 - f)$ and an artificial absorption term *via* the opacity $c\sigma_a = c\sigma f$ such that $\sigma_a + \sigma_s = \sigma$. The source term only depends on quantities evaluated at the beginning of the time step and has expression $c\sigma f \Theta$. Equation (10.44a) is *classically solved with the non-analog MC scheme of section 9.4 together with the source sampling strategy of section 9.9.1, see [110, 77, 301]*. The update of the material energy (and temperature) is made by instrumenting the MC resolution with a consistent track length estimator as in the previous section. The update is made once the tracking phase is over and every MC contributions tallied. The time step is then over.

Remark 10.1 *The description above can be found in many publications. We would like to focus briefly on the assumption made to go from the first line of (10.43) to its second one. Let us characterise this hypothesis. It can be summed up as $\forall t, s \in]0, t = \Delta t]$*

$$\frac{1}{t - 0} \iint_0^t u(\mathbf{x}, s, \omega) d\omega ds \approx \int u(\mathbf{x}, t, \omega) d\omega \iff \int u(\mathbf{x}, t, \omega) d\omega = cst = \int u(\mathbf{x}, s, \omega) d\omega. \quad (10.45)$$

In practice, during the resolution phase, the equality in (10.45) is only applied along a characteristic, i.e. we only use $\partial_s \int u(\mathbf{x} + c\omega s, s, \omega) d\omega = 0$ and $\iint_0^t u(\mathbf{x} + c\omega s, s, \omega) d\omega ds = \Delta t \int u(\mathbf{x} + c\omega t, t, \omega) d\omega$ rather than (10.45). In fact, the true linearisation hypothesis is closely related to the stiff equilibrium regime defined by⁸

$$\begin{aligned} & (\text{stiff equilibrium regime}) \quad \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int u(\mathbf{x}, t, \omega') d\omega', \\ & (\text{along a characteristic}) \quad \partial_s u(\mathbf{x} + c\omega s, s, \omega) + c\sigma(\mathbf{x})u(\mathbf{x} + c\omega s, s, \omega) = c\sigma(\mathbf{x}) \int u(\mathbf{x} + c\omega s, s, \omega') d\omega', \\ & (\text{isotropic as } \delta \rightarrow 0) \quad \partial_s \int u(\mathbf{x} + c\omega s, s, \omega) d\omega \approx 0. \end{aligned}$$

We briefly wanted to emphasize the hypothesis done in order to obtain (10.44) is closely related to plugging the solution of the stiff regime of interest along a characteristic to linearise the system, see point 4.) in section 10.1.3.

Now, we aim at analysing what can be asymptotically obtained (in the limit of an infinity of MC particles) from such linearisation. The latter is well-known in the literature and its asymptotical analysis, detailed in [77], leads to the following limit equation:

$$\partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \frac{1}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) = \frac{1}{\beta(\mathbf{x})\Delta t} (u^0(\mathbf{x}, t) - \Theta^0(\mathbf{x})). \quad (10.46)$$

The IMC scheme, in the limit of an infinity of MC particle to solve the linearized transport equation (10.44a), recovers the *diffusion limit* with the correct coefficient $\frac{1}{3\sigma}$, see (10.40), but misses the strict *equilibrium limit*. It introduces a numerical relaxation time $\beta\Delta t$ such that the smaller $\beta\Delta t$, the better the equilibrium limit is fulfilled. Note that the authors in [77, 282] emphasized the fact that in practice, time steps ensuring a good agreement for the local equilibrium limit are affordable for the IMC scheme (i.e. we can afford $\beta\Delta t \sim 0$ small enough). On another hand, the treatment of the source term is central

⁸relaxed state of (10.35).

for the limit regime we aim at capturing: a small inaccuracy in its sampling can introduce a significant *teleportation error*. The denomination has been introduced in [197] and is very well described in [301]. Until now, we only summed up results which can be found in many publications, we now would like to introduce some originality in the study, especially with respect to this teleportation error.

The analysis of the limit equation on time step $[0, t]$ for the IMC scheme has been carried on continuously with respect to the spatial variable $\mathbf{x} \in \mathcal{D}$ whereas many authors (see [301] and the references therein) admit the spatial discretisation of the source term in the IMC linearisation is closely related to the teleportation error. See [77, 69, 70, 301] and [282] for very pedagogical numerical examples. It is even confirmed by the fact that *tilts*, i.e. spatial (bi-)linear interpolatory reconstructions of Θ can lead to significant improvements, see [69, 70, 301]. In practice, a spatial discretisation is introduced (as in section 9.6, i.e. $\mathcal{D} = \bigcup_{i=1}^{N_x} \mathcal{D}_i$), together with assumptions such as constant Θ in each cell. At every beginning of time step we have an $\mathcal{O}(\Delta x)$ approximation of $\Theta(\mathbf{x}) = \sum_{i=1}^{N_x} \Theta^i \mathbf{1}_{\mathcal{D}_i}(\mathbf{x}) + \mathcal{O}(\Delta x)$ where $\Delta x = \max_{i \in \{1, \dots, N_x\}} (|\mathcal{D}_i|)$ and $\Theta^i = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} \Theta(\mathbf{x}) d\mathbf{x}$. This choice affects $\Theta(\mathbf{x})$ but also $\beta(\mathbf{x}) = \beta(\Theta(\mathbf{x}))$ and above all the *source sampling* of the MC particles at every beginning of time steps. The material of section 9.8.1 (together with the examples) put forward the importance of accurate samplings within cells. In the following paragraphs, we suggest

- taking into account a spatial $\mathcal{O}(\delta_x)$ discrepancy in the source term within cells (inaccurate source sampling),
- together with the asymptotic development $\mathcal{O}(\delta)$ of the linearized system (10.42).

Let us now revisit the limit equation under such condition.

Let us perform a Taylor development of $\Theta(\mathbf{x})$ with respect to a small spatial parameter δ_x . Note that we are also going to perform a Hilbert development afterward with respect to δ . For this, we write $\Theta(\mathbf{x}) = \Theta_0(\mathbf{x}) + \Theta_1(\mathbf{x})\delta_x + \mathcal{O}(\delta_x^2)$ with of course $\Theta_0(\mathbf{x}) = \sum_{i=1}^{N_x} \Theta_0^i \mathbf{1}_{\mathcal{D}_i}(\mathbf{x})$ and $\Theta_1(\mathbf{x}) = \partial_x \Theta(\mathbf{x})$. The superscripts refer to the terms in the development with respect to δ (as before). Assume constant opacities $\sigma(\mathbf{x}) = \sigma$, then for small δ_x , we have

$$\begin{aligned} c\sigma f \Theta(\mathbf{x}) &= \underset{\delta_x \sim 0}{=} \frac{c\sigma \Theta_0}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} + c\sigma \delta_x \Theta_1 \frac{1 + c\sigma \Delta t \Theta_0^{3/4}}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2} + \mathcal{O}(\delta_x^2), \\ c\sigma(1-f) &= \underset{\delta_x \sim 0}{=} \frac{4c^2 \sigma^2 \Delta t \Theta_0^{3/4}}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} + c\sigma \delta_x \Theta_1 \frac{3c\sigma \Delta t}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2 \Theta_0^{1/4}} + \mathcal{O}(\delta_x^2), \\ c\sigma f &= \underset{\delta_x \sim 0}{=} \frac{c\sigma}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} - c\sigma \delta_x \Theta_1 \frac{3c\sigma \Delta t}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2 \Theta_0^{1/4}} + \mathcal{O}(\delta_x^2). \end{aligned} \quad (10.47)$$

The above developments are then plugged in the collisional part of (10.42) solved on time step $[0, t]$. This yields (we drop the dependences for convenience)

$$\left\{ \begin{aligned} \partial_t u + c\omega \partial_x u + c\sigma u &= \frac{c\sigma \Theta_0}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} + c\sigma \delta_x \Theta_1 \frac{1 + c\sigma \Delta t \Theta_0^{3/4}}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2} + \mathcal{O}(\delta_x^2) \\ &\quad + \left(\frac{4c^2 \sigma^2 \Delta t \Theta_0^{3/4}}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} + \frac{3c^2 \sigma^2 \Delta t \Theta_1 \delta_x}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2 \Theta_0^{1/4}} + \mathcal{O}(\delta_x^2) \right) \int u, \end{aligned} \right. \quad (10.48a)$$

$$\left\{ \begin{aligned} \partial_t E &= \left(\frac{c\sigma}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} - \delta_x \Theta_1 \frac{3c^2 \sigma^2 \Delta t}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2 \Theta_0^{1/4}} + \mathcal{O}(\delta_x^2) \right) \int u \\ &\quad - \frac{c\sigma \Theta_0}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} - c\sigma \delta_x \Theta_1 \frac{1 + c\sigma \Delta t \Theta_0^{3/4}}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2} + \mathcal{O}(\delta_x^2). \end{aligned} \right. \quad (10.48b)$$

Note that (10.48b) is built ensuring conservation of (10.48) on time step $[0, t]$. Now use the fact that

$c\frac{\mathcal{T}}{\mathcal{D}} = \mathcal{O}(\frac{c^*}{\delta})$, $c\sigma\frac{\mathcal{T}}{\lambda} = \mathcal{O}(\frac{c^*\sigma^*}{\delta^2})$ in (10.48) so that we have

$$\begin{aligned} \delta^2\partial_t u + \delta c\omega\partial_{\mathbf{x}} u + c\sigma u &= \frac{\delta^2 c\sigma\Theta_0}{\delta^2 + 4\Delta t c\sigma\Theta_0^{3/4}} + c\sigma \frac{\delta^4 + \delta^2\Delta t c\sigma\Theta_0^{3/4}}{(\delta^2 + 4\Delta t c\sigma\Theta_0^{3/4})^2} \delta_{\mathbf{x}}\Theta_1 \\ &+ \left(\frac{4\Delta t(c\sigma)^2\Theta_0^{3/4}}{\delta^2 + 4\Delta t c\sigma\Theta_0^{3/4}} + \frac{3\delta^2\Delta t(c\sigma)^2}{(\delta^2 + 4\Delta t c\sigma\Theta_0^{3/4})^2\Theta_0^{1/4}} \Theta_1\delta_{\mathbf{x}} \right) \int u d\omega + \mathcal{O}(\delta_{\mathbf{x}}^2). \end{aligned} \quad (10.49)$$

Assuming $\delta \rightarrow 0$ and keeping only the $\mathcal{O}(\delta^4)$ orders, we obtain

$$\begin{aligned} \delta^2\partial_t u(\mathbf{x}, t, \omega) + \delta c\omega\partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) &= \delta^2 \frac{\Theta_0^{1/4}(\mathbf{x})}{4\Delta t} - \delta^2 \frac{\delta_{\mathbf{x}}\Theta_1(\mathbf{x})}{16\Delta t\Theta_0^{3/4}(\mathbf{x})} \\ &+ \left(c\sigma(\mathbf{x}) - \frac{\delta^2}{4\Delta t\Theta_0^{3/4}(\mathbf{x})} + \delta^2 \frac{3\Theta_1(\mathbf{x})\delta_{\mathbf{x}}}{16\Delta t\Theta_0^{7/4}(\mathbf{x})} \right) \int u(\mathbf{x}, t, \omega) d\omega + \mathcal{O}(\delta_{\mathbf{x}}^2) + \mathcal{O}(\delta^4). \end{aligned} \quad (10.50)$$

The leading order terms allows identifying the asymptotic regime for the linearized equation (10.50)

$$\left\{ \begin{array}{l} u^0(\mathbf{x}, t, \omega) = \int u^0(\mathbf{x}, t, \omega) d\omega, \\ c\omega\partial_{\mathbf{x}} u^0(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u^1(\mathbf{x}, t, \omega) = 0, \\ \partial_t u^0(\mathbf{x}, t, \omega) + c\omega\partial_{\mathbf{x}} u^1(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u^2(\mathbf{x}, t, \omega) = \frac{1}{4\Delta t(\Theta_0^0(\mathbf{x}))^{3/4}} \left(\Theta_0^0(\mathbf{x}) - \int u^0(\mathbf{x}, t, \omega) d\omega \right) \\ - \frac{\delta_{\mathbf{x}}\Theta_1^0(\mathbf{x})}{16\Delta t(\Theta_0^0(\mathbf{x}))^{3/4}} + \int u^2(\mathbf{x}, t, \omega) d\omega + \frac{3}{16} \frac{1}{\Delta t(\Theta_0^0(\mathbf{x}))^{7/4}} \Theta_1^0(\mathbf{x})\delta_{\mathbf{x}} \int u^0(\mathbf{x}, t, \omega) d\omega + \mathcal{O}(\delta_{\mathbf{x}}^2). \end{array} \right. \quad (10.51)$$

Integrating the previous relations over angles and combining them (just as in the beginning of section 10.2) yields

$$\begin{aligned} \partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \frac{c}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) &= \frac{1}{4\Delta t(\Theta_0^0(\mathbf{x}))^{3/4}} (\Theta_0^0(\mathbf{x}) - u^0(\mathbf{x}, t)) \\ &+ \frac{1}{16\Delta t(\Theta_0^0(\mathbf{x}))^{7/4}} (3u^0(\mathbf{x}, t) - \Theta_0^0(\mathbf{x})) \delta_{\mathbf{x}}\Theta_1^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^2). \end{aligned} \quad (10.52)$$

Besides, the leading order with respect to δ in (10.48b) ensures $\int u^0(\mathbf{x}, t, \omega) d\omega = \Theta_0^0(\mathbf{x}) + \delta_{\mathbf{x}}\Theta_1^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^2)$. Together with the isotropy of u^0 , we have

$$\begin{aligned} \partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \frac{c}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) &= -\frac{\delta_{\mathbf{x}}}{8\Delta t(\Theta_0^0(\mathbf{x}))^{3/4}} \partial_{\mathbf{x}}\Theta^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^2), \\ &= -\frac{\delta_{\mathbf{x}}}{2\Delta t} \partial_{\mathbf{x}}\Theta_0^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^2), \\ &= -\frac{\delta_{\mathbf{x}}}{2\Delta t} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) + \mathcal{O}(\delta_{\mathbf{x}}^2). \end{aligned} \quad (10.53)$$

The limit equation on time step $[0, t]$ of the IMC scheme obtained with an infinitely accurate MC approximation for (10.44a) but taking into account an $\mathcal{O}(\delta_{\mathbf{x}})$ discrepancy in the source term (during the source sampling phase) yields

$$\partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \frac{c}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) + \frac{\delta_{\mathbf{x}}}{2\Delta t} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) = \mathcal{O}(\delta_{\mathbf{x}}^2). \quad (10.54)$$

The limit equation (10.54) is an advection-diffusion one. The velocity of the advection operator depends on the discretisation parameters Δt and $\delta_{\mathbf{x}}$. They even compete during time step $[0, t]$. On one hand, taking Δt the smaller possible ensures recovering the *equilibrium* limit (first term of the right hand side of (10.52)), but imposes a finer and finer spatial discretisation (perturbed advection velocity in (10.54)). In fact, the error can be very important for steep gradients of u^0 (front of a Marshak wave for example for which $\partial_{\mathbf{x}} u^0 \gg 1$). It is well-known the mechanism accumulates discrepancies proportionally to the number of time steps (due to the cycle-to-cycle differences of magnitude $\frac{\delta_{\mathbf{x}}}{\Delta t} \partial_{\mathbf{x}} u^0$, see [301]). The behaviour of the IMC scheme is close to the behaviour of the Dufort-Frankel scheme for parabolic equations. For this scheme, if $C\delta_{\mathbf{x}} = \Delta t \rightarrow 0$, the scheme is inconsistent. It is possible to force $\Delta t = C$

the greater possible (limit of stability) and make sure $\delta_x \rightarrow 0$ to obtain a converging $\mathcal{O}(\delta_x)$ scheme.

Let us detail another way to put forward the sensitivity to a small error with respect to a spatial perturbation of Θ during the source sampling phase. Let us consider the linearized IMC system (10.48) and subtract the system below (we drop the dependences for convenience)

$$\begin{cases} \partial_t \tilde{u} + c\omega \partial_x \tilde{u} + c\sigma \tilde{u} = \\ \frac{c\sigma \Theta_0}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} + \left(\frac{4c^2 \sigma^2 \Delta t \Theta_0^{3/4}}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} \right) \int \tilde{u}, \end{cases} \quad (10.55a)$$

$$\partial_t \tilde{E} = \left(\frac{c\sigma}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} \right) \int \tilde{u} - \frac{c\sigma \Theta_0}{1 + 4c\sigma \Delta t \Theta_0^{3/4}}. \quad (10.55b)$$

In the above expression, we truncated Θ to its first term Θ_0 . Let us introduce $e_1 = u - \tilde{u}$ and $e_2 = E - \tilde{E}$. They correspond to the errors due to a discrepancy in the spatial discretisation (source sampling) of Θ on time step $[0, \Delta t]$ with respect to a spatially accurate IMC temporal linearisation. We have (we focus on the first equation as it is self-consistent)

$$\begin{aligned} \partial_t e_1 + c\omega \partial_x e_1 + c\sigma e_1 &= c\sigma \delta_x \Theta_1 \frac{1 + c\sigma \Delta t \Theta_0^{3/4}}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2} \\ &\quad + \frac{4c^2 \sigma^2 \Delta t \Theta_0^{3/4}}{1 + 4c\sigma \Delta t \Theta_0^{3/4}} \int e_1 d\omega + \frac{3c^2 \sigma^2 \Delta t \Theta_1 \delta_x}{(1 + 4c\sigma \Delta t \Theta_0^{3/4})^2 \Theta_0^{1/4}} \int u. \end{aligned}$$

We can rewrite it more concisely

$$T(e_1(\mathbf{x}, t, \omega)) = c\sigma(\mathbf{x}) \delta_x \Theta_1(\mathbf{x}) \frac{1 + c\sigma(\mathbf{x}) \Delta t \Theta_0^{3/4}(\mathbf{x})}{(1 + 4c\sigma(\mathbf{x}) \Delta t \Theta_0^{3/4}(\mathbf{x}))^2} + \frac{3c^2 \sigma(\mathbf{x})^2 \Delta t \Theta_1(\mathbf{x}) \delta_x}{(1 + 4c\sigma(\mathbf{x}) \Delta t \Theta_0^{3/4}(\mathbf{x}))^2 \Theta_0^{1/4}(\mathbf{x})} \int u(\mathbf{x}, t, \omega) d\omega.$$

In the above expression, if the sources are infinitely accurately sampled (no spatial discrepancies in Θ), $T(e_1(\mathbf{x}, t, \omega)) = 0$. In this case, we recover the temporal IMC discretization. This is precisely this latter property we will aim at having *by construction* in the next sections. Now, let us compute

$$\begin{aligned} T(e_1(\mathbf{x}, t, \omega)) &\stackrel{c\sigma(\mathbf{x}) \sim \frac{1}{\delta^2}}{=} \frac{1}{\delta^2} \delta_x \Theta_1(\mathbf{x}) \frac{1 + \frac{1}{\delta^2} \Delta t \Theta_0^{3/4}(\mathbf{x})}{(1 + 4 \frac{1}{\delta^2} \Delta t \Theta_0^{3/4}(\mathbf{x}))^2} + \frac{3 \frac{1}{\delta^4} \Delta t \Theta_1(\mathbf{x}) \delta_x}{(1 + 4 \frac{1}{\delta^2} \Delta t \Theta_0^{3/4}(\mathbf{x}))^2 \Theta_0^{1/4}(\mathbf{x})} \int u(\mathbf{x}, t, \omega) d\omega, \\ &\stackrel{\delta \sim 0}{=} \frac{\delta_x}{4\delta^2} \partial_x \Theta(\mathbf{x}) + \mathcal{O}(\delta^0). \end{aligned}$$

The above expression tends to show the IMC linearisation is not AP (cf. definition 9.1) due to the spatial discrepancy (δ_x). Furthermore, (10.54) shows that any such discrepancy will be amplified along the time-steps. Any small spatial error on the emission will be amplified in the equilibrium diffusion limit. In order to design AP MC scheme, care has to be taken to avoid such behaviour during the MC phase. This will be the aim of the next sections but first, we would like to study the effect of having resort to source tilting.

As emphasized in many publications, see [301] and the references therein, if the emission is non uniform within the cell, a reconstruction method can be introduced to estimate $c\sigma f\Theta$ in (10.42) at every beginning of time steps. This is commonly called a *tilt* (see [301, 69, 70]) and it has been experimentally observed it reduces teleportation errors. In fact, it corresponds to a second order approximation of $\Theta(\mathbf{x}) = \Theta^{tilt}(\mathbf{x}) + \mathcal{O}(\delta_x^2)$. The same asymptotic developments (Taylor of order 2 with respect to δ_x and Hilbert one with respect to δ) applying an accurate tilt can be performed. The second order equivalent

of (10.47) reads

$$\begin{aligned}
c\sigma f\Theta(\mathbf{x}) &= \underset{\delta_{\mathbf{x}} \sim 0}{=} \frac{c\sigma\Theta_0}{1 + 4c\sigma\Delta t\Theta_0^{3/4}} + \delta_{\mathbf{x}}^2 c\sigma\Theta_2 \frac{1 + c\sigma\Theta_0^{3/4}\Delta t}{(1 + 4c\sigma\Theta_0^{3/4}\Delta t)^2} + \mathcal{O}(\delta_{\mathbf{x}}^3), \\
c\sigma(1 - f) &= \underset{\delta_{\mathbf{x}} \sim 0}{=} \frac{4c^2\sigma^2\Delta t\Theta_0^{3/4}}{1 + 4c\sigma\Delta t\Theta_0^{3/4}} + \delta_{\mathbf{x}}^2 \frac{3c^2\sigma^2\Theta_2\Delta t}{\Theta_0^{1/4}(1 + 4c\sigma\Theta_0^{3/4}\Delta t)^2} + \mathcal{O}(\delta_{\mathbf{x}}^3), \\
c\sigma f\beta(\Theta(\mathbf{x})) &= \underset{\delta_{\mathbf{x}} \sim 0}{=} \frac{4c\sigma\Theta_0^{3/4}}{1 + 4c\sigma\Delta t\Theta_0^{3/4}} + \delta_{\mathbf{x}}^2 \frac{3c\sigma\Theta_2}{\Theta_0^{1/4}(1 + 4c\sigma\Theta_0^{3/4}\Delta t)^2} + \mathcal{O}(\delta_{\mathbf{x}}^3), \\
c\sigma f\beta(\Theta(\mathbf{x}))\Theta(\mathbf{x}) &= \underset{\delta_{\mathbf{x}} \sim 0}{=} \frac{4c\sigma\Theta_0^{7/4}}{1 + 4c\sigma\Delta t\Theta_0^{3/4}} + \delta_{\mathbf{x}}^2 c\sigma\Theta_2 \sqrt{\Theta_0} \frac{7\Theta_0^{1/4} + 16\Theta_0 c\sigma\Delta t}{(1 + 4c\sigma\Theta_0^{3/4}\Delta t)^2} + \mathcal{O}(\delta_{\mathbf{x}}^3),
\end{aligned} \tag{10.56}$$

with $\Theta_2(\mathbf{x}) = \partial_{\mathbf{x}\mathbf{x}}^2\Theta(\mathbf{x})$. Plugged in the transport equation linearized on time step $[0, t]$ and going through the same steps as before leads to

$$\begin{aligned}
\partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \frac{c}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} u^0(\mathbf{x}, t) &= -\frac{\delta_{\mathbf{x}}^2}{8\Delta t(\Theta_0^0(\mathbf{x}))^{3/4}} \partial_{\mathbf{x}\mathbf{x}}^2 \Theta^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^3), \\
&= -\frac{\delta_{\mathbf{x}}^2}{2\Delta t} \partial_{\mathbf{x}\mathbf{x}}^2 \Theta_0^0(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^3), \\
&= -\frac{\delta_{\mathbf{x}}^2}{2\Delta t} \partial_{\mathbf{x}\mathbf{x}}^2 u^0(\mathbf{x}, t) + \mathcal{O}(\delta_{\mathbf{x}}^3).
\end{aligned} \tag{10.57}$$

Re-arranging the terms in (10.57) produces expression

$$\partial_t u^0(\mathbf{x}, t) - \partial_{\mathbf{x}} \left[\frac{c}{3\sigma(\mathbf{x})} + \frac{\delta_{\mathbf{x}}^2}{2\Delta t} \right] \partial_{\mathbf{x}} u^0(\mathbf{x}, t) = \mathcal{O} \left(\frac{\delta_{\mathbf{x}}^3}{\Delta t} \right). \tag{10.58}$$

It is a diffusion equation (up to order $\mathcal{O}(\delta_{\mathbf{x}}^3)$). The asymptotic diffusion coefficient is $\frac{c}{3\sigma(\mathbf{x})} + \frac{\delta_{\mathbf{x}}^2}{2\Delta t}$ and depends on discretisation parameters $\delta_{\mathbf{x}}$ and Δt . Once again, both discretisation parameters compete during time step $[0, t]$. This is attenuated by the fact that $\delta_{\mathbf{x}}$ is squared but it still accumulates discrepancies proportionally to the number of time steps due to the cycle-to-cycle differences of magnitude $\frac{\delta_{\mathbf{x}}^2}{2\Delta t} \partial_{\mathbf{x}\mathbf{x}}^2 u^0$. The same remark as above still applies here: if $\Delta t = C\delta_{\mathbf{x}}^2$ with $\delta_{\mathbf{x}}$ going to zero, the tilted IMC scheme is inconsistent (behaviour to be compared with the Dufort-Frankel scheme for parabolic equations). On another hand, if $\frac{\delta_{\mathbf{x}}}{\Delta t} = C$ is kept constant and $\delta_{\mathbf{x}}$ goes to zero, a convergence behaviour can still be observed.

The above analysis shows that an accurate third order (tilt) reconstruction (i.e. $\Theta(\mathbf{x}) = \Theta^{tilt}(\mathbf{x}) + \mathcal{O}(\delta_{\mathbf{x}}^3)$) is mandatory to recover the correct diffusion coefficient for the regime $\delta \rightarrow 0$. This observation is all the more interesting that in plasma physics, many PIC codes rely on a third order spatial discretisation of the electromagnetic fields (often third order Splines, see [25, 24, 29] for example). This is probably for the same reason as in the previous section: the linearisation induces modification of the propagation waves for first and second order operators. Of course, I did not study precisely the MC scheme of PIC codes but this should be interesting to do so with the above point of view. Maybe the material of the next section could also benefit their resolution. Due to

- the appearance of competing discretisation parameters $(\frac{\delta_{\mathbf{x}}^k}{\Delta t})_{k=1,2,\dots}$ at every reconstruction order k (even if attenuated from $\frac{\delta_{\mathbf{x}}}{\Delta t}$ to $\frac{\delta_{\mathbf{x}}^2}{\Delta t}$ from order 1 to order 2),
- and the possible error accumulations along the cycles,

instead of introducing an additional reconstruction, we prefer building MC schemes which do not suffer teleportation errors. Based on these remarks, in the following sections, we present two new MC schemes, three if we count the (approximated) one of remark 10.3, and detail their asymptotic properties.

10.2.2 Two Asymptotic Preserving MC schemes for photon transport

At this stage, it may be tempting to try to apply strictly the same methodology as in section 10.1. In other words, plug the solution of the asymptotic regime of interest, here the equilibrium diffusion limit,

in the MC scheme *via* a change of variable ($u = Uf$) as in [3]. This will be briefly done in remark 10.3. We first want to emphasize the fact that for system (10.30) there are different presentation possibilities, related to more relevant linearisations of (10.30) on time step $[0, t]$.

In this section, we present two MC schemes to solve system (10.30) bearing interesting asymptotic properties (as $\delta \rightarrow 0$). Depending on the linearisations, those properties differ. Their common point remains an accurate *diffusion limit* capturing behaviour (and *no teleportation error* but the notions are closely related). The two MC schemes are presented so the reader can pick the solver having the properties he finds the more relevant. The first linearisation ensures capturing the *diffusion limit*, is conservative but does not capture exactly the *equilibrium* one (only up to $\mathcal{O}(\Delta t)$). The second captures the *equilibrium diffusion* limit but does not ensure exact conservation of energy at the end of the time step (only up to $\mathcal{O}(\Delta t)$).

The study of section 10.2.1 showed a reconstruction of $E^4 = \Theta$ is mandatory to ensure recovering the diffusion limit. For this, we built high order reconstruction of this field from a spatial Taylor serie. In this section, we suggest building MC schemes which do not need this spatial reconstruction (or one may say they reconstruct it *on-the-fly* during the MC resolution).

A conservative Asymptotic Preserving MC scheme for the *diffusion* limit

We aim at solving (10.34) together with capturing regime (10.40) characterised by $\delta \rightarrow 0$. We here suggest introducing the variable $e(\mathbf{x}, t, \omega)$ defined by $\int e(\mathbf{x}, t, \omega) d\omega = E(\mathbf{x}, t)$. System (10.34) can then be rewritten in term of unknowns (u, e) by

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \left(\int e(\mathbf{x}, t, \omega) d\omega \right)^4, \\ \partial_t e(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \left(\int u(\mathbf{x}, t, \omega) d\omega - \left(\int e(\mathbf{x}, t, \omega) d\omega \right)^4 \right). \end{array} \right. \quad (10.59a)$$

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t)) \int e(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t))e(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int u(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.60b)$$

Integrating (10.59b) with respect to ω (such that $\int d\omega = 1$) allows recovering (10.35b). Besides, introduce $\eta(E(\mathbf{x}, t)) = E^3(\mathbf{x}, t)$. Then it can also be rewritten

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t)) \int e(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t))e(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int u(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.60a)$$

$$\left\{ \begin{array}{l} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t)) \int e(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(E(\mathbf{x}, t))e(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int u(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.60b)$$

System (10.60) is still nonlinear. On time step $[0, t]$, we consider linearisations corresponding to a particular choice of $\eta(E(\mathbf{x}, t)) = E^3(\mathbf{x}, t) = (\int e(\mathbf{x}, t, \omega) d\omega)^3$ with respect to the time variable, i.e. explicit, implicit etc. Independently of this choice, the linearized system is conservative as

$$\partial_t \left(\int u(\mathbf{x}, t, \omega) d\omega + \int e(\mathbf{x}, t, \omega) d\omega \right) + \partial_{\mathbf{x}} \int c\omega u(\mathbf{x}, t, \omega) d\omega = 0.$$

An explicit choice for η together with the non-analog⁹ scheme of section 9.4 lead to the known MC scheme of [11]. In practice, it is computationally unusable due to the small time steps it needs for stability [301]. For relevant and efficient numerical tricks to be able to take bigger time steps with this same scheme, we rely on the work of [282] summed up in the last paragraph of this section. We here want to focus on how to build an MC scheme for the resolution of system (10.60) of unknowns (u, e) . Discretising e with an MC approximation ensures we do not need to reconstruct a density or use source sampling and avoid, by construction, the teleportation error.

Let us assume a linearisation based on a choice¹⁰ of $\eta(E(\mathbf{x}, t)) \approx \eta(\mathbf{x}, t)$ on time step $[0, t]$. The

⁹and *not* the semi-analog scheme as in the title of [11].

¹⁰explicit or implicit, this only affects the resolution, not the MC scheme.

linearisation of system (10.60) of solution¹¹ (ϕ, e_m) is then given by

$$\begin{cases} \partial_t \phi(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\phi(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(\mathbf{x}, t) \int e_m(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e_m(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(\mathbf{x}, t)e_m(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int \phi(\mathbf{x}, t, \omega) d\omega. \end{cases} \quad (10.61a)$$

$$(10.61b)$$

System (10.61) is now linear and can be solved with an MC scheme. It has the same structure as a multigroup transport equation in neutronics for example, see [173, 268]. But in this case, there are only two groups and the basis functions are analytical, given by $\phi_0(v) = \delta_c(v)$ and $\phi_1(v) = \delta_0(v)$. Let us clarify this point. To solve (10.61), let us build $u(\mathbf{x}, t, \omega, v)$, a new unknown depending on one more dimension and on unknowns (ϕ, e) solutions of (10.61). For this, the variable v is chosen such that

$$u(\mathbf{x}, t, \mathbf{v}) = \phi(\mathbf{x}, t, \omega)\delta_c(v) + e_m(\mathbf{x}, t, \omega)\delta_0(v).$$

In fact, v is nothing more than a velocity which can be c for photons or 0 for matter. Let us now build the linear equation satisfied by $u(\mathbf{x}, t, \omega, v)$. Expression (10.61) can be rewritten (we drop the dependences for conciseness)

$$\partial_t \underbrace{(\phi\delta_c + e_m\delta_0)}_u + v\omega \partial_{\mathbf{x}} \underbrace{(\phi\delta_c + e_m\delta_0)}_u = -c\sigma(\phi\delta_c + \eta e_m\delta_0) + c\sigma \left(\eta \int e_m\delta_c + \int \phi\delta_0 \right). \quad (10.62)$$

It remains to make u appear in the collisional part. For the moment the integration is only over the angular distribution. Let us introduce the Kronecker symbols

$$\delta_{0,c}(v', v) = \delta_0(v)\delta_c(v') \quad \text{and} \quad \delta_{c,0}(v', v) = \delta_c(v)\delta_0(v'), \quad (10.63)$$

and rewrite (10.62) as

$$\begin{aligned} \partial_t u(\mathbf{x}, t, v, \omega) + v\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + c\sigma(\mathbf{x})(\delta_c(v) + \eta(\mathbf{x}, t)\delta_0(v))u(\mathbf{x}, t, v, \omega) = \\ + c\sigma \iint [u(\mathbf{x}, t, v', \omega')\delta_{0,c}(v', v) + u(\mathbf{x}, t, v', \omega')\eta(\mathbf{x}, t)\delta_{c,0}(v', v)] dv' d\omega'. \end{aligned} \quad (10.64)$$

One can check that choosing $v = c$ in (10.64) allows recovering (10.61a) and choosing $v = 0$ in (10.64) leads to (10.61b). We can identify scattering and total cross-sections to rewrite (10.64) under the general form encountered all along chapter 9:

$$\begin{aligned} \partial_t u(\mathbf{x}, t, v, \omega) + v, \omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + \sigma_t(\mathbf{x}, t, v, \omega)u(\mathbf{x}, t, v, \omega) = \\ \iint \sigma_s(\mathbf{x}, t, v, v')u(\mathbf{x}, t, v', \omega')dv' d\omega'. \end{aligned} \quad (10.65)$$

In the above expression, we have

$$\sigma_t(\mathbf{x}, t, v) = c\sigma(\mathbf{x})(\delta_c(v) + \eta(\mathbf{x}, t)\delta_0(v)), \quad \text{and} \quad \sigma_s(\mathbf{x}, t, v, v') = c\sigma(\mathbf{x})(\delta_{0,c}(v', v) + \eta(\mathbf{x}, t)\delta_{c,0}(v', v)).$$

Let us rewrite the scattering part as $\sigma_s(\mathbf{x}, t, v)P_s(\mathbf{x}, t, v, v') = \sigma_s(\mathbf{x}, t, v, v')$, this implies

$$\begin{aligned} \sigma_s(\mathbf{x}, t, v) &= \int \sigma_s(\mathbf{x}, t, v, v')dv', \\ &= \int c\sigma(\mathbf{x})(\delta_{0,c}(v', v) + \eta(\mathbf{x}, t)\delta_{c,0}(v', v))dv', \\ &= c\sigma(\mathbf{x})(\eta(\mathbf{x}, t)\delta_{c,0}(v, 0) + \delta_{0,c}(v, c)), \\ &= c\sigma(\mathbf{x})(\eta(\mathbf{x}, t)\delta_c(v) + \delta_0(v)). \end{aligned}$$

¹¹We insist on the change of notation:

- (u, e) is solution of the nonlinear system (10.60),
- whereas (ϕ, e_m) is solution of the linearized system (10.61) on time step $[0, t]$.

By definition of P_s we have

$$\begin{aligned} P_s(\mathbf{x}, t, v, v') &= \frac{\sigma_s(\mathbf{x}, t, v, v')}{\sigma_s(\mathbf{x}, t, v)}, \\ &= \frac{\delta_{0,c}(v', v) + \eta(\mathbf{x}, t)\delta_{c,0}(v', v)}{\delta_0(v) + \eta(\mathbf{x}, t)\delta_c(v)}. \end{aligned}$$

The above expression can be considerably simplified by noticing that

$$\text{for } v = 0, P_s(\mathbf{x}, t, 0, v') = \delta_c(v'), \text{ and for } v = c, P_s(\mathbf{x}, t, c, v') = \delta_0(v').$$

Practically, the later expressions imply the scattering term systematically makes an MC particle change state at a collision:

- if a particle represents matter (i.e. $v = 0$), the outer particle represents a photon with probability 1.
- Conversely if a particle representing a photon (with $v = c$) encounters a collision, it is transformed into a matter MC particle with probability 1.

Now, we are interested in a direct resolution of (10.64) on time step $[0, t]$. We consequently apply the material of section 9.5 and consider the adjoint form of (10.64) given by

$$\begin{aligned} -\partial_t u(\mathbf{x}, t, v, \omega) - v, \omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + \sigma_t(\mathbf{x}, t, v)u(\mathbf{x}, t, v, \omega) &= \\ \iint \sigma_s(\mathbf{x}, t, v, v')P_s(\mathbf{x}, t, v, v')u(\mathbf{x}, t, v', \omega')d\omega'dv'. \end{aligned} \quad (10.66)$$

Just as in section 9.5, we introduce

$$\sigma_S(\mathbf{x}, t, v, v') = \sigma_S(\mathbf{x}, t, \mathbf{v})P_S(\mathbf{x}, t, v, v') = \sigma_s(\mathbf{x}, t, \mathbf{v}')P_s(\mathbf{x}, t, v, v').$$

The latter can be characterised computing

$$\begin{aligned} \sigma_S(\mathbf{x}, t, v) &= \int \sigma_s(\mathbf{x}, t, v')P_s(\mathbf{x}, t, v, v')dv', \\ &= \int c\sigma(\mathbf{x})(\eta(\mathbf{x}, t)\delta_c(v') + \delta_0(v'))\frac{\delta_{0,c}(v', v) + \eta(\mathbf{x}, t)\delta_{c,0}(v', v)}{\delta_0(v) + \eta(\mathbf{x}, t)\delta_c(v)}, \\ &= c\sigma(\mathbf{x})\eta(\mathbf{x}, t)\frac{\delta_c(v)}{\delta_0(v) + \eta(\mathbf{x}, t)\delta_c(v)} + c\sigma(\mathbf{x})\eta(\mathbf{x}, t)\frac{\delta_0(v)}{\delta_0(v) + \eta(\mathbf{x}, t)\delta_c(v)}, \\ &= c\sigma(\mathbf{x})\delta_c(v) + c\sigma(\mathbf{x})\eta(\mathbf{x}, t)\delta_0(v) = \sigma_t(\mathbf{x}, t, v). \end{aligned}$$

Few calculations, similar to the already performed one before to identify P_s , show that

$$P_S(\mathbf{x}, t, v, v') = P_s(\mathbf{x}, t, v, v') = 1 - \delta_v(v').$$

Suppose one wants to apply the (direct) non-analog MC scheme of section 9.4 to solve (10.61). Then the samplings are quite simple:

- the time interaction is sampled from σ_S , i.e. $c\sigma$ if the MC particle represents a photon (i.e. if $v = c$) or $c\sigma\eta$ if the MC particle represents matter (i.e. if $v = 0$).
- We have $\sigma_A(\mathbf{x}, t, v) = 0, \forall \mathbf{x} \in \mathcal{D}, t \in [0, t], v \in \{0, c\}$ so that the weight of the MC particle does not change all along the tracking phase.
- Finally, if the MC particle encounters an interaction, it changes of state with probability 1 (i.e. systematically), from photon to matter and matter to photon.

Let us show why it is a satisfying linearisation for our coupled system. On time step $[0, t]$, the previously

presented MC scheme (for $N_{MC} \rightarrow \infty$) recovers

$$\begin{cases} \partial_t \phi + \frac{1}{\delta} \omega \partial_{\mathbf{x}} \phi = \frac{1}{\delta^2} \sigma \left(\eta \int e_m - \phi \right), \\ \partial_t e_m = \frac{1}{\delta^2} \sigma \left(\int \phi - \eta e_m \right). \end{cases} \quad (10.67)$$

Performing a Hilbert development in the first equation of the above linearized system yields

$$\begin{cases} \partial_t \begin{pmatrix} 0 \\ 0 \\ \phi_0 \delta^2 \\ + \sum_1 \phi_i \delta^{i+2} \end{pmatrix} + \omega \partial_{\mathbf{x}} \begin{pmatrix} 0 \\ \phi_0 \delta \\ \phi_1 \delta^2 \\ \sum_2 \phi_i \delta^{i+1} \end{pmatrix} = \\ \sigma \left[\int \begin{pmatrix} \eta_0 e_0 \\ (\eta_0 e_1 + \eta_1 e_0) \delta \\ (\eta_0 e_2 \eta_1 e_1 + \eta_2 e_0) \delta^2 \\ \sum_{i+j > 2} \eta_i e_j \delta^{i+j} \end{pmatrix} - \begin{pmatrix} \phi_0 \\ \phi_1 \delta \\ \phi_2 \delta^2 \\ \sum_3 \phi_i \delta^i \end{pmatrix} \right]. \end{cases} \quad (10.68)$$

It leads to (we used the notations $(E_i = \int e_i, \Phi_i = \int \phi_i)_{i \in \mathbb{N}}$)

$$\begin{cases} \text{(raw results from (10.68))} & \text{(integrated results with respect to } \omega) \\ \eta_0 E_0 = \phi_0, & \eta_0 E_0 = \Phi_0, \\ \omega \partial_{\mathbf{x}} \phi_0 = \sigma(\eta_0 E_1 + \eta_1 E_0 - \phi_1), & \frac{1}{3} \partial_{\mathbf{x}} \Phi_0 = -\Phi_1, \\ \partial_t \phi_0 + \omega \partial_{\mathbf{x}} \phi_1 = \sigma(\eta_0 E_2 + \eta_1 E_1 + \eta_2 E_0 - \phi_2), & \partial_t \Phi_0 - \partial_{\mathbf{x}} \frac{1}{3\sigma} \partial_{\mathbf{x}} \Phi_0 = \sigma(\eta_0 E_2 + \eta_1 E_1 + \eta_2 E_0 - \Phi_2). \end{cases}$$

The same development of the second equation of (10.68) yields

$$\begin{cases} \partial_t \begin{pmatrix} 0 \\ 0 \\ e_0 \delta^2 \\ + \sum_1 \phi_i \delta^{i+2} \end{pmatrix} = \sigma \left[- \begin{pmatrix} \eta_0 e_0 \\ (\eta_0 e_1 + \eta_1 e_0) \delta \\ (\eta_0 e_2 \eta_1 e_1 + \eta_2 e_0) \delta^2 \\ \sum_{i+j > 2} \eta_i e_j \delta^{i+j} \end{pmatrix} + \int \begin{pmatrix} \phi_0 \\ \phi_1 \delta \\ \phi_2 \delta^2 \\ \sum_3 \phi_i \delta^i \end{pmatrix} \right]. \end{cases} \quad (10.69)$$

It leads to

$$\begin{cases} \text{(raw results from (10.69))} & \text{(integrated results with respect to } \omega) \\ \eta_0 e_0 = \int \phi_0, & \eta_0 E_0 = \Phi_0, \\ \eta_0 e_1 + \eta_1 e_0 = \int \phi_1, & \eta_0 E_1 + \eta_1 E_0 = \Phi_1, \\ \partial_t e_0 = \sigma(-\eta_0 e_2 - \eta_1 e_1 - \eta_2 e_0 + \int \phi_2), & \partial_t E_0 = \sigma(-\eta_0 E_2 - \eta_1 E_1 - \eta_2 E_0 + \Phi_2). \end{cases}$$

Some equations are redundant (hence no incompatibility) but we finally obtain the asymptotic limit for the linearized system (10.61):

$$\begin{cases} \Phi_0(\mathbf{x}, t) = \eta_0(\mathbf{x}, t) E_0(\mathbf{x}, t) = E_0^4(\mathbf{x}, t) + \mathcal{O}(\Delta t), \\ \partial_t(E_0(\mathbf{x}, t) + \Phi_0(\mathbf{x}, t)) - \partial_{\mathbf{x}} \frac{1}{3\sigma(\mathbf{x})} \partial_{\mathbf{x}} \Phi_0(\mathbf{x}, t) = 0. \end{cases} \quad (10.70)$$

In summary, an asymptotic analysis of the above MC scheme yields, to leading order, a valid (explicit or implicit, the above calculations are general enough to be independent of this choice) *conservative* discretisation of the equilibrium diffusion equation (10.70). The MC scheme does not introduce 'source sampling' as a resolution strategy and consequently does not suffer the teleportation error (avoiding potential competing discretisation parameters, no cycle-to-cycle error explosion). However, the leading order radiation intensity is not given by a planckian at the local end of time-step, indeed $\Phi_0 = \eta_0 E_0 = E_0^4 + \mathcal{O}(\Delta t) \neq E_0^4$ in general. Thus, the above method only has the *diffusion* limit, and not the *equilibrium diffusion* one, or only up to order $\mathcal{O}(\Delta t)$.

A non-conservative Asymptotic Preserving MC scheme for the *equilibrium diffusion* limit

The MC scheme presented in the previous paragraph is *conservative*, capture the *diffusion limit*, but not the *equilibrium* one. We here aim at solving (10.34) together with capturing the *equilibrium diffusion* regime (10.40) characterised by $\delta \rightarrow 0$. For this, we introduce variable $\theta(\mathbf{x}, t, \omega)$ defined by $\int \theta(\mathbf{x}, t, \omega) d\omega = \Theta(\mathbf{x}, t) = E^4(\mathbf{x}, t)$. System (10.34) can then be rewritten in term of unknowns (u, θ)

$$\begin{cases} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int \theta(\mathbf{x}, t, \omega) d\omega, \\ \partial_t \theta(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\beta(\Theta(\mathbf{x}, t)) \left(\int u(\mathbf{x}, t, \omega) d\omega - \int \theta(\mathbf{x}, t, \omega) d\omega \right). \end{cases} \quad (10.71)$$

System (10.71) is still nonlinear. On time step $[0, t]$, we consider linearisations corresponding to a particular choice of $\beta(\Theta(\mathbf{x}, t)) = 4\Theta^{3/4}(\mathbf{x}, t) = 4(\int \theta(\mathbf{x}, t, \omega) d\omega)^{3/4}$ with respect to the time variable, i.e. explicit, implicit, etc. Once again, the discussion on an explicit or implicit choice for β is not the purpose of this section. We focus on how to build an MC scheme for the resolution of system (10.71) of unknowns (u, θ) . Discretising θ with an MC approximation ensures we do not need to reconstruct a density or use source sampling and allows avoiding, by construction, the teleportation error.

Let us assume a linearisation based on a choice¹² of $\beta(\Theta(\mathbf{x}, t)) \approx \beta(\mathbf{x}, t)$ on time step $[0, t]$: the linearisation of system (10.71) of solution (ϕ, θ_m) is then given by

$$\begin{cases} \partial_t \phi(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\phi(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int \theta_m(\mathbf{x}, t, \omega) d\omega, \\ \partial_t \theta_m(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\beta(\mathbf{x}, t)\theta_m(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\beta(\mathbf{x}, t) \int \phi(\mathbf{x}, t, \omega) d\omega. \end{cases} \quad (10.72)$$

System (10.72) is now linear and can be solved with an MC scheme. To solve it, introduce

$$u(\mathbf{x}, t, v, \omega) = \phi(\mathbf{x}, t, \omega)\delta_c(v) + \theta_m(\mathbf{x}, t, \omega)\delta_0(v),$$

where v is nothing more than a velocity which can be c for photons or 0 for matter. Expression (10.72) can be rewritten (we drop the dependences for conciseness)

$$\partial_t \underbrace{(\phi\delta_c + \theta_m\delta_0)}_u + v\omega \partial_{\mathbf{x}} \underbrace{(\phi\delta_c + \theta_m\delta_0)}_u = -c\sigma(\phi\delta_c + \beta\theta_m\delta_0) + c\sigma \left(\int \theta_m\delta_c + \beta \int \phi\delta_0 \right). \quad (10.73)$$

To make sure u appears in the collisional part, introduce the same Kronecker symbols as in (10.63) and rewrite (10.73) as

$$\begin{aligned} \partial_t u(\mathbf{x}, t, v, \omega) + v\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + c\sigma(\mathbf{x})(\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v))u(\mathbf{x}, t, v, \omega) = \\ + c\sigma(\mathbf{x}) \iint [u(\mathbf{x}, t, v', \omega')\delta_{c,0}(v', v) + u(\mathbf{x}, t, v', \omega')\beta(\mathbf{x}, t)\delta_{0,c}(v', v)] dv' d\omega'. \end{aligned} \quad (10.74)$$

One can check that choosing $v = c$ in (10.74) allows recovering (10.72a) and choosing $v = 0$ in (10.74) leads to (10.72b). We can identify scattering and total cross-sections to rewrite (10.74) under the general form encountered all along chapter 9:

$$\begin{aligned} \partial_t u(\mathbf{x}, t, v, \omega) + v\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + \sigma_t(\mathbf{x}, t, v)u(\mathbf{x}, t, v, \omega) = \\ \iint \sigma_s(\mathbf{x}, t, v, v')u(\mathbf{x}, t, v', \omega')dv' d\omega'. \end{aligned} \quad (10.75)$$

In the above expression, we introduced

$$\sigma_t(\mathbf{x}, t, v) = c\sigma(\mathbf{x})(\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)), \text{ and } \sigma_s(\mathbf{x}, t, v, v') = c\sigma(\mathbf{x})(\delta_{c,0}(v', v) + \beta(\mathbf{x}, t)\delta_{0,c}(v', v)).$$

¹²explicit or implicit, this only affects the resolution, not the MC scheme.

Let us rewrite the scattering part as $\sigma_s(\mathbf{x}, t, v)P_s(\mathbf{x}, t, v, v') = \sigma_s(\mathbf{x}, t, v, v')$. This implies

$$\begin{aligned}\sigma_s(\mathbf{x}, t, v) &= \iint \sigma_s(\mathbf{x}, t, v, v') dv' d\omega', \\ &= \int c\sigma(\mathbf{x})(\delta_{c,0}(v', v) + \beta(\mathbf{x}, t)\delta_{0,c}(v', v)) dv', \\ &= c\sigma(\mathbf{x})(\delta_{c,0}(v, 0) + \beta(\mathbf{x}, t)\delta_{0,c}(v, 0)) + c\sigma(\mathbf{x})(\delta_{c,0}(v, c) + \beta(\mathbf{x}, t)\delta_{0,c}(v, c)), \\ &= c\sigma(\mathbf{x})(\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)).\end{aligned}$$

By definition of P_s we have

$$\begin{aligned}P_s(\mathbf{x}, t, v, v') &= \frac{\sigma_s(\mathbf{x}, t, v, v')}{\sigma_s(\mathbf{x}, t, v)}, \\ &= \frac{\delta_{c,0}(v', v) + \beta(\mathbf{x}, t)\delta_{0,c}(v', v)}{\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)}.\end{aligned}$$

Few calculations, similar to the already performed one before to identify P_s , show that

$$P_s(\mathbf{x}, t, v, v') = 1 - \delta_v(v'),$$

exactly as in the previous section. Now, we are interested in a direct resolution of (10.74) on time step $[0, t]$. We apply the material of section 9.5 and consider the adjoint form of (10.75) given by

$$\begin{aligned}-\partial_t u(\mathbf{x}, t, v, \omega) - v\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, v, \omega) + \sigma_t(\mathbf{x}, t, v, \omega)u(\mathbf{x}, t, v, \omega) = \\ \iint \sigma_s(\mathbf{x}, t, v)P_s(\mathbf{x}, t, v, v')u(\mathbf{x}, t, v', \omega')d\omega'dv'.\end{aligned}\tag{10.76}$$

Just as in section 9.5, we introduce

$$\sigma_S(\mathbf{x}, t, v, v') = \sigma_S(\mathbf{x}, t, v, \omega)P_S(\mathbf{x}, t, v, v') = \sigma_s(\mathbf{x}, t, v, \omega')P_s(\mathbf{x}, t, v, v').$$

It can be characterised computing

$$\begin{aligned}\sigma_S(\mathbf{x}, t, v) &= \iint \sigma_s(\mathbf{x}, t, v, v')P_s(\mathbf{x}, t, v, v')dv'd\omega', \\ &= \int c\sigma(\mathbf{x})(\delta_c(v') + \beta(\mathbf{x}, t)\delta_0(v')) \frac{\delta_{c,0}(v', v) + \beta(\mathbf{x}, t)\delta_{0,c}(v', v)}{\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)} dv', \\ &= c\sigma(\mathbf{x})\beta(\mathbf{x}, t) \frac{\delta_c(v)}{\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)} + c\sigma(\mathbf{x}) \frac{\beta(\mathbf{x}, t)\delta_0(v)}{\delta_c(v) + \beta(\mathbf{x}, t)\delta_0(v)}, \\ &= c\sigma(\mathbf{x})(\beta(\mathbf{x}, t)\delta_c(v) + \delta_0(v)).\end{aligned}$$

Furthermore, few calculations similar to the already performed one before to identify P_s , show that

$$P_S(\mathbf{x}, t, v, v') = P_s(\mathbf{x}, t, v, v') = 1 - \delta_v(v').$$

With the above expressions of the cross-sections, the non-analog MC schemes of section 9.4 for the *direct* resolution of (10.72) on time step $[0, t]$ implies

- sampling the interaction time from σ_S , i.e. $c\sigma\beta$ if the MC particle represents a photon (i.e. if $v = c$),
- or $c\sigma$ if the MC particle represents matter (i.e. if $v = 0$).

We have

$$\sigma_A(\mathbf{x}, t, v) = \sigma_t(\mathbf{x}, t, v) - \sigma_S(\mathbf{x}, t, v) = c\sigma(\mathbf{x})((1 - \beta(\mathbf{x}, t))\delta_c(v) - (1 - \beta(\mathbf{x}, t))\delta_0(v)).$$

It describes the weight modification along the flight path of any MC particles, see section 9.5. Concerning the scattering part, at each interaction, the MC particle changes of state, from photon to matter and matter to photon with probability 1.

Let us study the asymptotic properties of the above linearisation in the limit $\delta \rightarrow 0$. On time step $[0, t]$, the previously presented MC scheme (in the limit $N_{MC} \rightarrow \infty$) allows solving

$$\begin{cases} \partial_t \phi + \frac{1}{\delta} \mu \partial_x \phi = \frac{1}{\delta^2} \sigma \left(\int \theta_m - \phi \right), \\ \partial_t \theta_m = \frac{1}{\delta^2} \sigma \beta \left(\int \phi - \theta_m \right). \end{cases} \quad (10.77)$$

Performing a Hilbert development in the first equation of the above linearized system yields

$$\begin{cases} \partial_t \begin{pmatrix} 0 \\ 0 \\ \phi_0 \delta^2 \\ \sum_1 \phi_i \delta^{i+2} \end{pmatrix} + \mu \partial_x \begin{pmatrix} 0 \\ \phi_0 \delta \\ \phi_1 \delta^2 \\ \sum_2 \phi_i \delta^{i+1} \end{pmatrix} = \sigma \left[\int \begin{pmatrix} \theta_0 \\ \theta_1 \delta \\ \theta_2 \delta^2 \\ \sum_{i>2} \theta_i \delta^i \end{pmatrix} - \begin{pmatrix} \phi_0 \\ \phi_1 \delta \\ \phi_2 \delta^2 \\ \sum_3 \phi_i \delta^i \end{pmatrix} \right]. \end{cases} \quad (10.78)$$

It leads to (we used the notations $(\Theta_i = \int \theta_i, \Phi_i = \int \phi_i)_{i \in \mathbb{N}}$)

$$\begin{cases} (\text{raw results from (10.78)}) & (\text{integrated with respect to } \omega) \\ \int \theta_0 = \phi_0, & \Theta_0 = \Phi_0, \\ \mu \partial_x \phi_0 = \sigma \left(\int \theta_1 - \phi_1 \right), & \frac{1}{3} \partial_x \Phi_0 = -\sigma \Phi_1, \\ \partial_t \phi_0 + \mu \partial_x \phi_1 = \sigma \left(\int \theta_2 - \phi_2 \right), & \partial_t \Phi_0 - \partial_x \frac{1}{3\sigma} \partial_x \Phi_0 = \sigma(\Theta_2 - \Phi_2). \end{cases}$$

The same development of the second equation of (10.77) yields

$$\begin{cases} \partial_t \begin{pmatrix} 0 \\ 0 \\ \theta_0 \delta^2 \\ \sum_1 \phi_i \delta^{i+2} \end{pmatrix} = \sigma \left[- \begin{pmatrix} \beta_0 \theta_0 \\ (\beta_0 \theta_1 + \beta_1 \theta_0) \delta \\ (\beta_0 \theta_2 + \beta_1 \theta_1 + \beta_2 \theta_0) \delta^2 \\ \sum_{i+j>2} \beta_i \theta_j \delta^{i+j} \end{pmatrix} + \int \begin{pmatrix} \beta_0 \phi_0 \\ (\beta_0 \phi_1 + \beta_1 \phi_0) \delta \\ (\beta_0 \phi_2 + \beta_1 \phi_1 + \beta_2 \phi_0) \delta^2 \\ \sum_{i+j>2} \beta_j \phi_i \delta^{i+j} \end{pmatrix} \right] \end{cases} \quad (10.79)$$

It leads to

$$\begin{cases} (\text{raw results from (10.79)}) & (\text{integrated with respect to } \omega) \\ \theta_0 = \int \phi_0, & \Theta_0 = \Phi_0, \\ \beta_0 \theta_1 + \beta_1 \theta_0 = \beta_0 \Phi_1 + \beta_1 \Phi_0, & \Theta_1 = \Phi_1 \\ \partial_t \theta_0 = \sigma(-\beta_0 \Theta_2 - \beta_1 \Theta_1 - \beta_2 \Theta_0 + \beta_0 \Phi_2 + \beta_1 \Phi_1 + \beta_2 \Phi_0), & \frac{1}{\beta_0} \partial_t \Theta_0 = -\sigma(\Theta_2 - \Phi_2). \end{cases}$$

Combining the two asymptotical analysis, we finally obtain the asymptotic limit for the linearized system (10.72):

$$\begin{cases} \Phi_0 = \Theta_0 = E_0^4, \\ \partial_t (E_0 + E_0^4) - \partial_x \frac{1}{3\sigma} \partial_x E_0^4 = \mathcal{O}(\Delta t). \end{cases} \quad (10.80)$$

In summary, an asymptotic analysis of the above MC scheme yields, to leading order, a valid (explicit or implicit, the above calculations are general enough to be independent of this choice) *non-conservative* discretisation of the equilibrium diffusion equation (10.80). The MC scheme does not introduce 'source sampling' as a resolution strategy and consequently avoids teleportation error (avoiding potential competing discretisation parameters, no cycle-to-cycle error explosion). The leading order radiation intensity is given by a planckian at the local end of time-step, indeed $\Phi_0 = E_0^4$. Thus, the above method captures the *equilibrium diffusion* limit. Conservation is only ensured up to $\mathcal{O}(\Delta t)$.

Few words on how to be able to take larger time steps for the two AP linearisations

The original idea comes from X. Valentin and H. Jourdren. It consists in applying a similar methodology as in the IMC framework. Let us introduce wisely the equivalent of the Fleck factor (see section 10.2.1)

for the previous linearisation. We here sketch the idea for the Asymptotic-Preserving scheme obtained from linearisation (10.61). The equivalent for the (Asymptotic-Preserving) linearisation (10.72) can be obtained in a very similar manner and will not be presented.

The starting point is consequently (10.61) reminded here:

$$\left\{ \begin{array}{l} \partial_t \phi(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\phi(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(\mathbf{x}, t) \int e_m(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e_m(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(\mathbf{x}, t)e_m(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int \phi(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.81a)$$

$$\left\{ \begin{array}{l} \partial_t \phi(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\phi(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})\eta(\mathbf{x}, t) \int e_m(\mathbf{x}, t, \omega) d\omega, \\ \partial_t e_m(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta(\mathbf{x}, t)e_m(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x}) \int \phi(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.81b)$$

Applying the same implicitation as in the Fleck and Cummings methodology supposes choosing $\eta(\mathbf{x}, t) = \eta^{n+1}(\mathbf{x})$ and making sure the closure equation for the energy ensures conservation of the total energy of the system. Beforehand, $\eta^{n+1}(\mathbf{x})$ must be evaluated. To do so, we first identify the equation it solves (we here drop the dependences and recall no linearisation is assumed for the moment)

$$\begin{aligned} \partial_t E_m^4 &= E_m \partial_t \eta + \eta \partial_t E_m &= 4E_m^3 \partial_t E_m = 4\eta c\sigma \left(\int \phi - \eta E_m \right), \\ &= E_m \partial_t \eta + \eta c\sigma \left(\int \phi - \eta E_m \right) &= 4E_m^3 \partial_t E_m = 4\eta c\sigma \left(\int \phi - \eta E_m \right). \end{aligned} \quad (10.82)$$

We then have

$$\partial_t \eta(\mathbf{x}, t) = 3\eta(\mathbf{x}, t)c\sigma(\mathbf{x}) \left(\frac{1}{E_m(\mathbf{x}, t)} \int \phi(\mathbf{x}, t, \omega) d\omega - \eta(\mathbf{x}, t) \right). \quad (10.83)$$

Let us introduce the **explication** and **implication** hypothesis

$$\partial_t \eta(\mathbf{x}, t) = 3\eta^n(\mathbf{x})c\sigma(\mathbf{x}) \left(\frac{1}{E_m(\mathbf{x}, t)} \int \phi(\mathbf{x}, t, \omega) d\omega - \eta^{n+1}(\mathbf{x}) \right). \quad (10.84)$$

We now integrate the above equation with respect to time to write

$$\begin{aligned} \eta^{n+1}(\mathbf{x}) &= \eta^n(\mathbf{x}) + 3\eta^n(\mathbf{x})c\sigma(\mathbf{x}) \left(\int_0^t \frac{1}{E_m(\mathbf{x}, \tau)} \int \phi(\mathbf{x}, \tau, \omega) d\omega - \Delta t \eta^{n+1}(\mathbf{x}) \right), \\ &\approx \eta^n(\mathbf{x}) + 3\eta^n(\mathbf{x})c\sigma(\mathbf{x})\Delta t \left(\frac{1}{E_m(\mathbf{x}, t)} \int \phi(\mathbf{x}, t, \omega) d\omega - \eta^{n+1}(\mathbf{x}) \right). \end{aligned} \quad (10.85)$$

We finally have

$$\begin{aligned} \eta^{n+1}(\mathbf{x}) &= \eta^n(\mathbf{x}) \frac{1}{1 + 3\eta^n(\mathbf{x})c\sigma(\mathbf{x})\Delta t} + \frac{3\eta^n(\mathbf{x})c\sigma(\mathbf{x})\Delta t}{1 + 3\eta^n(\mathbf{x})c\sigma(\mathbf{x})\Delta t} \frac{1}{E_m(\mathbf{x}, t)} \int \phi(\mathbf{x}, t, \omega) d\omega, \\ &= \eta^n(\mathbf{x})V^n(\mathbf{x}, \Delta t) + (1 - V^n(\mathbf{x}, \Delta t)) \frac{1}{E_m(\mathbf{x}, t)} \int \phi(\mathbf{x}, t, \omega) d\omega. \end{aligned} \quad (10.86)$$

In the above expression, we introduce the *Valentin's factor* V^n : it plays exactly the same role as the Fleck one, except it is fitted for linearisation (10.81).

Remark 10.2 *The same remark as remark 10.1 can be made about equations (10.85) and the hypothesis made to go from its first line to its second.*

Plugging the expression of η^{n+1} in (10.81a) and introducing the second equation to satisfy conservation on the time step leads to

$$\left\{ \begin{array}{l} \partial_t \phi(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} \phi(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\phi(\mathbf{x}, t, \omega) = \\ c\sigma(\mathbf{x})\eta^n(\mathbf{x})V^n(\mathbf{x}, \Delta t) \int e_m(\mathbf{x}, t, \omega) d\omega + c\sigma(\mathbf{x})(1 - V^n(\mathbf{x}, \Delta t)) \int \phi(\mathbf{x}, t, \omega) d\omega, \end{array} \right. \quad (10.87a)$$

$$\left. \begin{array}{l} \partial_t e_m(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})\eta^n(\mathbf{x})V^n(\mathbf{x}, \Delta t)e_m(\mathbf{x}, t, \omega) = c\sigma(\mathbf{x})V^n(\mathbf{x}, \Delta t) \int \phi(\mathbf{x}, t, \omega) d\omega. \end{array} \right. \quad (10.87b)$$

The above linearized system has to be solved on time step $[0, t]$ with an MC scheme, very similar to

the one presented before (after linearisation (10.61)), except an artificial scattering is introduced. This artificial scattering is such that a radiative MC particle is not necessarily changed into a matter one. There is no additional difficulty to build an MC scheme for system (10.87), it consists in applying the material of the previous chapter 9. There is no additional difficulty to show the new linearisation bears relevant properties with respect to the *conservative equilibrium diffusion* regime. The same applies for the question of the teleportation error. The new scheme does not suffer it (as the previous analysis have been carried out independently of an explicit or implicit choice of η and β).

10.2.3 Summary

In this section, we deepened the analysis of the IMC scheme [110] together with its improved *tilted* versions [69, 70, 301]. In particular, we formally showed how a tilt behaves with respect to the regime $\delta \rightarrow 0$. Tilting ensures better numerical approximations but intrinsically makes the Δt and $\Delta \mathbf{x}$ discretisation parameters compete during each time steps. It leads to cycle-to-cycle error accumulations (see [301]). To avoid such competing behaviour, we detailed the construction of two Asymptotic Preserving MC scheme for the stiff regime (10.40). They are based on avoiding any small spatial discrepancies in the emission along the cycles of the MC resolution. They both capture the *diffusion* limit but differ with respect to the *equilibrium conservative* one. They are not subject to the teleportation error and the different discretisation parameters do not compete. This ensures converging approximations as $N_{MC} \rightarrow \infty$ together with $\Delta t \rightarrow 0$ and $\Delta \mathbf{x} \rightarrow 0$ independently.

We finally would like to mention a *third* possibility based on the few observations of this section. The above analysis shows that without the introduction of any spatial discrepancy, the IMC linearisation does capture the equilibrium diffusion limit. *Being able to design an MC scheme for the IMC linearisation avoiding this spatial discrepancy should consequently give satisfactory results.* In section (9.9.2), we detailed the construction of a new AP scheme in the stiff source regime. Its application to the IMC linearisation (10.42) also avoids the teleportation error and gives very satisfactory results in the equilibrium diffusion regime. We consider the material of both sections 9.9.2 and 10.2.1 makes its construction straightforward but we give few details on how it has been numerically tested in the following remark.

Remark 10.3 (*A third AP MC scheme in the diffusion limit based on the IMC linearisation for legacy IMC codes*) *In this remark, we focus on the limited number of modifications to apply in an IMC code to be able to apply the material of section 9.9.2 and avoid the teleportation error. Let us focus on the self-consistent equation (10.44a) obtained from the IMC linearisation (10.44):*

$$\begin{aligned} \partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma(\mathbf{x})u(\mathbf{x}, t, \omega) = \\ c\sigma(\mathbf{x})f(\mathbf{x}, \Delta t)\Theta(\mathbf{x}) + c\sigma(\mathbf{x})(1 - f(\mathbf{x}, \Delta t)) \int u(\mathbf{x}, t, \omega) d\omega. \end{aligned}$$

It can easily be recast as

$$\partial_t u(\mathbf{x}, t, \omega) + c\omega \partial_{\mathbf{x}} u(\mathbf{x}, t, \omega) + c\sigma_t(\mathbf{x})u(\mathbf{x}, t, \omega) = c\sigma_a(\mathbf{x})S(\mathbf{x}) + c\sigma_s(\mathbf{x}) \int u(\mathbf{x}, t, \omega) d\omega, \quad (10.88)$$

to fit in the notations and structure of section 9.9.2. Equation (10.88) is simpler than in section 9.9.2 in the sense the source term depends only on the spatial variable \mathbf{x} during the considered time step. It is easy, from the simulation code in which the IMC method is developed, to test the strategy of section 9.9.2. We here describe the few modifications needed:

- the weight modification along the flight path in a given cell i is given by (cf. (9.116))

$$K_i(t) = \frac{U_i(t)}{U_i(0)} = e^{-c\sigma_a^i t} + \frac{S_i}{U_i^0}(1 - e^{-c\sigma_a^i t}).$$

In the above expression, $U_i(t)$ corresponds to the analytical solution of

$$\begin{cases} \partial_t U_i(t) = -c\sigma_a^i U_i(t) + c\sigma_a^i S_i, \\ U_i(0) = U_i^0. \end{cases} \quad (10.89)$$

We assumed constant per cell quantities (opacities, sources, initial conditions) along each characteristic. The weight of each MC particles is multiplied by $K_i(t)$ between two events, the first one occurring at time 0, the second at time t (independently of their nature, census, cell exit, interaction) within cell i . In other words we have $w_p(t) = w_p(0)K_i(t)$ along the flight path of the MC particle in the time interval $[0, t]$.

- The contribution to matter in cell i occurring between times 0 and t (track length estimator) is then given by $w_p(0)(1 - K_i(t))$.
- Of course, without source term in cell i (if $S_i = 0$), we recover the classical weight modification $w_p(t) = w_p(0)e^{-c\sigma_a^i t}$ and track length estimator $w_p(t) = w_p(0)(1 - e^{-c\sigma_a^i t})$ in cell i for the non-analog MC scheme of section 9.5 with constant per cell opacities.
- We insist with this method, **source sampling** (cause of the teleportation error) **must be deactivated**.
- Finally, the sampling of the interaction time (for an MC particle in cell i) must be done according to (9.43) with

$$v\sigma_S^i(t) = c\sigma_s^i + \frac{S_i}{U_i(t)} = c\sigma_s^i + \frac{S_i}{U_i^0 e^{-c\sigma_a^i t} + S_i(1 - e^{-c\sigma_a^i t})}. \quad (10.90)$$

To sample the interaction time from an uniform random variable \mathcal{U} on $[0, 1]$, expression (9.53) must be inversed using (10.90). This inversion implies (to my knowledge) an iterative procedure (newton or fixed point). The latter can be computationally intensive especially in diffusive media (characterised by the need to compute many interaction times per MC particles). To avoid relying on such complex and costful process, we suggest approximating (10.90) by its mean on the time step $[0, \Delta t]$. In other words, we suggest performing the following additional hypothesis

$$v\sigma_S^i(t) \approx \overline{v\sigma_S^i} = \frac{1}{\Delta t} \int_0^{\Delta t} v\sigma_S^i(\alpha) d\alpha = \mathbb{E}[v\sigma_S^i(t_{\mathcal{U}_1})] \text{ where } t_{\mathcal{U}_1} \sim \mathcal{U}([0, \Delta t]). \quad (10.91)$$

To compute the above mean on-the-fly during the MC computations, we suggest applying an MC method. Just before sampling the interaction time, the new scheme relies on one more sampling:

- sample \mathcal{U}_1 from an uniform random variable on $[0, 1]$.
- Compute $t_{\mathcal{U}_1} = \mathcal{U}_1 \Delta t$ uniformly distributed in $[0, \Delta t]$.
- Use $t_{\mathcal{U}_1}$ to evaluate $v\sigma_S^i(t_{\mathcal{U}_1})$ from (10.90) (analytical formulae).
- Sample one realisation of the interaction time τ from

$$\tau = -\frac{\ln(\mathcal{U})}{v\sigma_S^i(t_{\mathcal{U}_1})} \text{ with } \mathcal{U} \sim \mathcal{U}([0, 1]).$$

- The above strategy ensures the population of MC particle, in mean, sees (10.91) as an opacity in cell i during time step $[0, \Delta t]$.

With the above modifications, it is easy implementing an IMC solver without tilt, without source sampling and without teleportation error from an already existing (legacy) IMC code. Numerical results on the classical benchmarks¹³ confirm

- the time linearisation (10.44) is efficient¹⁴,
- the loss of the AP character in the diffusion limit is mainly due to the source sampling phase¹⁵,

¹³Marshak waves for example.

¹⁴no dissuasive time steps for stability needed.

¹⁵triggering teleportation errors.

- all the more emphasized by the fact the interaction time sampling (9.43) can even be simplified to (10.91) without leading to significant errors on the numerical tests performed in the equilibrium diffusion regime.

This MC scheme has only been tackled briefly here because it relies on additional approximation (10.91) in comparison to the approaches of section 10.2.2.

Finally, we insist on the similarity of the methodology described here and the one presented in section 10.1 and [3]. System (10.89) typically refers to the reduced model in the stiff regime of interest hinted at in point 4) of section 10.1.3 to build an AP MC scheme.

About the construction of the MC scheme for the nonlinear Boltzmann equation:

- once a relevant linearisation chosen,
- it simply resumes to the application of the material of chapter 9.

The key idea is then to enrich the unknown u with an additional dimension (v in the previous sections). The MC scheme is insensitive to such increase. It then only remains to identify the resulting total and scattering opacities. Such general methodology can be applied in many other fields of applications (section 9.11 is a typical example with the uncertain linear Boltzmann equation).

Part IV

Conclusion

Chapter 11

Conclusion

The end is the beginning is the end is the ...

Contents

11.1 On Uncertainty Quantification (part II)	287
11.2 On Monte-Carlo resolution schemes (part III)	288

In this concluding chapter, we would like to come back to several aspects of the researches presented in this document. We insist on

- how they can be deepened,
- which tracks could have been followed and probably will in the future,
- and which of the previous topics lead to shorten the gap between long-term studies to *at hand* ones.

This section is brief, on purpose, to avoid speculations on the future hot topics and possible dead-ends. Just as the manuscript, the conclusion is divided into two parts, the first one dealing with uncertainty quantification, the second with Monte-Carlo schemes.

11.1 On Uncertainty Quantification (part II)

Let us begin by concluding remarks and perspectives for the uncertainty quantification topic. We mainly would like to come back on three points:

- the first point concerns the *ergocidity* property at the basis of polynomial chaos in the seminal work of Wiener [295]. Ergodicity has only been skimmed in this work, mainly because the notion is still unfamiliar to me. The literature about this mathematical and physical notion is furnished and complex and it is not easy being a self-taught newcomer in this area. Still, I feel the importance of it. Its underlying properties for modeling could greatly improve the understanding of complex physical (as in chapter 7 and paper [243]) or mathematical/numerical (as in [109]) phenomenon.
- The second point concerns mainly non-intrusive gPC¹, i.e. chapters 5, 6 and 8, and the construction of improved approximation methods having
 - a less important truncation error (see expression (5.5)),
 - in much more (non-smooth or multimodal, see the examples of chapter 8) situations.

¹and its derivations, collocation-gPC, Kriging-gPC etc., see chapter 5.

In chapter 6, a general inequality (see (6.3)) has been put forward. It has been exploited in section 6.1.1 *via* a particular choice of change of variable ensuring a gain in a new basis. This choice is not unique and there may exist some more relevant ones. For example, in section 6.1.1, we chose the change of variable $Z(X)$ such that $(u_1^Z)^2 = \sum_{k=1}^P (u_k^X)^2$ to make sure (6.3) becomes (6.5). Amongst the many other possibilities, we may be able to make sure every $(u_k^Z)_{k \in \{0, \dots, P\}}$ are related to every $(u_k^X)_{k \in \{0, \dots, P\}}$, i.e. with $\forall k \in \{0, \dots, P\}$ $u_k^Z = u_k^X \times K_k$ and $(K_k)_{k \in \{0, \dots, P\}}$ sufficiently well-suited to have inequality (6.5). Filters [162] could help finding relevant relations for example. Different choices may bear some different interesting properties.

- The last point concerns intrusive methods. For the Euler system, it is still not clear whether being intrusive is really an advantage. In the summaries of chapters 4 and 5, care has been taken to highlight the pros and cons of the two approaches in the same configuration. But the conclusions whether one approach is better than the other probably remains subjective. On another hand, in section 9.11 and above all in [241], we put forward intrusiveness can be much more efficient than non-intrusiveness for the resolution of the uncertain linear Boltzmann equation. In other words, intrusiveness, under certain conditions² helps a lot. It is sometimes worth opening the black-box code.

The last point, closely related to the MC resolution topic, will also be tackled in the next section.

11.2 On Monte-Carlo resolution schemes (part III)

Let us finish by concluding remarks and perspectives on Monte-Carlo resolution schemes. For this topic, we also would like to come back on three points:

- the first point concerns chapter 9 and the MC resolution of the linear Boltzmann equation. To my knowledge, this chapter is the most complete and furnished work available in the literature allowing to build from scratch consistent and converging MC schemes for the linear Boltzmann equation. Some parts can probably be found in different books or publications together with some more theoretical details but every practical/numerical³ ones are described in this document. The latter concern has been particularly important for me during these past (engineering) years. Every parts are originally written and we hope pedagogical enough to build simulation platforms having several MC schemes fitted for different regimes of interest. In this chapter, there are also some original (and efficient!) MC schemes⁴.
- Second, in chapter 10, we presented several Asymptotic-Preserving schemes for the resolution of the nonlinear Boltzmann equation. The scheme of section 10.1 is presented in [3]. The schemes of section 10.2 will be the purpose of future publications and will be accompanied by numerical examples for comparisons. In this chapter, the steps can be summed up as
 - first, identify a relevant linearisation with respect to the regime one wants to capture accurately. The latter must ensure taking large time steps: having in mind an MC scheme will be used to solve the linearised equation, to take advantage of *replication domain*⁵, one must make the communications at the end of time steps the scarcer possible, see [99].
 - Once the above step performed, it only remains to choose an adapted MC scheme (for example amongst the ones of chapter 9) for the now linearised equation. The identification of such MC scheme is not straightforward: for example in section 10.2.1, we put forward a very common MC discretisation (implying source sampling mainly) leading to a loss of the asymptotical equilibrium diffusion limit.

²Possibility to compute the gPC coefficients with an uncertain MC scheme for the system of interest as in section 9.11.

³For example, the accelerated Boltzmann equation is dealt with in [296, 297] but numerical aspects are eluded in the latters.

⁴The scheme of section 9.9.2 is used to solve the model described in section 10.2 for example and bear very interesting properties, see remark 10.3. The efficiency of the uncertain MC scheme of section 9.11 is demonstrated in [241].

⁵The most straightforward efficient (weak scalability) parallel strategy for MC scheme, see [99, 247, 2, 190, 187] and [99, 3, 241].

Now, regarding short term perspectives on this topic, we will probably tackle uncertainties in the nonlinear models of chapter 10. The combination of the efficient MC schemes of this chapter together with the new gPC based MC scheme of section 9.11 could lead to an almost immediate and efficient intrusive method. This will probably be the purpose of future publications.

- Finally, I would like to come back on the uncertain Euler system (tackled in part II) and the possibility to address it *via* BGK or Fokker-Planck models [193, 22, 194]. Those models can be solved with an MC scheme (see [298] for example). If a relevant linearisation of the model can be found, i.e. if we are able to take large time steps, its MC resolution could be enriched, as in section 9.11 and [241], to take efficiently uncertainties into account. This will also constitute future prospective tracks.

This ends the manuscript, hope you enjoyed it, thank you for reading.

Part V

Appendix

Appendix A

Analytical resolution of the uncertain Burgers' equation

Or what is behind the resolution of stochastic PDEs and those histograms...

The idea of this chapter is to analytically solve an uncertain propagation problem. It will allow

- identifying explicitly its different steps,
- and introducing progressively the important notions (probability measure, histogram, introduction of a Monte-Carlo resolution scheme for propagation etc.) casually used in the whole document.

In this short chapter, we focus on one of the simplest nonlinear conservation law, the Burgers' equation in 1D (spatial dimension). It corresponds to a particular choice of U and of the flux $f(U)$ in (2.1). They are given by $\forall x \in \mathcal{D}, \forall t \in [0, T]$

$$\begin{cases} \partial_t u(x, t) + \partial_x \frac{u^2(x, t)}{2} = 0, \\ u(x, 0) = u_0(x). \end{cases} \quad (\text{A.1})$$

Equation (A.1) must come with proper boundary conditions. For the moment, we focus on the previous Cauchy problem (A.1). The initial condition is given by

$$u_0(x) = u_H \mathbf{1}_{]-\infty, x_H]}(x) + \left(\frac{u_L - u_H}{x_L - x_H} (x - x_H) + u_H \right) \mathbf{1}_{[x_H, x_L]}(x) + u_L \mathbf{1}_{[x_L, \infty[}(x), \quad (\text{A.2})$$

with $x_H \neq x_L$ and $u_H \neq u_L$. Initial condition (A.2) is continuous and has two constant states, u_H for $x < x_H$ and u_L for $x_L < x$, separated by an affine part between $x \in [x_H, x_L]$ connecting state u_H to state u_L .

In a first paragraph of this chapter, we solve (A.1) together with (A.2) analytically. Equation (A.1) is deterministic. Its analytical resolution is classical, see [81, 260, 81]. It is briefly recalled in the following section. Let us define t^* as

$$t^* = -\frac{1}{\inf_{x \in \mathcal{D}} (\mathrm{d}_x u_0(x))},$$

then t^* is such that the solution is continuous for $t < t^*$ and exhibits a discontinuous behaviour for $t \geq t^*$. Given the explicit expression of the initial condition (A.2), we have $t^* = -\frac{x_L - x_H}{u_L - u_H}$, see [81, 260]. First, for $t < t^*$, the solution of (A.1) can be built applying the characteristic method [81, 260]. Let us introduce the change of variable

$$\begin{cases} \mathrm{d}_t x(t) = u(x(t), t), \\ x(0) = x_0. \end{cases}$$

Then if $\Phi(t) = u(x(t), t)$, we have $d_t \Phi(t) = \partial_t u(x(t), t) + d_t x(t) \partial_x u(x(t), t)$. Furthermore, $d_t \Phi(t) = 0$ if u is a smooth solution of (A.1). This implies $\Phi(t) = \Phi(0)$, $\forall t < t^*$. We then have

$$\Phi(t) = u(x(t), t) = \Phi(0) = u(x(0), 0) = u_0(x_0).$$

Finally, the solution $u(x, t)$ can be obtained inverting the relation $x = x_0 + u_0(x_0)t$ of unknown $x_0 = x_0(t, x)$. We do not detail the computations here. The solution for $t < t^*$ is given by

$$u(x, t) = \begin{aligned} &+u_H \mathbf{1}_{]-\infty, x_H - u_H t]}(x) \\ &+ \left(\frac{u_L - u_H}{x_L + u_L t - u_H t - x_H} (x - u_H t - x_H) + u_H \right) \mathbf{1}_{]x_H - u_H t, x_L - u_L t]}(x) \\ &+ u_L \mathbf{1}_{]-\infty, x_L - u_L t]}(x). \end{aligned} \quad (\text{A.3})$$

After t^* , the solution exhibits a discontinuous behaviour [81, 260] and can be obtained applying Rankine-Hugoniot's relation [81, 260]: let us introduce $x^*(0) = x_H + u_H t^* = x_L + u_L t^*$, $x^*(t) = x^*(0) + \frac{u_H + u_L}{2}(t - t^*)$, we then have for $t \geq t^*$

$$u(x, t) = u_H \mathbf{1}_{]-\infty, x^*(t)]}(x) + u_L \mathbf{1}_{]-\infty, x^*(t)]}(x). \quad (\text{A.4})$$

The solution of (A.1) with initial condition (A.2) is then given by¹

$$u(x, t) = \begin{aligned} &+ \mathbf{1}_{[0, t^*]}(t) \begin{bmatrix} +u_H \mathbf{1}_{]-\infty, x_H - u_H t]}(x) \\ + \left(\frac{u_L - u_H}{x_L + u_L t - u_H t - x_H} (x - u_H t - x_H) + u_H \right) \mathbf{1}_{]x_H - u_H t, x_L - u_L t]}(x) \\ + u_L \mathbf{1}_{]-\infty, x_L - u_L t]}(x) \end{bmatrix} \\ &+ \mathbf{1}_{[t^*, \infty)}(t) \begin{bmatrix} +u_H \mathbf{1}_{[-\infty, x^*(t)]}(x) \\ 0 \\ +u_L \mathbf{1}_{[x^*(t), \infty]}(x) \end{bmatrix}. \end{aligned} \quad (\text{A.5})$$

Few notations can be introduced to alleviate the above expression

$$\begin{cases} x^*(t) = x_0^* + \frac{u_H + u_L}{2}(t - t^*), \\ x_H(t) = \mathbf{1}_{[0, t^*]}(t)[x_H + u_H t] + \mathbf{1}_{[t^*, \infty)}(t)x^*(t), \\ x_L(t) = \mathbf{1}_{[0, t^*]}(t)[x_L + u_L t] + \mathbf{1}_{[t^*, \infty)}(t)x^*(t), \\ U(x, t) = \left(\frac{u_L - u_H}{x_L(t) - x_H(t)} (x - x_H(t)) + u_H \right), \text{ with } u_L \leq U(x, t) \leq u_H, \end{cases}$$

so that the solution can then be rewritten in a much simpler form

$$u(x, t) = \mathbf{1}_{[0, t^*]}(t) \begin{bmatrix} +u_H & \mathbf{1}_{]-\infty, x_H(t)]}(x) \\ +U(x, t) & \mathbf{1}_{]x_H(t), x_L(t)]}(x) \\ +u_L & \mathbf{1}_{[x_L(t), \infty]}(x) \end{bmatrix} + \mathbf{1}_{[t^*, \infty)}(t) \begin{bmatrix} +u_H & \mathbf{1}_{[-\infty, x^*(t)]}(x) \\ +u_L & \mathbf{1}_{[x^*(t), \infty]}(x) \end{bmatrix}. \quad (\text{A.6})$$

The dynamic of the *deterministic* solution is illustrated figure A.1. Each spatial profile corresponds to a particular time. The first continuous profile on the left corresponds to the initial one. As time passes,

- the slope of the affine part becomes steeper and steeper,
- the spatial interval in which it lives narrower and narrower,
- until a discontinuity forms and propagates at velocity $\frac{u_L + u_H}{2}$ (discontinuous profiles on the right hand side of abscissae $x = 2$ in figure A.1).

Such kind of solution, of low regularity, is encountered all along the present document.

In the second part of this chapter, we build an uncertainty quantification problem from (A.1) and its initial condition (A.2). The *stochastic* solution of the problem we build from the previous configuration can also be solved analytically. Suppose the solution no longer only depends on (x, t) but also explicitly on a random variable $X \sim \mathcal{G}(0, \sigma^2)$ where $\mathcal{G}(0, \sigma^2)$ denotes a gaussian of mean 0 and variance σ^2 . Here,

¹We insist it is only rewriting (A.3) and (A.4) in the same expression.

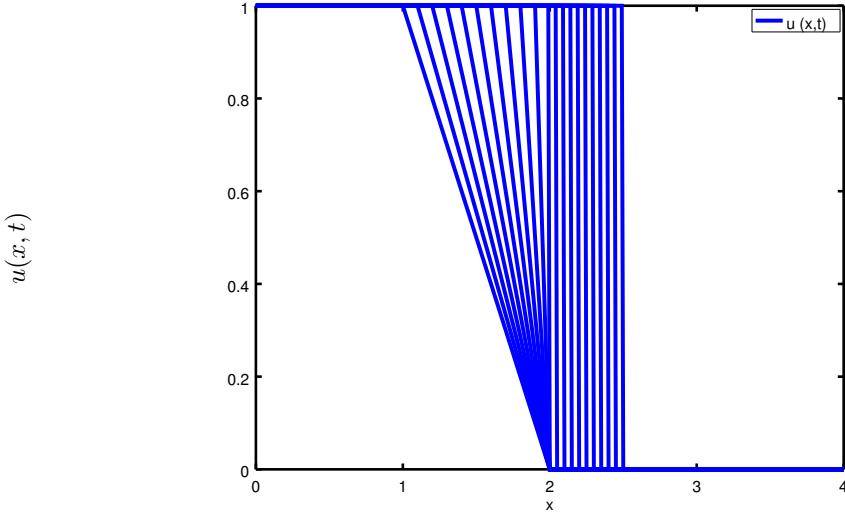


Figure A.1: Time evolution of the spatial profile for the solution of the deterministic Burgers' equation.

we suppose X models an uncertainty in the initial condition $u_0(x, X) = u_0(x - X)$. With the previous particular choice for X , the probability measure of X has expression $d\mathcal{P}_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$, with $\mu = 0$. Its cumulative density function (cdf) is given by

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right).$$

Besides, we have

$$\int_x^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = 1 - \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = 1 - F_X(x) = \frac{1}{2} \left(1 - \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right). \quad (\text{A.7})$$

The above expression will be useful later on². With the above modeling of the initial uncertainty, the uncertain Burgers' equation rewrites

$$\begin{cases} \partial_t u(x, t, X) + \partial_x \frac{u^2(x, t, X)}{2} = 0, \\ u_0(x, 0, X) = u_0(x, X) = u_0(x - X), \end{cases} \quad (\text{A.8})$$

In (A.8), we have

- $(x, t, X) \in \mathcal{D} \subset \mathbb{R} \times [0, T] \times \Omega$,
- $u_0(x, X)$ is the *uncertain* initial condition built from (A.2).

The unknown u now belongs to a probability space and the solution of the uncertainty quantification problem is a stochastic process, i.e. a random variable parameterized by both³ x and t . In this sense, solving a uncertainty quantification problem corresponds to the resolution of stochastic partial differential equations (SPDE). As a consequence, solving analytically the uncertainty quantification problem implies fully characterising the stochastic process $u(x, t, X)$, solution of the uncertain Burgers' equation (A.8). It resumes to characterising the random variable $u(x, t, X)$, $\forall(x, t)$. A random variable being fully characterised by its probability measure, we here aim at looking for $d\mathcal{P}_{u(x,t,X)}$, the measure of $X \rightarrow u(x, t, X)$. This implies looking for the probability of every admissible states of $u(x, t, X) \sim d\mathcal{P}_{u(x,t,X)}$.

²In the above expressions, the erf function is defined as $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-\frac{t^2}{2}} dt$.

³i.e. for fixed x, t , quantity $u(x, t, X)$ is a random variable.

The initial uncertainty is here modeled by a gaussian random variable $X \sim \mathcal{G}(0, \sigma^2)$. It affects the position of the affine part between the two constant states. Applying the same resolution strategy as above and taking X into account leads to a relatively simple solution of the stochastic Burgers equation (A.8). It is given by

$$u(x, t, X) = \mathbf{1}_{[0, t^*]}(t) \begin{bmatrix} +u_H & \mathbf{1}_{]-\infty, x_H(t)]}(x - X) \\ +U(x - X, t) & \mathbf{1}_{]x_H(t), x_L(t)]}(x - X) \\ +u_L & \mathbf{1}_{]x_L(t), \infty]}(x - X) \end{bmatrix} + \mathbf{1}_{[t^*, \infty[}(t) \begin{bmatrix} +u_H \mathbf{1}_{[-\infty, x^*(t)]}(x - X) \\ +u_L \mathbf{1}_{[x^*(t), \infty[}(x - X) \end{bmatrix}. \quad (\text{A.9})$$

In the particular chosen configuration, t^* does not depend on X and is still given by $t^*(X) = t^* = -\frac{x_L - x_H}{u_L - u_H}$. The maximum principle for Burgers equation [81, 260] ensures those states are within $u \in [u_L, u_H]$. The form of the solution (A.9) allows considering separately the states u_H, u_L and the interval $u \in]u_L, u_H[$:

- The probability of having $u(x, t, X) = u_H$ is denoted by $\mathbb{P}(u(x, t, X) = u_H)$ and is, by definition⁴, given by

$$\begin{aligned} \mathbb{P}(u(x, t, X) = u_H) &= \int \delta_{u_H}(u) d\mathcal{P}_{u(x, t, X)}(u) = \int_{-\infty}^{\infty} \delta_{u_H}(u(x, t, X)) d\mathcal{P}_X, \\ &= \int_{-\infty}^{\infty} \delta_{u_H}(u(x, t, y)) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} dy. \end{aligned}$$

Using expression (A.7) in the above formulae bears

$$\begin{aligned} \mathbb{P}(u(x, t, X) = u_H) &= \mathbf{1}_{[0, t^*]}(t) \int_{-\infty}^{\infty} \mathbf{1}_{]-\infty, x_H(t)]}(x - y) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} dy \\ &\quad + \mathbf{1}_{[t^*, \infty[}(t) \int_{-\infty}^{\infty} \mathbf{1}_{[-\infty, x^*(t)]}(x - y) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} dy, \\ &= \mathbf{1}_{[0, t^*]}(t) \int_{-\infty}^{\infty} \mathbf{1}_{]-\infty, x_H(t)]}(y) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2\sigma^2}} dy \\ &\quad + \mathbf{1}_{[t^*, \infty[}(t) \int_{-\infty}^{\infty} \mathbf{1}_{[-\infty, x^*(t)]}(y) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2\sigma^2}} dy, \\ &= \mathbf{1}_{[0, t^*]}(t) \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x_H(t) - x}{\sigma\sqrt{2}} \right) \right) + \mathbf{1}_{[t^*, \infty[}(t) \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x^*(t) - x}{\sigma\sqrt{2}} \right) \right), \\ &= w_{u_H}(x, t). \end{aligned}$$

In the following lines, $w_{u_H}(x, t)$ corresponds to the weight of the Dirac mass δ_{u_H} at $u(x, t, X) = u_H$.

- Very similar computations of the probability of having $u(x, t, X) = u_L$ leads to

$$\begin{aligned} \mathbb{P}(u(x, t, X) = u_L) &= \int \delta_{u_L}(u) d\mathcal{P}_{u(x, t, X)}(u), \\ &= \int_{-\infty}^{\infty} \delta_{u_L}(u(x, t, X)) d\mathcal{P}_X, \\ &= \int_{-\infty}^{\infty} \delta_{u_L}(u(x, t, y)) d\mathcal{P}_X(y), \\ &= \mathbf{1}_{[0, t^*]}(t) \int_{-\infty}^{\infty} \mathbf{1}_{[x_L(t), \infty]}(x - y) d\mathcal{P}_X(y) + \mathbf{1}_{[t^*, \infty[}(t) \int_{-\infty}^{\infty} \mathbf{1}_{[x^*(t), \infty]}(x - y) d\mathcal{P}_X(y), \\ &= \mathbf{1}_{[0, t^*]}(t) \int_{-\infty}^{\infty} \mathbf{1}_{]-\infty, x - x_L(t)]}(y) d\mathcal{P}_X(y) + \mathbf{1}_{[t^*, \infty[}(t) \int_{-\infty}^{\infty} \mathbf{1}_{]\infty, x - x^*(t)]}(y) d\mathcal{P}_X(y), \\ &= \mathbf{1}_{[0, t^*]}(t) \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x - x_L(t)}{\sigma\sqrt{2}} \right) \right) + \mathbf{1}_{[t^*, \infty[}(t) \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x - x^*(t)}{\sigma\sqrt{2}} \right) \right), \\ &= w_{u_L}(x, t). \end{aligned}$$

Once again, $w_{u_L}(x, t)$ is the weight of the Dirac mass δ_{u_L} at $u(x, t, X) = u_L$.

- The computations for the affine part are a little bit different. Let us rewrite $U(x, t) = a(t)x + b(t)$.

⁴we here only use simple probabilistic calculations, see [256].

This implicitly defines

$$\begin{cases} a(t) = \frac{u_L - u_H}{x_L(t) - x_H(t)} < 0, \\ b(t) = -\frac{u_L - u_H}{x_L(t) - x_H(t)} x_H(t) + u_H. \end{cases}$$

We consequently have $U(x - X, t) = a(t)x + b(t) - a(t)X = U(x, t) - a(t)X$. Then by definition of the probability of having $u(x, t, X)$ within interval $[u_L, u_H]$ we have

$$\begin{aligned} \mathbb{P}(u_L < u(x, t, X) < u_H) &= \mathbb{P}(u(x, t, X) = \mathbf{1}_{[x_H(t), x_L(t)]}(x - X)U(x - X, t)), \\ &= \mathbb{P}(u(x, t, X) = \mathbf{1}_{[x-x_L(t), x-x_H(t)]}(X)[U(x, t) - a(t)\sigma^2 \mathcal{G}]), \\ &= \mathbb{P}(u(x, t, X) = \mathbf{1}_{[x-x_L(t), x-x_H(t)]}(\sigma^2 \mathcal{G})(U(x, t) - a(t)\sigma^2 \mathcal{G})), \quad (\text{A.10}) \\ &= \mathbb{P}\left(u(x, t, X) = \mathbf{1}_{\left[\frac{x-x_H(t)}{\sigma^2}, \frac{x-x_L(t)}{\sigma^2}\right]}(\mathcal{G})(U(x, t) - a(t)\sigma^2 \mathcal{G})\right). \end{aligned}$$

The latter expression ensures $u(x, t, X)$ can be expressed in term of truncated gaussian law: let

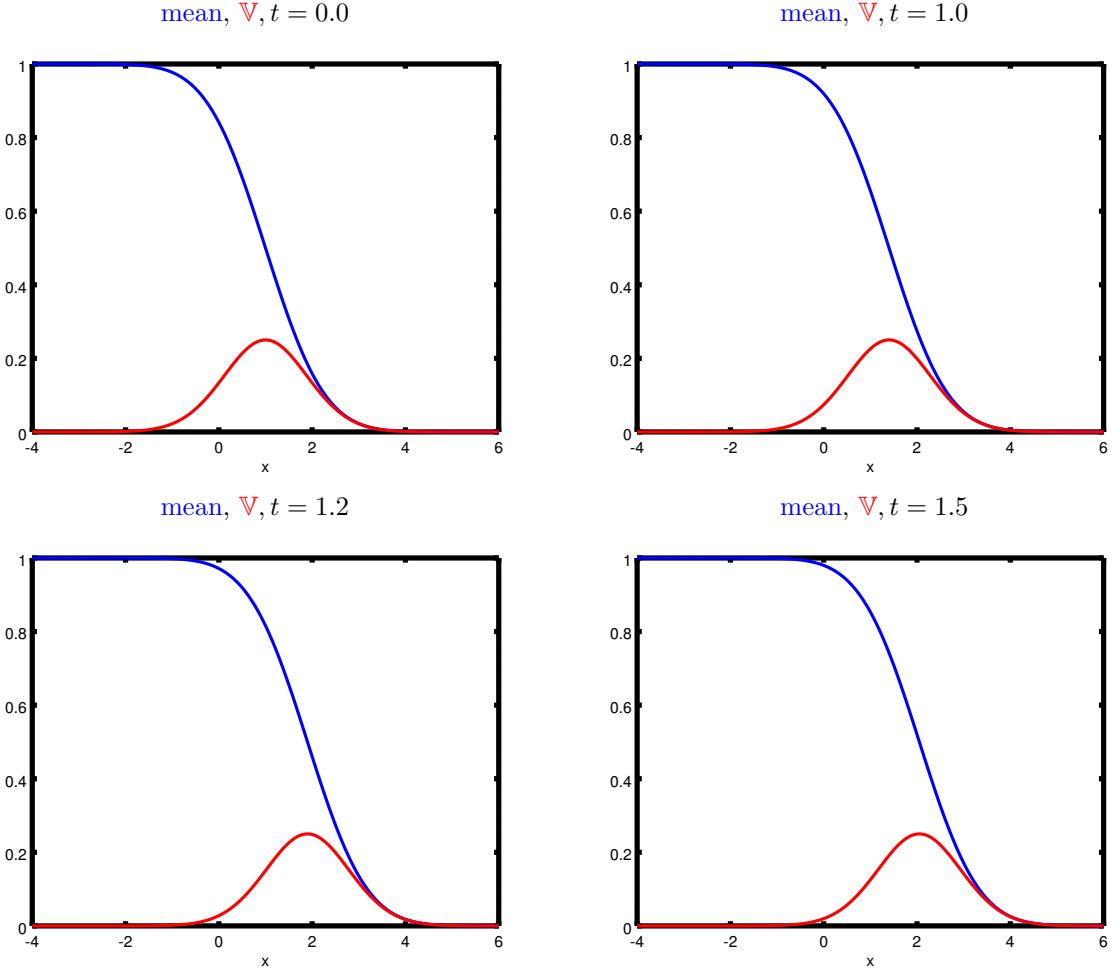


Figure A.2: Mean and Variance spatial profiles of $u(x, t, X)$ solution of the uncertain Burgers' equation (A.8) for different times.

us introduce a new gaussian variable \bar{X} of mean $U(x, t)$ and variance $a(t)\sigma^2$, then the probability measure conditioned to $u(x, t, X) \in [u_L, u_H]$ is given by

$$d\mathcal{P}_{u(x, t, X) \in [u_L, u_H]}(u) = \frac{1}{K_{L,H}} \mathbf{1}_{[U(x, t) - a(t)(x - x_H(t)), U(x, t) - a(t)(x - x_L(t))]}(u) \frac{1}{\sqrt{-2\pi a(t)\sigma^2}} e^{-\frac{(U(x, t) - u)^2}{-2a(t)\sigma^2}} du.$$

In the above expression, the normalization constant $K_{L,H}$ is simply defined such that $\forall(x,t)$ $\int d\mathcal{P}_{u(x,t,X)}(u)du = 1 - w_{u_H}(x,t) - w_{u_L}(x,t)$.

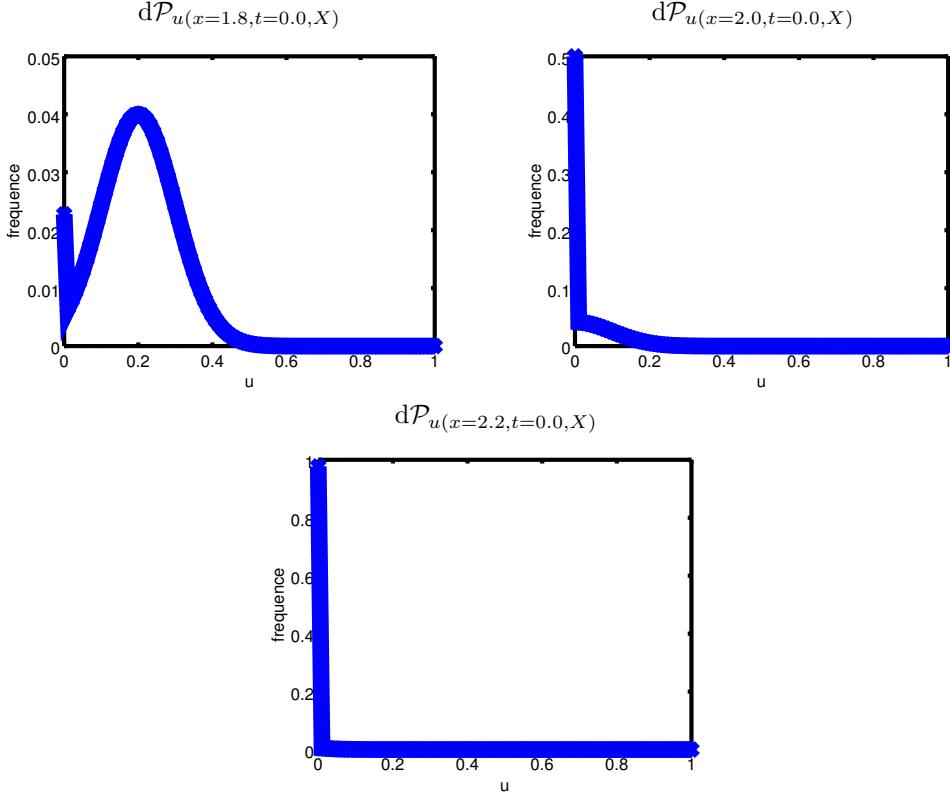


Figure A.3: Pdfs of the random variables $u(x = 1.8, t = 0.0, X)$, $u(x = 2.0, t = 0.0, X)$ and $u(x = 2.2, t = 0.0, X)$.

Finally, introduce $w_{]u_L, u_H[}(x,t) = \frac{1}{K_{L,H}} \frac{1}{\sqrt{-2\pi a(t)\sigma^2}}$, then the probability measure of $u(x,t,X)$ is fully characterised by the sum of measures

$$d\mathcal{P}_{u(x,t,X)}(u) = \begin{aligned} & +w_{u_H}(x,t)\delta_{u_H}(u) \\ & +w_{]u_L, u_H[}(x,t)\mathbf{1}_{[U(x,t)-a(t)(x-x_H(t)), U(x,t)-a(t)(x-x_L(t))]}(u)e^{-\frac{(U(x,t)-u)^2}{-2a(t)\sigma^2}}du \\ & +w_{u_L}(x,t)\delta_{u_L}(u). \end{aligned} \quad (\text{A.11})$$

Recall $w_{u_H}(x,t) + w_{]u_L, u_H[}(x,t) + w_{u_L}(x,t) = 1$ holds $\forall(x,t)$. From (A.11), every statistical observable of interest can be obtained. For example, the expressions of the high order moments $(M_k)_{k \in \mathbb{N}}$ of $u(x,t,X)$ can be calculated analytically $\forall(x,t)$ from

$$M_k(x,t) = \int u^k d\mathcal{P}_{u(x,t,X)}(u).$$

The first moment $M_1(x,t)$ (mean) and the variance $\mathbb{V}[u](x,t) = M_2(x,t) - M_1^2(x,t)$ are displayed figure A.2. We do not detail their expressions here but we display their spatial profiles for $u_L = 0, u_H = 1, x_L = 1, x_H = 2, \sigma = 1$ and for several times $t = 0.0, t = 1.0, t = 1.2$ and $t = 1.5$. We recall that for the previous choices of u_L, u_H, x_L, x_H , the critical time at which a discontinuous solution appears is $t^* = 1$, whatever the value of the realisation of X . On figure A.2, it is interesting noticing how the dynamics of the solution is hidden by the important amount of uncertainty: the slope of the mean or even the variance are only very slightly affected by the appearance of a shock (after $t^* = 1$). Mean and variance only seem to be advected and nonlinear behaviours are almost impossible to identify.

On another hand, from (A.11), it is also possible to display the probability density function (pdf)

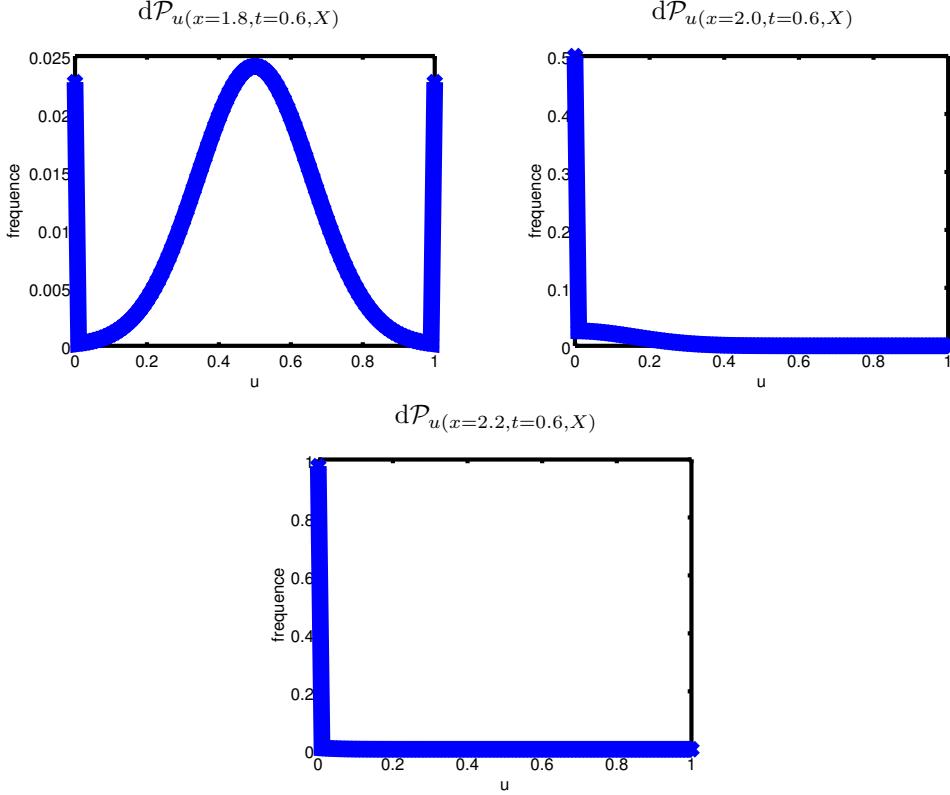


Figure A.4: Pdfs of the random variables $u(x = 1.8, t = 0.6, X)$, $u(x = 2.0, t = 0.6, X)$ and $u(x = 2.2, t = 0.6, X)$.

of $u(x, t, X)$ for several spatial locations $x = 1.8$, $x = 2.0$, $x = 2.2$ and times $t = 0.0$, $t = 0.6$, $t = 1.4$ in the same conditions (figures A.3–A.4–A.5). On figure A.3 for example, for $x = 1.8, t = 0.0$ (top left picture), the initial condition behaves as a Gaussian random variable plus a Dirac mass at $u = u_L = 0$. At $x = 2.0$ on the top right picture of figure A.3, the Dirac mass has a more important probability. At position $x = 2.2$, the solution is deterministic with state $u_L = 0$ having probability 1. As time passes, the pdf of the solution changes: in the top left picture of figure A.4, the two Dirac masses at states $u_L = 0$ and $u_H = 1$ have a non-zero probability whereas the pdfs at the other locations are only slightly affected. At time $t = 1.4$ on figure A.5, the top pictures testify of the deterministic behaviour of the solution for positions $x = 1.8$ and $x = 2.0$ (state $u_H = 1$ with probability 1) whereas the solution behaves as the binomial law (two Dirac masses only) at $x = 2.2$ in the bottom picture of the same figure. If the nonlinear behaviour of the solution was hidden when considering the mean and variance, this is not the case when considering the pdfs: the discontinuous behaviour induces the dynamical appearance of Dirac masses in the random variables $X \rightarrow u(x, t, X)$ for some x as time t increases.

Up to this point, few comments can be made: first, the complete characterization of a stochastic process, solution of an uncertainty propagation (typically the construction of (A.11)), is complex. In our case, it even strongly relies on the availability of an analytical solution of the deterministic model of interest. Obviously, such analytical solution is usually far from being available. Analytical solutions are not available anymore in this document (reason why this study is considered singular and is in the appendix). Nonetheless, before explaining how numerical solution can be handled to capture the same kind of solutions as in figures A.2–A.3–A.4–A.5, i.e. mean, variance, histograms etc., we would like to make few comments on the structure of the uncertain solutions in a hyperbolic context. With the deterministic analytical solution, we emphasized the appearance of discontinuous solutions, dynamically, as time passes. With the uncertain problem, we emphasized first, with the study of the mean and of the variance, that the solutions may seem smoothed out (figure A.2). They are not in reality, as testifies the study of the pdf of the stochastic process (figures A.4–A.5). In practice, this is of capital interest as

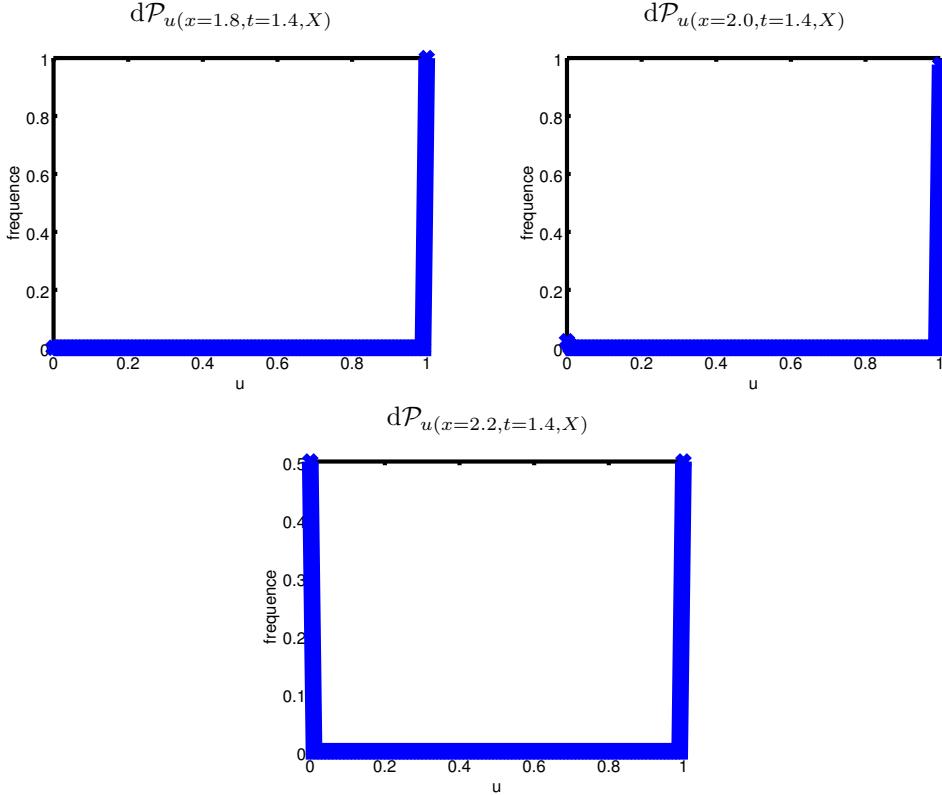
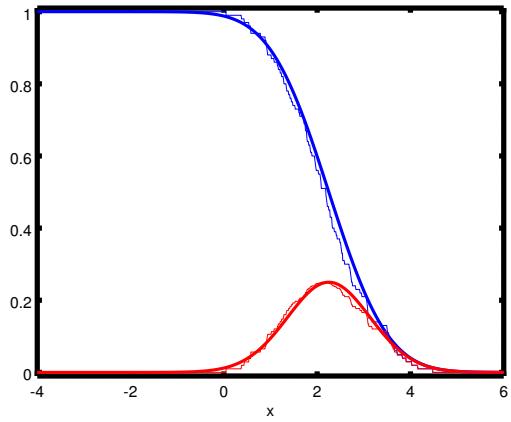


Figure A.5: Pdfs of the random variables $u(x = 1.8, t = 1.4, X)$, $u(x = 2.0, t = 1.4, X)$ and $u(x = 2.2, t = 1.4, X)$.

the discontinuous behaviour in the stochastic space typically generates important threshold effects with respect to the uncertain parameter. This justifies why, in this document, we consider mean and variance are not relevant enough quantities when dealing with uncertain systems of conservation laws.

Now, to be able to perform the same interpretations as above, there is another possibility, less calculatory, more computational: instead of calculating the probability measure (A.11) (complex process), it is possible to introduce a numerical method to compute every statistical observables of interest. Having access to the analytical solution (A.9), the idea is to apply a Monte-Carlo method, i.e. sample N_{MC} realisations of X and apply (A.9) to them before a postprocessing step. Figure A.6 compares the previous results, obtained from the analytical characterization of the probability measure, i.e. from (A.11), to Monte-Carlo approximated ones obtained from the sampling of (A.9). Figure A.6 (left) compares the results on the mean and variance spatial profiles at $t = 1.5$ for $N_{MC} = 100$ Monte-Carlo samples. The dynamics of the solution is well captured by the Monte-Carlo approximation (the smooth curves are the references) but the numerical solutions exhibit a noisy behaviour, less and less identifiable as N_{MC} is increased. On figure A.6 (right), the analytical pdf at $x = 1.8$ and $t = 0.5$ is compared to the Monte-Carlo approximated one, commonly called a histogram, obtained with $N_{MC} = 10000$ samples. Both are in good agreement, the two Dirac masses are captured together with the central part, even if more samples would be needed to, for example, accurately estimate the probability of having u within interval $[0.3, 0.4]$. Of course, when an analytical solution such as (A.9) is not available, one must rely on a simulation device which must be run for the N_{MC} realisations of the Monte-Carlo method. It usually implies dealing with longer restitution times and an additional discretization error. Those practical issues are discussed all along the present document.

mean, $\mathbb{V}, t = 1.5, N_{MC} = 100$



$d\mathcal{P}_{u(x=1.8,t=0.5,X)}, N_{MC} = 10000$

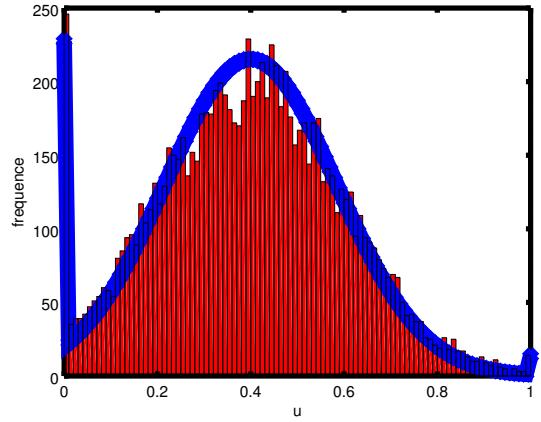


Figure A.6: Comparisons of mean, variance and pdf obtained from the analytical probability measure (A.11) and approximated from (A.9) together with a Monte-Carlo method.

Appendix B

Statistical hypothesis testing in a nutshell

An example of use of an uncertainty propagation study in a V&V context

Contents

B.1 A (too brief and general) presentation of statistical hypothesis testing . . .	300
B.2 Statistical hypothesis testing and uncertainty propagation for V&V	301

In this short appendix, we briefly present the main principle of statistical hypothesis testing. It is first generally presented in section B.1. The section, in any way, does not represent a substitute to complete works on statistics (see [256, 1] and the references therein or [250] for an interesting industrial application in neutronics). The aim of section B.1 is to enumerate the basic bricks necessary to introduce properly statistical hypothesis testing as can be needed after an uncertainty propagation. Section B.2 then corresponds to an application of the material of section B.1 on the problem of comparing efficiently and rigorously experimental and numerical results tackled in section 1.2 (more precisely when δ_X can not be made arbitrary small in section 1.2.4).

B.1 A (too brief and general) presentation of statistical hypothesis testing

Statistical hypothesis testing is of crucial importance for anyone willing to compare two data sets $(Y_i)_{i \in \{1, \dots, N_Y\}}$ and $(X_i)_{i \in \{1, \dots, N_X\}}$ obtained from two systems, two devices. In this section, we present a crude summary of its main steps. These steps will be illustrated in the next section on the problem of comparing experimental and numerical results in a V&V context (cf. the end of section 1.2.4). When one has to compare two sets of data, generally, first comes

1. one or several questions: are my two systems, from which are drawn my two data sets, equivalent? Are they fundamentally different? Of how much do they differ? Or do they have the same performances? As many questions any physician would be eager to answer when comparing experimental results to simulation ones for example.
2. Suppose we are interested in the question

$$\text{"do my systems have the same average performances?"} \quad (\text{B.1})$$

The statistical hypothesis testing framework implies stating two hypothesis. They will be confronted *via* the two data sets at hand. Those hypothesis are:

H_0 the null hypothesis. It is associated with a contradiction to a theory one would like to prove.

H_1 The alternative hypothesis. It is associated with a theory one would like to prove.

Relative to question (B.1), in general, we hope for our systems to have the same performances and do not expect any important differences. The null hypothesis would correspond to "my two systems do not have the same average performances". The alternative hypothesis would rather be "my two systems have significantly different average performances". Note that H_1 does not need to be the contraposition of H_0 .

3. The next step corresponds to the *efficient and accurate* mathematical and statistical modeling of the two hypothesis from the two data sets. This is where uncertainty propagation plays an important role. Emphasis is made on the need for accuracy and efficiency because without them, the conclusion at the end of the tests may be wrong. This modeling step will be illustrated in the next section.
4. The last step corresponds to the definition of relevant statistical tools to compare our hypothesis, their relevance, their significance. The comparison is said to be *statistically significant* if the relationship between the data sets leads to the null hypothesis having its probability of occurrence under a certain threshold probability, the significance level. The relevant statistical tools must be able to determine the probability of a rejection of the null hypothesis for an *a priori* chosen significance level. Those tools are the type I and type II errors:

Type I : the type I error/risk corresponds to the probability α of having the null hypothesis falsely rejected.

Type II: the type II error/risk corresponds to the probability β of having the null hypothesis falsely assumed to be true.

By specifying a threshold probability on the admissible risk of making a type I error, the type II error can be estimated and the statistical decision process can be controlled. One can decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 under a certain value of the type II error, and to accept or "fail to reject" the hypothesis otherwise. The strength of the test is defined by $1 - \beta$ and corresponds to the probability of rejecting H_0 when H_1 is true.

With the above lines, we quite abruptly introduced the general framework. We mainly highlighted four points. We suggest identifying those points in the context presented in section 1.2.4 regarding V&V and the use of uncertainty analysis as a tool to compare experimental and numerical results.

B.2 Statistical hypothesis testing and uncertainty propagation for V&V

In the previous sections, we presented general guidelines for statistical hypothesis testing. In this section, we focus on its application to compare rigorously experimental and numerical results. More precisely, we are going to focus on the situation where δ_X can not be made arbitrary small in section 1.2.4.

Let us come back to the problem addressed in section 1.2.4 and assume we are in situation summed up by (1.47) recalled below

$$U_{\text{exp}} - U_{\mathcal{M}^X} = \delta_0 = \underbrace{\delta_{\Delta}}_{\mathcal{O}(\Delta^\zeta) \ll 1} + \delta_{\mathcal{M}^X} - \underbrace{\delta_X}_{\ll 1} = \delta_{\mathcal{M}^X} - \delta_X. \quad (\text{B.2})$$

In the above expression, recall that (see section 1.2.4 for all the details)

- vector X models probabilistically uncertain parameters from an experimental setting.
- Scalar $U_{\text{exp}} = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} U_{\text{exp}}^i$ is the mean of N_{exp} experiments.

- Scalar $\delta_X = \frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} \delta_X^i$ is a random variable modeling probabilistically the noise in the N_{exp} experiments.
- Scalar $U_{\mathcal{M}_\Delta^X}$ is the numerical solution of model \mathcal{M}^X .
- Scalar δ_Δ is the numerical error during the resolution of \mathcal{M}^X .
- Scalar $\delta_{\mathcal{M}^X}$ is finally a flaw in the model which we would like to confirm or not or even to quantify.

Equation (B.2) expresses the fact the numerical error δ_Δ can be made arbitrary small (cf. the $\mathcal{O}(\Delta^\zeta) \ll 1$ term) but not the experimental noise δ_X (cf. the $\ll 1$ term). *Still, we would like to be able to extract some information from our set of experimental data $(U_{\text{exp}}^i, \delta_X^i)_{i \in \{1, \dots, N_{\text{exp}}\}}$ to validate or invalidate hypothesis (1.33) having some computational device at hand.* The latter computational device is typically a simulation code giving access to random variable $U_{\mathcal{M}_\Delta^X}(X)$, result of an uncertainty propagation of the uncertainty from parameter X through model \mathcal{M}^X . Note that having access to $U_{\mathcal{M}_\Delta^X}(X)$ leads to having access to a new data set $\{U_{\mathcal{M}_\Delta^X}(X_i) = U_{\mathcal{M}_\Delta^X}^i\}_{i \in \{1, \dots, N\}}$. In agreement with the notations of section 1.2.4, we decompose $U_{\mathcal{M}_\Delta^X}(X) = U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X$ as a sum of its mean $U_{\mathcal{M}_\Delta^X}$ and a fluctuation term $\tilde{\delta}_X$.

With the few lines, we wanted to integrate the problem of section 1.2.4 into the framework briefly depicted in the previous section. We have access to two data sets and we want to answer the question of the validity of model \mathcal{M}^X to represent some experiments. Let us revisit the above summed-up problem of section 1.2.4 applying the material of section B.1. Let us go through the same steps as before:

1. in section 1.2, we hinted at validating or invalidating hypothesis (1.33). The equivalent question in this context would be

$$\text{"is my model } \mathcal{M}^X \text{ relevant to represent my physical observations?"} \quad (\text{B.3})$$

Question (B.3) is probably too complex to be answered. But in general, we will be fine with having the answer of a much less ambitious one such as

$$\text{"do random variables } U_{\text{exp}} + \delta_X \text{ and } U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X \text{ have the same mean?"} \quad (\text{B.4})$$

We will consider that if they do, the answer to question (B.3) will be 'yes, model \mathcal{M}^X is relevant'. Question (B.4) will be considered in the following steps.

2. Let us now formulate some hypothesis relative to question (B.4). We are looking for a flaw in the model so hopefully, our experimental and numerical results do not have the same performances and $U_{\text{exp}} + \delta_X$ and $U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X$ do not have the same mean value. As a consequence, some relevant null and alternative hypothesis can be stated as

$$H_0 \text{ "Random variables } U_{\text{exp}} + \delta_X \text{ and } U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X \text{ have the same mean".}$$

$$H_1 \text{ "Finer experimental settings (i.e. a better control of the fluctuations of } X \text{) may lead to significant differences for the means of random variables } U_{\text{exp}} + \delta_X \text{ and } U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X \text{".}$$

Once hypothesis H_0 and H_1 stated, it remains to model them. This is when uncertainty propagation plays an important role.

3. The modeling of the two above hypothesis comes with the characterization of random variables $U_{\text{exp}} + \delta_X$ and $U_{\mathcal{M}_\Delta^X} + \tilde{\delta}_X$. Suppose their respective probability measures $d\mathcal{P}_{\delta_X}$ and $d\mathcal{P}_{\tilde{\delta}_X}$ are available:

- the characterization of $\delta_X \sim d\mathcal{P}_{\delta_X}$ may come from the application of the GUM's guidelines, see [112].
- The characterization of $\tilde{\delta}_X \sim d\mathcal{P}_{\tilde{\delta}_X}$ may come from any of the propagation methods presented in part II of this document for example.

- The question now is how can we model H_0 and H_1 from $d\mathcal{P}_{\delta_X}$ and $d\mathcal{P}_{\tilde{\delta}_X}$? Let us introduce the *statistics of the test*

$$\tilde{\delta}_{\mathcal{M}^X} = \delta_X - \tilde{\delta}_X \sim d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}. \quad (\text{B.5})$$

It is also called the *decision variable* in the litterature [256, 250]. Having access to $d\mathcal{P}_{\delta_X}$ and $d\mathcal{P}_{\tilde{\delta}_X}$ implies we have access to the probability measure¹ of $\tilde{\delta}_{\mathcal{M}^X}$, denoted by $d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}$. From $d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}$, we are going to model H_0 and H_1 .

- Under hypothesis H_1 , $\tilde{\delta}_{\mathcal{M}^X} = \delta_X - \tilde{\delta}_X \sim d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}$ has mean $U_{\text{exp}} - U_{\mathcal{M}^X_\Delta}$. As a consequence, $d\mathcal{P}_{H_1} \sim d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}$ and the modeling of H_1 is simple in this case.
- Under hypothesis H_0 , the decision variable $\tilde{\delta}_{\mathcal{M}^X} = \delta_X - \tilde{\delta}_X$ has zero mean. As a consequence, the statistical modeling of H_0 comes with a translation of the mean of $d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}$, i.e. we have $d\mathcal{P}_{H_0}(u) \sim d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}(u - (U_{\text{exp}} - U_{\mathcal{M}^X_\Delta}))$.

The definition of probability measures $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$ modeling hypothesis H_0 and H_1 ends this step. Note that more elaborate hypothesis may lead to more complex modeling but the idea remains the same.

4. The last step corresponds to the exploitation of $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$ and the computation, for example, of a type II risk β given a certain significance level (type I risk) α .
 - Assume α chosen, then the *critical region* (region in which H_0 is rejected) is defined by the interval $[U_\alpha, \infty[$ such that

$$\int \mathbf{1}_{[U_\alpha, \infty[}(u) d\mathcal{P}_{H_0}(u) = \alpha.$$

In order to determine U_α , one needs to inverse the cumulative density function of $d\mathcal{P}_{H_0}$.

- Once U_α obtained, the type II risk β can be computed:

$$\beta = \int \mathbf{1}_{]-\infty, U_\alpha]}(u) d\mathcal{P}_{H_1}(u).$$

It is obtained integrating $d\mathcal{P}_{H_1}$ on the complementary of the critical region $] -\infty, U_\alpha]$.

Risk β corresponds to the probability of falsely accepting H_0 . Risk α corresponds to the probability of falsely rejecting H_0 . As a consequence, for a given α , the smaller β is, the more efficient the test is. Note that α and β are closely related (*via* U_α): one can decide to decrease β but this will lead to an increase of α . In practice, the only way to decrease β without increasing α is to reduce the variances of probability measures $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$. In other words, it is in agreement with making sure $\delta_X \ll 1$ and reducing the fluctuations of the uncertain parameters X of the experimental setting.

To end this section, we suggest illustrating the previous purpose. Suppose the characterization of $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$ done according to the material of the above points. Assume furthermore that

- we performed some experiments and some calculations to characterize $d\mathcal{P}_{\tilde{\delta}_X}$ and $d\mathcal{P}_{\delta_X}$.
- Their difference of means gives $U_{\text{exp}} - U_{\mathcal{M}^X_\Delta} = \frac{3}{2}$.
- The convolution of $d\mathcal{P}_{\delta_X}$ and $d\mathcal{P}_{\tilde{\delta}_X}$ leads to the following expression of the decision variable's measure:

$$d\mathcal{P}_{\tilde{\delta}_{\mathcal{M}^X}}(u) = \frac{3}{4} d\mathcal{P}_{\mathcal{G}}(x, 0, \sigma_{\text{exp}}) + \frac{1}{4} d\mathcal{P}_{\mathcal{G}}(x, 4\sigma_{\text{exp}}, 2\sigma_{\text{exp}}).$$

In the above expression, $d\mathcal{P}_{\mathcal{G}}$ denotes the gaussian measure defined by

$$d\mathcal{P}_{\mathcal{G}}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

¹The probability density function (pdf) of a sum of two random variables is obtained convoluting their pdfs, see [256].

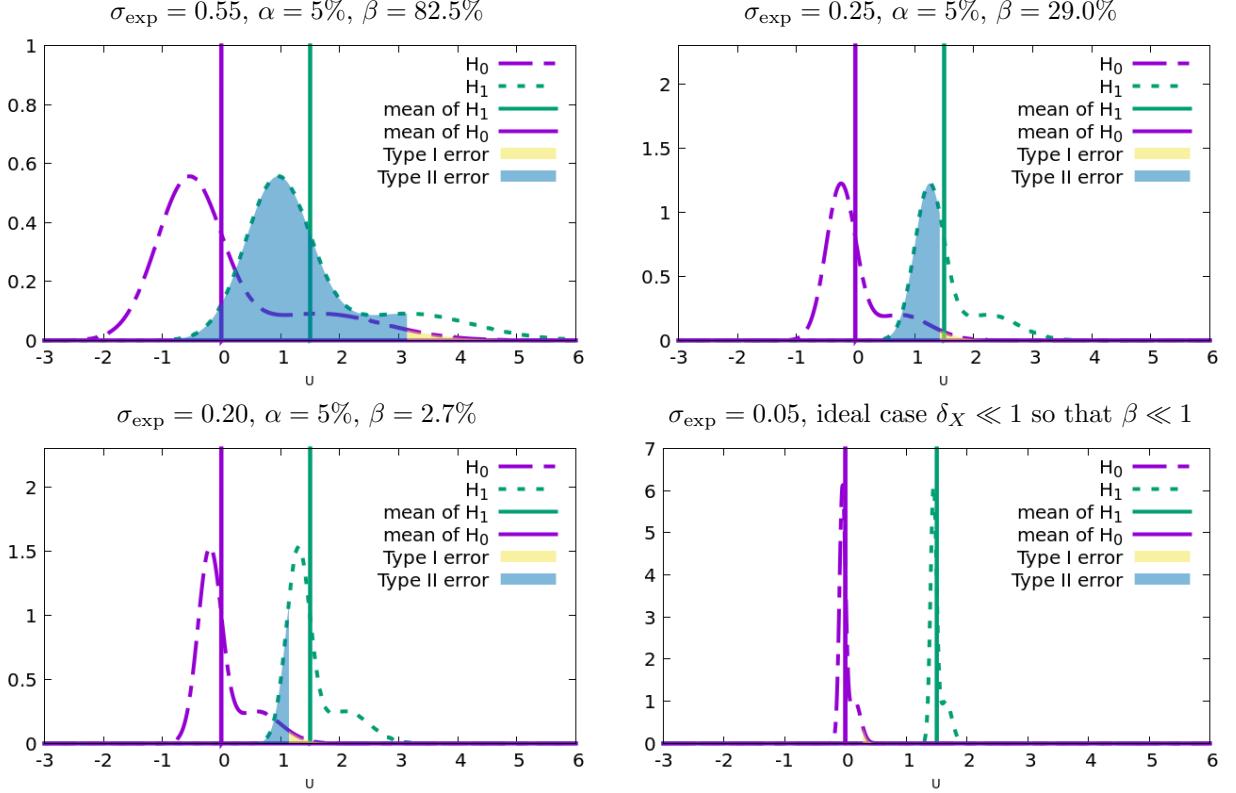


Figure B.1: The figure presents the mean of H_0 , the mean of H_1 , the distributions $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$ and the area corresponding to the type I and type II risks (α and β), for several values of the experimental variance σ_{exp} and a chosen type I error of $\alpha = 5\%$. The numerical values of the corresponding type II errors are also displayed.

In this case, the decision variable's distribution is a mixture of gaussian depending only on σ_{exp} . The choice of having a mixture of gaussian as a decision variable is here arbitrary. It makes the plots of figure B.1 fancier but will also put forward some interesting points later on.

- Now, according to the previous material and the considered hypothesis, we have:

$$\begin{aligned} d\mathcal{P}_{H_0}(u) &= d\mathcal{P}_{\tilde{\delta}_{M_X}}(u - (U_{\text{exp}} - U_{M_X})), \\ d\mathcal{P}_{H_1}(u) &= d\mathcal{P}_{\tilde{\delta}_{M_X}}(u). \end{aligned}$$

It now only remains to exploit the results.

Figure B.1 displays

- the mean of H_0 , which by hypothesis is always zero,
- the mean of H_1 , which is given by the difference $U_{\text{exp}} - U_{M_X} = \frac{3}{2}$,
- the distributions $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$,
- the area corresponding to the type I and type II risks (α and β),

for several values of the experimental variance σ_{exp} and a chosen type I error of $\alpha = 5\%$. The numerical values of the corresponding type II errors are also displayed above each picture.

Figure B.1 (top-left) presents the results of the statistical hypothesis test for $\sigma_{\text{exp}} = 0.55$ and $\alpha = 5\%$: the type II error is $\beta = 82.5\%$, see the important area under the curve $d\mathcal{P}_{H_1}$. The risk of accepting falsely H_0 , i.e. that the means of the experimental and numerical results are the same, is important. For

such level of experimental noise σ_{exp} , one can not either reject hypothesis H_0 . Thanks to the quantified risk β , one can decide to work on the reduction of the experimental noise (by better controlling the fluctuations of parameters X for example). Figure B.1 (top-right) presents the same curves but with $\sigma_{\text{exp}} = 0.25$: the type II risk has considerably decreased to $\beta = 29.0\%$. Note also that the (arbitrary) choice of having $d\mathcal{P}_{H_0}$ and $d\mathcal{P}_{H_1}$ based on a mixture of gaussian allows having distributions for which the means are different from the maximum likelihood. With this remark, we wanted to put forward the fact that hypothesis H_0 and H_1 must be stated according to the relevant statistical observable, their choice is crucial to fully harness the potential of statistical hypothesis testing. With such asymmetrical distributions, we may have had to work with quantiles rather than with the mean. Now, by decreasing the experimental noise to $\sigma_{\text{exp}} = 0.20$ (figure B.1 bottom-left), i.e. of only 20% with respect to the previous plot, a drastic improvement is made with respect to the type II risk as it drops from $\beta = 29.0\%$ to $\beta = 2.7\%$. Finally, by once again decreasing the noise to $\sigma_{\text{exp}} = 0.05$ (figure B.1 bottom-right), we can verify we recover the ideal case where $\delta_X \ll 1$, tackled in section 1.2.4, leading to having $\beta \ll 1$. In a way, such methodology only generalizes the ideal one, to be able to deal with situations where δ_X can not be made arbitrary small and help decide whether model \mathcal{M}^X is relevant enough or if some new studies must be carried out.

Bibliography

- [1] Statistical hypothesis testing. https://en.wikipedia.org/wiki/Statistical_hypothesis_testing. Accessed: 2018-02-08. pages 19, 301
- [2] Brunner T. A. and P. S. Brantley. An efficient, robust, domain-decomposition algorithm for particle Monte Carlo. *Journal of Computational Physics*, 228(10):3882–3890, 2009. pages 289
- [3] A. Bernede and G. Poëtte. An Unsplit Monte-Carlo solver for the resolution of the linear Boltzmann equation coupled to (stiff) Bateman equations. *Journal of Computational Physics*, 354:211 – 241, 2018. pages 17, 20, 181, 183, 184, 185, 187, 195, 211, 212, 213, 255, 256, 259, 261, 262, 263, 264, 265, 266, 267, 276, 286, 289
- [4] R. Abgrall. A Simple, Flexible and Generic Deterministic Approach to Uncertainty Quantifications in Non Linear Problems: Application to Fluid Flow Problems. *Rapport de Recherche INRIA*, 2007. pages 49
- [5] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964. pages 9, 31, 42, 82
- [6] S. Acharya and N. Zabaras. Uncertainty Propagation in Finite Deformations - A Spectral Stochastic Lagrangian Approach. *Comp. Meth. Appl. Mech. Engrg.*, 195:2289–2312, 2006. pages 31
- [7] N. I. Akhiezer. *The Classical Moment Problem*. Oliver and Boyd, 1965. pages 44, 45, 46, 79, 114, 115, 116
- [8] Graham W. Alldredge, Cory D. Hauck, and André L. Tits. High-order entropy-based closures for linear transport in slab geometry II: A computational study of the optimization problem. *SIAM J. Scientific Computing*, 34(4), 2012. pages 163
- [9] J. Anderson, L. Pal, and I. Pazsit. On the Feynman-alpha formula for fast neutrons. *ArXiv e-prints*, May 2011. pages 191
- [10] R. Dautray annd J.L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques*, volume tome 1-6. Masson, 1984. pages 165
- [11] C. Arhens and E. Larsen. A 'semi-analog' monte-carlo method for grey radiative transfer problems. Salt Lake City, Utah: American Nuclear Society, 2001. pages 276
- [12] R. Askey and J. Wilson. Some Basic Hypergeometric Polynomials that Generalize Jacobi Polynomials. *Memoirs Amer. Math. Soc., AMS, Providence RI*, 319, 1985. pages 40
- [13] American Society of Mechanical Engineers ASME V&V 20-2009. Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer. *ASME*, 2009. pages 13, 180, 195
- [14] G.A. Athanassoulis and P.N. Gavriilidis. The truncated hausdorff moment problem solved by using kernel density functions. *Probabilistic Engineering Mechanics*, 17(3):273–291, 2002. doi:10.1016/S0266-8920(02)00012-7. pages 46, 115, 116
- [15] A. Atkinson, A. Donev, and R. Tobias. *Optimum Experimental Designs, With SAS*. Oxford Statistical Science Series. OUP Oxford, 2007. pages 75, 84, 85, 89

- [16] Christian Aussourd. Styx: a multidimensional AMR S_N scheme. *Nuclear science and engineering*, 143(3):281–290, 2003. pages 11, 73, 163
- [17] Morillon B. *Méthode de Monte Carlo non analogue, application à la simulation des neutrons*. Ph. d. thesis, Université de Paris 11, Orsay, FRANCE (Université de soutenance), 1995. pages 234, 247, 250, 251
- [18] J. P. Babuel-Peyrissac. *Équations cinétiques des fluides et des plasmas*. Cours et documents de mathématiques et de physique. Gordon and Breach, 1974. pages 4, 5, 163, 189, 191, 216
- [19] F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *ArXiv e-prints*, January 2013. pages 93, 94
- [20] F. Bachoc. Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. *ArXiv e-prints*, December 2014. pages 94
- [21] François Bachoc. *Estimation paramétrique de la fonction de covariance dans le modèle de Krigeage par processus Gaussiens : application à la quantification des incertitudes en simulation numérique*. PhD thesis, 2013. Thèse de doctorat dirigée par Garnier, Josselin Mathématiques appliquées Paris 7 2013. pages 92, 93, 94
- [22] Céline Baranger, Gentien Marois, Jordane Mathe, Julien Mathiaud, and Luc Mieussens. A BGK model for polyatomic gas flows at high temperature. In *2017 SIAM Conference on Analysis of Partial Differential Equations*, Baltimore, United States, December 2017. pages 12, 290
- [23] A. Barth, Ch. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011. pages 233
- [24] R Barthelmé and E Sonnendrücker. *Le problème de conservation de la charge dans le couplage des équations de Vlasov et de Maxwell*. PhD thesis, Strasbourg, Strasbourg, 2005. Presented on 8 Jul 2005. pages 163, 214, 216, 275
- [25] C. Benedetti, A. Sgattoni, G. Turchetti, and P. Londrillo. **ALaDyn**: A High-Accuracy PIC Code for the Maxwell Vlasov Equations. *IEEE Transactions on Plasma Science*, 36(4):1790–1798, Aug 2008. pages 163, 214, 216, 275
- [26] M. Berveiller, B. Sudret, and M. Lemaire. Stochastic Finite Element: a Non Intrusive Approach by Regression. *Rev. Eur. Méc. Num.*, 15(1-2-3):81–92, 2006. pages 84
- [27] P.L. Bhatnagar, E.P. Gross, and M. Krook. A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems. *Physical Review*, 94:511–525, 1954. pages 12, 163
- [28] G.A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Number vol. 1 in Molecular Gas Dynamics and the Direct Simulation of Gas Flows. Clarendon Press, 1994. pages 229, 230
- [29] Charles K. Birdsall and A. Bruce Langdon. *Plasma physics via computer simulation*. the Adam Hilger series on plasma physics, 1991. Bristol, Philadelphia and New York. pages 163, 214, 216, 275
- [30] Alexandre Birolleau. *Résolution de problème inverse et propagation d'incertitudes: application à la dynamique des gaz compressibles*. PhD thesis, Paris 6, 2014. pages 15, 104, 112, 128, 130, 132, 136, 138
- [31] Alexandre Birolleau, Gaël Poëtte, and Didier Lucor. Adaptive bayesian inference for discontinuous inverse problems, application to hyperbolic conservation laws. *Communications in Computational Physics*, 16(1):134, 2014. pages 15, 19, 42, 74, 128, 130, 132, 138, 140, 145, 148, 152
- [32] M. Bisi and L. Desvillettes. From reactive boltzmann equations to reaction-diffusion systems. *Journal of Statistical Physics*, 124(2):881–912, 2006. pages 4

- [33] X. Blanc, C. Bordin, G. Kluth, and G. Samba. Variance reduction method for particle transport equation in spherical geometry. *Journal of Computational Physics*, 364:274 – 297, 2018. pages 233, 247
- [34] G. Blatman. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. Thèse de doctorat, Université Blaise Pascal - Clermont II, 2009. pages 48, 82, 237
- [35] G. Blatman and B. Sudret. Sparse Polynomial Chaos Expansions and Adaptive Stochastic Finite Elements using a Regression Approach. *C. R. Méc.*, 336:518–523, 2008. pages 48, 53, 84
- [36] G. Blatman, B. Sudret, and M. Berveiller. Quasi-Random Numbers in Stochastic Finite Element Analysis. *Mech. Ind. proofs*, 8:289–297, 2007. pages 53, 156
- [37] A. V. Bobylev and K. Nanbu. Theory of collision algorithms for gases and plasmas based on the boltzmann equation and the landau-fokker-planck equation. *Phys. Rev. E*, 61:4576–4586, Apr 2000. pages 229, 230
- [38] K. Bogues, C. R. Morrow, and T. N. L. Patterson. An Implementation of the Method of Er-makov and Zolothukin for Multidimensional Integration and Interpolation. *Numerische Mathe-matik*, 37:49–60, 1981. pages 234, 237
- [39] C. Boudesocque-Dubois and J.-M. Clarisse. Investigation of Linear Perturbation Growth in a Planar Ablation Flow. In *ECLIM: 27th European Conference on Laser Interaction with Matter*. SPIE, 2002. (to appear). pages 123, 130
- [40] S. Boyaval, C. LeBris, T. Lelièvre, Y. Maday, N. Nguyen, and A. Patera. Reduced Basis techniques for stochastic problems. *Arch. Comput. Meth. Eng.*, 17:435–454, 2012. pages 253
- [41] Sébastien Boyaval and Tony Lelièvre. Variance reduction method for parametrized stochastic differential equations using the reduced basis paradigm. *Commun. Math. Sci.*, 8(3):735–762, 2010. pages 253
- [42] J.P. Boyd. *Chebyshev and Fourier Spectral Methods: Second Revised Edition*. Dover Books on Mathematics. Dover Publications, 2001. pages 38, 41
- [43] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010. pages 53
- [44] Marc Briant and Esther S. Daus. The boltzmann equation for a multi-species mixture close to global equilibrium. *Archive for Rational Mechanics and Analysis*, 222(3):1367–1443, 2016. pages 5, 6
- [45] Eugene D. Brooks, III, Michael Scott McKinley, Frank Daffin, and Abraham Szöke. Symbolic implicit monte carlo radiation transport in the difference formulation: a piecewise constant discretization. *J. Comput. Phys.*, 205:737–754, May 2005. pages 253
- [46] S. Brooks, A. Gelman, G. Jones, and X.L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2011. pages 185
- [47] Thomas A Brunner. Forms of approximate radiation transport. *Sandia report*, 2002. pages 163
- [48] Christophe Buet and Bruno Després. Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics, 2003. pages 9, 14, 269
- [49] Christophe Buet, Bruno Després, and Emmanuel Franck. An asymptotic preserving scheme with the maximum principle for the m1 model on distorted meshes. *Comptes Rendus Mathematique*, 350(11):633 – 638, 2012. pages 213
- [50] Christophe Buet, Bruno Després, and Emmanuel Franck. Design of asymptotic preserving fi-nite volume schemes for the hyperbolic heat equation on unstructured meshes. *Numer. Math.*, 122(2):227–278, October 2012. pages 213

- [51] Christophe Buet, Bruno Després, and Emmanuel Franck. Asymptotic preserving schemes on distorted meshes for friedrichs systems with stiff relaxation: application to angular models in linear transport. *Journal of Scientific Computing*, 62(2):371–398, 2015. pages 213
- [52] Cacucci and Dan G. *Handbook of Nuclear Engineering*, volume 1. Springer Science & Business Media, 2010. pages 163, 164, 190, 229, 257
- [53] Russel E Caflisch. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49, 1998. pages 77
- [54] Russel E Caflisch, William J Morokoff, and Art B Owen. Valuation of mortgage backed securities using brownian bridges to reduce effective dimension, 1997. pages 77
- [55] R.H. Cameron and W.T. Martin. The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals. *Annals of Math.*, 48:385–392, 1947. pages v, 30, 31, 32, 34, 35, 42, 47, 114, 160
- [56] Thomas Camminady, Martin Frank, Kerstin Kpper, and Jonas Kusch. Ray effect mitigation for the discrete ordinates method through quadrature rotation. *Journal of Computational Physics*, 382:105 – 123, 2019. pages 11, 163
- [57] LL Carter and CA Forest. Nonlinear radiation transport simulation with an implicit monte carlo method. *LA-5038, Los Alamos National Laboratory*, 1973. pages 270
- [58] Julien Cartier. *Mixed-hybrid finite element method for the transport equation;br / β and diffusion approximation of transport problems*. Theses, Université d’Orléans, April 2006. pages 165
- [59] J. Castor. *Radiation hydrodynamics*. Cambridge University Press, 2004. pages 14, 163, 213, 214, 215, 216, 218, 220, 267, 268, 269
- [60] C. Cercignani. *Theory and Application of the Boltzmann Equation*. Scottish Academic Press, 1975. pages 4
- [61] F. Chaland and G. Samba. Discrete ordinates method for the transport equation preserving one-dimensional spherical symmetry in two-dimensional cylindrical geometry. *Nuclear Science and Engineering*, 182(4):417–434, 2016. pages 11, 73, 163
- [62] S. Chapman and T.G. Cowling. *The mathematical theory of non-uniform gases: An account of the kinetic theory of viscosity, thermal conduction, and diffusion in gases*. Cambridge University Press, 1960. pages 8
- [63] G. Chen, C. Levermore, and T. Liu. Hyperbolic Conservation Laws with Stiff Relaxation Terms and Entropy. *Comm. Pure Appl. Math.*, 47:787–830, 1994. pages 66, 115
- [64] E.W. Cheney and W.A. Light. *A Course in Approximation Theory*. Graduate studies in mathematics. American Mathematical Soc. pages 90
- [65] A. J. Chorin. Hermite Expansions in Monte Carlo Computations. *J. Comput. Phys.*, 8:472–482, 1971. pages 232, 234
- [66] A.J. Chorin. Hermite Expansions in Monte Carlo Computation. *J. Comp. Phys.*, 8:472–482, 1971. pages 31, 40, 234
- [67] A.J. Chorin. Gaussian Fields and Random Flow. *J. Fluid. Mech.*, 63:21–32, 1974. pages 31, 40
- [68] J.-M. Clarisse, S. Jaouen, and P.-A. Raviart. A Godunov-Type Method in Lagrangian Coordinates for Computing Linearly-Perturbed Planar-Symmetric Flows of Gas Dynamics. *J. Comp. Phys.*, 198:80–105, 2004. pages 123, 130
- [69] Mathew A. Cleveland and Nick Gentile. Mitigating teleportation error in frequency-dependent hybrid implicit monte carlo diffusion methods. *Journal of Computational and Theoretical Transport*, 43(1-7):6–37, 2014. pages 210, 272, 274, 284

- [70] J.-F. Clouet and G. Samba. Asymptotic diffusion limit of the symbolic monte-carlo method for the transport equation. *Journal of Computational Physics*, 195(1):293 – 319, 2004. pages 210, 272, 274, 284
- [71] Mireille Coste-Delclaux, Cheikh DIOP, Anne NICOLAS, and Bernard BONIN. *Neutronique*. E-dén, Une monographie de la Direction de l'énergie nucléaire. CEA Saclay; Groupe Moniteur, June 2013. pages 186
- [72] T. Crestaux. Polynômes de Chaos pour la Propagation et la Quantification d'Incertitudes. Technical report, CEA, 2006. pages 48, 81, 82
- [73] M. Dahmani. *Résolution des équations de la cinétique des réacteurs par la méthode nodale mixte duale utilisant le modèle quasi-statique amélioré et implémentation dans le code CRONOS*. PhD thesis, Faculté des Sciences de Rabat, Maroc, 1999. pages 211, 212, 213
- [74] Gautier Dakin. *Couplage fluide-structure d'ordre (très) élevé pour des schémas volumes finis 2D Lagrange-projection*. Theses, Université Pierre et Marie Curie - Paris VI, November 2017. pages 16, 17
- [75] M. K. Deb, I. M. Babuska, and J. T. Oden. Solution of Stochastic Partial Differential Equations using Galerkin Finite Element Techniques. *Comp. Meth. Appl. Mech. Engrg.*, 190:6359–6372, 2001. pages 53
- [76] Bert J. Debusschere, Habib N. Najm, Philippe P. Pébay, Omar M. Knio, Roger G. Ghanem, and Olivier P. Le Maître. Numerical Challenges in the Use of Polynomial Chaos Representations for Stochastic Processes. *J. Sci. Comp.*, 26:698–719, 2004. pages 40, 54, 56
- [77] J. D. Densmore and E. W. Larsen. Asymptotic equilibrium diffusion analysis of time-dependent Monte Carlo methods for grey radiative transfer. *Journal of Computational Physics*, 199:175–204, September 2004. pages 256, 260, 269, 270, 271, 272
- [78] B. Després. Inégalités entropiques pour un solveur de type Lagrange + convection des équations de l'hydrodynamique. Technical Report 2822, CEA, 1997. (in French). pages 9
- [79] B. Després. Lagrangian systems of conservation laws. Invariance properties of Lagrangian systems of conservation laws, approximate Riemann solvers and the entropy condition. *Numer. Math.*, 89:99–134, 2001. pages 9
- [80] B. Després and P. Dossantos-Uzarralde. Conférence "Incertitudes et Simulations". In département de physique théorique et appliquée CEA/DAM Ile de France, Département sciences de la simulation et de l'information, editor, *Conférence "Incertitudes et simulations"*, 2007. pages 56, 71
- [81] Bruno Després. *Numerical Methods for Eulerian and Lagrangian Conservation Laws*. Springer, Birkhauser, 2017. pages 14, 25, 51, 52, 68, 125, 292, 293, 295
- [82] Bruno Després. Polynomials with bounds and numerical approximation. *Numerical Algorithms*, pages 1–31, 2017. pages 122
- [83] Bruno Després and Benoît Perthame. Uncertainty propagation;intrusive kinetic formulations of scalar conservation laws. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):980–1013, 2016. pages 71, 72
- [84] Bruno Després, Gaël Poëtte, and Didier Lucor. *Robust Uncertainty Propagation in Systems of Conservation Laws with the Entropy Closure Method*, volume 92 of *Lecture Notes in Computational Science and Engineering*. Uncertainty Quantification in Computational Fluid Dynamics, 2013. pages 19, 57, 59, 60, 61, 62, 63, 67, 68, 71
- [85] Bruno Després and Emmanuel Trélat. TWO-SIDED SPACE-TIME L 1 APPROXIMATION AND OPTIMAL CONTROL OF POLYNOMIAL SYSTEMS. working paper or preprint, 2017. pages 32

- [86] S. Destercke, D. Dubois, and E. Chojnacki. Unifying Practical Uncertainty Representations: I. Generalized p-boxes. *Int. J. Appr. Reas.*, 49:649–663, 2008. pages 15, 26
- [87] Sébastien Destercke, Didier Dubois, and Eric Chojnacki. Possibilistic information fusion using maximal coherent subsets. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), London (UK), 23/07/2007-26/07/2007*, pages 445–450, <http://www.ieee.org/>, 2007. IEEE. pages 15, 26
- [88] Sébastien Destercke, Didier Dubois, and Eric Chojnacki. Possibilistic Information Fusion Using Maximal Coherent Subsets. *IEEE Transactions on Fuzzy Systems*, 17(1):79–92, 2009. pages 15, 26
- [89] L. Desvillettes, R. Monaco, and F. Salvarani. A kinetic model allowing to obtain the energy law of polytropic gases in the presence of chemical reactions. *Eur. J. Mech. B Fluids*, 24(2):21923, 2005. pages 5
- [90] H. Dette and W.J. Studden. *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis*. A Wiley-Interscience publication. Wiley, 1997. pages 84, 89
- [91] R. J. DiPerna and P.-L. Lions. On the cauchy problem for boltzmann equations: global existence and weak stability. *Ann. of Math.*, 2(130):321366, 1989. 2. pages 6
- [92] Qiang Du, Max D Gunzburger, and Lili Ju. Voronoi-based finite volume methods, optimal voronoi meshes, and pdes on the sphere. *Computer methods in applied mechanics and engineering*, 192(35):3933–3957, 2003. pages 11, 73, 163
- [93] B. Dubroca, P. Charrier, G. Duffa, and R. Turpault. Multigroup Model for Radiating Flows during Atmospheric Hypersonic Re-Entry. *Proc. of Inter. Workshop on radiation of high temperature gases in atmospheric entry*, 338:103–110, 2003. pages 163
- [94] B. Dubroca and J. Feugeas. Etude théorique et numérique d'une hiérarchie de modèles aux moments pour le transfert radiatif. *Academie des Sciences Paris Comptes Rendus Serie Sciences Mathématiques*, 329:915–920, November 1999. pages 12, 163
- [95] J. Dufek and W. Gudowski. Stochastic approximation for Monte-Carlo Calculation of Steady-State Conditions in Thermal Reactors. *Nucl. Sci. Ener.*, 152:274–283, 2006. pages 183, 255, 256, 261
- [96] J. Dufek and V. Valtavirta. Time step length versus efficiency of Monte-Carlo burnup calculations. *Preprint submitted to Annals of Nuclear Energy*, 2014. pages 183, 256, 261
- [97] Jan Dufek, Dan Kotlyar, and Eugene Shwageraus. Numerical stability of the predictor-corrector method in Monte Carlo burnup calculations of critical reactors. *Annals of Nuclear Energy*, 56:34–38, 2013. pages 183, 256, 261
- [98] Jan Dufek, Dan Kotlyar, Eugene Shwageraus, and Jaakkko Leppänen. The stochastic implicit Euler method: A stable coupling scheme for Monte Carlo burnup calculations. *Annals of Nuclear Energy*, 60:295–300, 2013. pages 183, 255, 256, 261
- [99] D. Dureau and G. Poëtte. Hybrid Parallel Programming Models for AMR Neutron Monte-Carlo Transport. In *Joint International Conference on Supercomputing in Nuclear Applications + Monte-Carlo*, number 04202 in Parallelism and HPC, Monte-Carlo, 2013. pages 17, 20, 207, 289
- [100] A Dutfoy and R Lebrun. Practical approach to dependence modelling using copulas. *Journal of Risk and Reliability*, 223(4):347–361, 2009. pages 48, 222
- [101] C. Enaux. *Analyse Mathématique et Numérique d'un Modèle Multifluide Multivitesse pour l'Interpénétration de Fluides Miscibles*. Thèse de doctorat, École Centrale Paris, 2007. pages 26
- [102] James F. Epperson. On the runge example. *The American Mathematical Monthly*, 94(4):329–341, 1987. pages 90

- [103] S. M. Ermakov and V.G. Zolotukhin. Polynomial Approximations and the Monte Carlo Method. *Teor. Veroyatnost.*, 5:473–476, 1960. pages 40, 232, 234, 237
- [104] Ernst, Oliver G., Mugler, Antje, Starkloff, Hans-Jörg, and Ullmann, Elisabeth. On the convergence of generalized polynomial chaos expansions. *ESAIM: M2AN*, 46(2):317–339, 2012. pages 40, 105
- [105] Donald Estep. Practical analysis in one variable, 2002. pages 32
- [106] R. Eymard, T. Gallouet, and R. Herbin. *Finite Volume Methods, in Handbook of Numerical Analysis*, volume 7. P.G. Ciarlet, J. L. Lions eds, 2006. pages 11
- [107] V.V. Fedorov. *Theory Of Optimal Experiments*. Probability and Mathematical Statistics. Elsevier Science, 1972. pages 75, 84, 85, 86, 89, 108
- [108] F. Filbet and S. Jin. An Asymptotic Preserving Scheme for the ES-BGK model. *ArXiv e-prints*, March 2010. pages 213
- [109] U. S. Fjordholm, R. Käppeli, S. Mishra, and E. Tadmor. Construction of approximate entropy measure valued solutions for hyperbolic systems of conservation laws. *ArXiv e-prints*, February 2014. pages 32, 136, 138, 288
- [110] J.A. Fleck and J.D. Cummings. An implicit monte carlo scheme for calculating time and frequency dependent nonlinear radiation transport. *Journal of Computational Physics*, 8(3):313 – 342, 1971. pages 270, 271, 284
- [111] J. Foo and G.E. Karniadakis. Multi-element probabilistic collocation method in high dimensions. *J. Comput. Phys.*, 229:1536–1557, 2010. pages 84, 90
- [112] International Organization for Standardization. *Guide to the expression of uncertainty in measurement (GUM)*, volume ISO draft guide DGUIDE99998. International Organization for Standardization, Geneva, 2008. pages 17, 18, 303
- [113] Emmanuel Franck, Christophe Buet, and Bruno Després. Asymptotic preserving finite volumes discretization for non-linear moment model on unstructured meshes. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, pages 467–474. Springer Berlin Heidelberg, 2011. pages 213
- [114] P. Frauenfelder, C. Schwab, and R.A. Todor. Finite Element for Elliptic Problems with Stochastic Coefficients. *Comp. Meth. Appl. Mech. Engrg.*, 194:205–228, 2004. pages 53
- [115] B. Ganapathysubramanian and N. Zabaras. Sparse Grid Collocation Schemes for Stochastic Natural Convection Problems. *J. Comp. Phys.*, 225:652–685, 2007. pages 84, 90
- [116] C. Kristopher Garrett and Cory D. Hauck. A comparison of moment closures for linear kinetic transport equations: The line source benchmark. *Transport Theory and Statistical Physics*, 42(6–7), 2013. pages 11, 163
- [117] Walter Gautschi. *Orthogonal polynomials: applications and computation*, volume 5. Oxford University Press, 1996. pages 9, 31, 42, 43, 46, 78, 79, 80, 82, 84, 115
- [118] P.N. Gavriliadis and G.A. Athanassoulis. Moment data can be analytically completed. *Probabilistic Engineering Mechanics*, 18(4):329–338, 2003. doi:10.1016/j.probengmech.2003.07.001. pages 46, 115
- [119] S. L. Gavrilyuk, N. Favrie, and R. Saurel. Modelling wave dynamics of compressible elastic materials. *J. Comput. Phys.*, 227(5):2941–2969, February 2008. pages 9
- [120] F. Genz, A. C. Thomasset. A package for testing multiple integration subroutines. In P. Keast and G. Fair-weather, editors, *Numerical Integration*, pages 337–340, Dordrecht, 1987. Kluwer. pages 156
- [121] M. I. Gerritsma, J.-B. van der Steen, P. Vos, and G. E. Karniadakis. Time-dependent generalized polynomial chaos. *J. Comput. Physics*, pages 8333–8363, 2010. pages 31, 138

- [122] R. G. Ghanem and J. Red-Horse. Propagation of Uncertainty in Complex Physical Systems using a Stochastic Finite Elements Approach. *Physica D*, 133:137–144, 1999. pages 53
- [123] R.G. Ghanem. Ingredients for a General Purpose Stochastic Finite Element Formulation. *Comp. Meth. Appl. Mech. Eng.*, 168:19–34, 1999. pages 53
- [124] R.G. Ghanem and P. Spanos. *Stochastic Finite Elements: a Spectral Approach*. Springer-Verlag, 1991. pages 31, 38, 53
- [125] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: a Spectral Approach*. Dover, 1991. pages 53
- [126] S. Glasstone G.I. Bell. *Nuclear Reactor Theory*. Van Nostrand Reinhold Company, New York, N.Y. 10001, 1970. pages 255
- [127] F. Golse and G. Allaire. *Transport et Diffusion*. 2015. Polycopié de cours. pages 165, 216, 260
- [128] François Golse. *The Boltzmann equation and its hydrodynamic limits*, volume 2 of *Evolutionary equations*. 2005. pages 4, 6, 8
- [129] Gene H. Golub and Gerard Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton University Press, Princeton, NJ, USA, 2009. pages 42, 115
- [130] S. Gottlieb, J.H. Jung, and S. Kim. A Review of David Gottlieb's Work on the Resolution of the Gibbs Phenomenon. *Commun. Comp. Phys.*, 9:497–519, 2011. pages 117
- [131] Philip T. Gressman and Robert M. Strain. Global classical solutions of the boltzmann equation with long-range interactions. *Proceedings of the National Academy of Sciences*, 107(13):5744–5749, 2010. arXiv:1002.3639Freely accessible. Bibcode:2010PNAS..107.5744G. doi:10.1073/pnas.1001185107. PMC 2851887Freely accessible. PMID 20231489. pages 6
- [132] Bal Guillaume. *Couplage d'équations et homogeneisation en transport neutronique*. PhD thesis, Université Paris 6, France, 1997. pages 165
- [133] Sébastien Guisset, Stéphane Brull, Bruno Dubroca, and Emmanuel D'Humieres. Asymptotic-preserving numerical scheme for the electronic M1 model in the diffusive limit. *Mathematical Modelling and Numerical Analysis*, 2017. pages 12, 163
- [134] J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12):701–702, December 1964. pages 77
- [135] D. C. Handscomb. Remarks on Monte Carlo Integration Method. *Numerische Mathematik*, 6:261–168, 1964. pages 234
- [136] Cory Hauck and Ryan McClarren. Positive p_n closures. *SIAM J. Sci. Comput.*, 32(5):2603–2626. pages 163
- [137] Cory D Hauck, C David Levermore, and André L Tits. Convex duality and entropy-based moment closures: Characterizing degenerate densities. *SIAM Journal on Control and Optimization*, 47(4):1977–2015, 2008. pages 163
- [138] M. Héliot. Sensitivity analysis for complex models. Technical Report ???, CEA DAM CESTA, under the supervision of G. Poëtte, 2018. pages 15, 18
- [139] A. Henry. The Application of Reactor Kinetics to the Analysis of Experiments. *Nucl. Sci. Engng*, 3:52–70, 1958. pages 211, 212, 213
- [140] A. Henry and N.J Curlee. Verification of a Method for Treating Neutron Space-Time Problems. *Nucl. Sci. Engng*, 4:727–744, 1958. pages 211, 212, 213
- [141] F. Hermeline. A discretization of the multigroup pn radiative transfer equation on general meshes. *J. Comp. Phys.*, 313:549–582, 2016. pages 12, 54, 163

- [142] T.D. Hien and M. Kleiber. Stochastic Finite Element Modeling in Linear Transient Heat Transfer. *Comp. Meth. Appl. Mech. Engrg.*, 144:111–124, 1997. pages 53
- [143] D. Hilbert. Begründung der kinetischen gastheorie. *Math. Ann.*, 72:562–577, 1912. pages 8, 195, 211, 262, 269
- [144] Philippe Humbert. Stochastic neutronics with panda deterministic code. *Nuc1. Math. And Compo Sciences: A Century in Review, A Century Anew*, pages 6–11, 2003. pages 189
- [145] Bertrand Iooss and Paul Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Dellino, Gabriella and Meloni, Carlo, Springer US, Boston, MA, 2015. pages 28, 230
- [146] Adam Glenn Irvine, Iain D. Boyd, and Nicholas A. Gentile. Reducing the spatial discretization error of thermal emission in implicit monte carlo simulations. *Journal of Computational and Theoretical Transport*, 45(1-2):99–122, 2016. pages 210
- [147] Aarno Isotalo. *Computational Methods for Burnup Calculations with Monte Carlo Neutronics*. PhD Thesis, Aalto University School of Science Department of Applied Physics, 2013. pages 255
- [148] A.E. Isotalo, J. Leppänen, and J. Dufek. Preventing Xenon Oscillations in Monte Carlo Burnup Calculations by Enforcing Equilibrium Xenon Distribution. *Annals of Nuclear Energy*, 60:78–85, 2013. pages 183, 255, 256, 261
- [149] S. Jaouen J.-M. Clarisso and P.-A. Raviart. A Godunov Type Method in Lagrangian Coordinates for Computing Linearly-Perturbed Planar-Symmetric Flows of Gas Dynamics. *J. Comp. Phys.*, 198:80–105, 2004. pages 123, 130
- [150] L.Masse S. Jaouen and B. Canaud. Hydrodynamic Instabilities in Ablative Tamped Flows. *Phys. Plas.*, 13:122701, 2006. pages 123, 130
- [151] S. Jaouen. A Purely Lagrangian Method for Computing Linearly-Perturbed Flows in Spherical Geometry. *J. Comp. Phys.*, 225:464–490, 2007. pages 123, 130
- [152] M. Jardak, C.H. Su, and G.E. Karniadakis. Spectral polynomial chaos solutions of the stochastic advection equation. *Journal of Scientific Computing*, 17(1):319–338, 2002. pages 53, 57
- [153] S. Jin, H. Lu, and L. Pareschi. Efficient Stochastic Asymptotic-Preserving IMEX Methods for Transport Equations with Diffusive Scalings and Random Inputs. *ArXiv e-prints*, March 2017. pages 223
- [154] H. Jourdren and S. Del Pino. Arbitrary High-Order Schemes for the Linear Advection and Wave Equations: Application to Hydrodynamics and Aeroacoustics. *C.R. Acad. Sci. paris, Ser. I*, 342:441–446, 2006. pages 9, 16, 17, 26, 134
- [155] Michael Junk. Maximum Entropy for Reduced Moment Problems. *Math. Mod. Meth. Appl. Sci.* pages 63, 64, 67, 115
- [156] Samuel Karlin and Lloyd S. Shapley. *Geometry of moment spaces*. American Mathematical Society, 1953. pages 44, 45, 46, 114, 115, 116
- [157] A. Keese. *A General Purpose Framework for Stochastic Finite Elements*. Ph. d. thesis, Mathematik und Informatik der Technischen Universität Braunschweig, 2004. pages 53
- [158] Pierrick Kersaudy, Bruno Sudret, Nadège VARSIER, Odile Picon, and Joe Wiart. A new surrogate modeling technique combining Kriging and polynomial chaos expansions – Application to uncertainty analysis in computational dosimetry. *Journal of Computational Physics*, 286:130–117, January 2015. pages 48, 84, 92, 93, 94, 96, 102
- [159] Yu L. Klimontovich. *Statistical Physics*. CRC Press, 1986. pages 4
- [160] Gilles Kluth. *Analyse mathmatique et numrique de systmes hyperlastiques et introduction de la plasticit*. PhD thesis, 2008. Thse de doctorat dirige par Desprs, Bruno et Frey, Pascal Mathmatiques appliques Paris 6 2008. pages 9

- [161] Tobias Kunz, Andrea Thomaz, and Henrik Christensen. Hierarchical rejection sampling for informed kinodynamic planning in high-dimensional spaces. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 89–96, 2016. pages 200
- [162] J. Kusch, G. W. Alldredge, and M. Frank. Maximum-principle-satisfying second-order intrusive polynomial moment scheme. *arXiv preprint arXiv:1712.06966*, 2017. pages 57, 64, 69, 71, 72, 289
- [163] J. Kusch and M. Frank. Intrusive methods in uncertainty quantification and their connection to kinetic theory. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, pages 1–16, 2018. pages 71, 72
- [164] Jonas Kusch, Ryan G. McClarren, and Martin Frank. Filtered stochastic galerkin methods for hyperbolic equations. *J. Comput. Phys.*, 2018. pages 71, 72
- [165] B. Lapeyre, E. Pardoux, and R. Sentis. *Méthodes de Monte Carlo pour les équations de transport et de diffusion*. Number 29 in Mathématiques & Applications. Springer-Verlag, 1998. pages 36, 76, 163, 169, 188, 227, 231, 234, 241, 242, 247, 254, 256, 257
- [166] Loic Le Gratiet, Claire Cannamela, and Bertrand Iooss. A Bayesian approach for global sensitivity analysis of (multi-fidelity) computer codes. working paper or preprint, June 2013. pages 28
- [167] O. Le Maître, M. Reagan, H. Najm, R. Ghanem, and O. Knio. A Stochastic Projection Method for Fluid Flow. II. Random Process. *J. Comp. Phys.*, 181:9–44, 2002. pages 40
- [168] O. Le Maître, M. Reagan, H. Najm, R. Ghanem, and O. Knio. Multi-Resolution Analysis of Wiener-Type Uncertainty Propagation Schemes. *J. Comp. Phys.*, 197:502–531, 2004. pages 52, 246
- [169] R. Lebrun and A. Dutfoy. A Generalization of the Nataf Transformation to Distributions with Elliptical Copula. *Prob. Eng. Mech.*, 24,2:172–178, 2009. pages 48, 222
- [170] R. Lebrun and A. Dutfoy. An Innovating Analysis of the Nataf Transformation from the Copula viewpoint. *Prob. Eng. Mech.*, 24,3:312–320, 2009. pages 48
- [171] Pierre L’Ecuyer. Uniform random number generators: A review, 1997. pages 184
- [172] C. D. Levermore. Moment Closure Hierarchies for Kinetic Theories. *J. Stat. Phys.*, 83:5/6, 1996. pages 115
- [173] E. E. Lewis and W. F. Miller Jr. *Computational Methods of Neutron Transport*. John Wiley and Son New York, 1984. pages 76, 163, 168, 171, 186, 187, 231, 243, 257, 277
- [174] R. W. Lindquist. Relativistic transport theory. *Annals of Physics*, 37:487–518, May 1966. pages 215, 218
- [175] A. Loeven, J. Witteveen, and H. Bijl. Efficient Uncertainty Quantification using a Two-Step Approach with Chaos Collocation. *ECCOMAS CFD*, 2006. pages 84, 90
- [176] G.J.A. Loeven and H. Bijl. Airfoil Analysis with Uncertain Geometry Using the Probabilistic Collocation Method. *49th AIAA Aerospace Sciences Meeting and Exhibit*, AIAA 2008-2070, 2008. pages 84, 90
- [177] G.J.A Loeven, J. A. S. Witteveen, and H. Bijl. Probabilistic Collocation: an Efficient Non Intrusive Approach for Arbitrarily Distributed Parametric Uncertainties. *45th AIAA Aerospace Sciences Meeting and Exhibit*, AIAA 2007-317, 2007. pages 84, 90
- [178] R. B. Lowrie, J. E. Morel, and J. A. Hittinger. The coupling of radiation and hydrodynamics. *Astrophysical Journal*, 521:432–450, August 1999. pages 14, 269
- [179] D. Lucor, C. Enaux, H. Jourdren, and P. Sagaut. Multi-Physics Stochastic Design Optimization: Application to Reacting Flows and Detonation. *Comp. Meth. Appl. Mech. Eng.*, 196:5047–5062, 2007. pages 31

- [180] D. Lucor, J. Meyers, and P. Sagaut. Sensitivity Analysis of LES to Subgrid-Scale-Model Parametric Uncertainty using Polynomial Chaos. *J. Fluid Mech.*, 585:255–279, 2007. pages 237
- [181] Depinay J. M. *Automatisation de Méthodes de Réduction de Variance pour la Résolution de l'équation de Transport*. Ph. d. thesis, Université de Paris 11, Orsay, FRANCE (Université de soutenance), 2000. pages 234, 247, 250
- [182] P. Maire, R. Abgrall, J. Breil, and J. Ovadia. A cell-centered lagrangian scheme for two-dimensional compressible flow problems. *SIAM Journal on Scientific Computing*, 29(4):1781–1824, 2007. pages 9
- [183] Pierre-Henri Maire, Rémi Abgrall, Jérôme Breil, Raphael Loubère, and Bernard Reboulet. A nominally second-order cell-centered lagrangian scheme for simulating elasticplastic flows on two-dimensional unstructured grids. *Journal of Computational Physics*, 235:626 – 665, 2013. pages 9
- [184] S. Maire. Reducing Variance using Iterated Control Variates. *J. Stat. Comp. Sim.*, 73(1):1–30, 2003. pages 234
- [185] O. P. Le Maître and O. M. Knio. Uncertainty Propagation using Wiener-Haar Expansions. *J. Comp. Phys.*, 197:28–57, 2004. pages 31, 49, 52, 246
- [186] O. P. Le Maître and O. M. Knio. A Stochastic Particle-Mesh Scheme for Uncertainty Propagation in Vortical Flows. *J. Comp. Phys.*, 226:645–671, 2007. pages 31, 52
- [187] Amitava Majumdar. Parallel performance study of monte carlo photon transport code on shared-, distributed-, and distributed-shared-memory architectures. 2000. pages 289
- [188] F. H. Maltz and D. L. Hitzl. Variance reduction in Monte Carlo computations using multi-dimensional hermite polynomials. *J. Comput. Phys.*, Volume 32(3):345–376, 1979. pages 234
- [189] A. Mangeney, F. Califano, C. Cavazzoni, and P. Travnicek. A numerical scheme for the integration of the vlasov-maxwell system of equations. *Journal of Computational Physics*, 179(2):495 – 538, 2002. pages 214
- [190] et al Martin, William R. Monte carlo photon transport on shared memory and distributed memory parallel processors. *International Journal of High Performance Computing Applications*, 1.3:57–74, 1987. pages 289
- [191] J.-M. Martinez, J. Cahen, A. Millard, D. Lucor, F. Huvelin, J. Ko, and N. Poussineau. Modélisation des Incertitudes par Polynômes de Chaos – Étude d'un Écoulement en Milieux Poreux. Technical Report Rapport DM2S/DIR/RT/06-006/A, CEA-CEMRACS, 2006. pages 75, 81, 82
- [192] L. Mathelin and O. P. Le Maître. A Posteriori Error Analysis for Stochastic Finite Element Solutions of Fluid Flows with Parametric Uncertainties. *ECCOMAS CFD*, 2006. pages 53
- [193] Julien Mathiaud. *Models and methods for complex flows: application to atmospheric reentry and particle / fluid interactions*. Habilitation à diriger des recherches, University of Bordeaux, June 2018. pages 12, 290
- [194] Julien Mathiaud and Luc Mieussens. A Fokker-Planck model of the Boltzmann equation with correct Prandtl model. In *30th International Symposium on Rarefied Gas Dynamics*, Victoria, Canada, July 2016. pages 290
- [195] C. Mazeran and B. Després. Lagrangian gas dynamics in two dimensions and Lagrangian systems. *Arch. Rat. Mech. Anal.*, 2005. (To be published). pages 9
- [196] Ryan G McClarren and Cory D Hauck. Robust and accurate filtered spherical harmonics expansions for radiative transfer. *Journal of Computational Physics*, 229(16):5597–5614, 2010. pages 54, 163
- [197] Michael Scott McKinley, Eugene D. Brooks III, and Abraham Szoke. Comparison of implicit and symbolic implicit monte carlo line transport with frequency weight vector extension. *Journal of Computational Physics*, 189(1):330 – 349, 2003. pages 210, 272

- [198] L. R. Mead and N. Papanicolaou. Maximum Entropy in the Problem of Moments. *J. Math. Phys.*, 25 (8), 1984. pages 63, 64, 115
- [199] Boukhmes Mechitoua. Tokaimura criticality accident: Point model stochastic neutronic interpretation. *ANS Annual Meeting*, 2001. pages 189, 190
- [200] Boukhmes Méchitoua. On the fictitious aspect of the critical state. *JAERI-Conf*, 019, 2003. JP0450146. pages 163, 189, 190, 191
- [201] J. Le Meitour, D. Lucor, and J-C Chassaing. Prediction of Stochastic Limit Cycle Oscillations using an Adaptive Polynomial Chaos Method. *Journal of Aeroelasticity and Structural Dynamics*, 2(1):3–22, 2010. pages 246
- [202] J. Mercer. Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations. *Philos. Trans. Roy. Soc.*, 209, 1909. pages 133
- [203] D. Mihalas and B. W. Mihalas. *Foundations of Radiation Hydrodynamics*. Dover books on physics. Dover Publications, 1999. pages 14, 163, 213, 214, 215, 218, 220, 267, 268, 269
- [204] Frederik Riis Mikkelsen. Probabilistic aspects of moment sequences: The method of moments. pages 44
- [205] Guillaume Morel, Christophe Buet, and Bruno Després. Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P 1 model. working paper or preprint, December 2017. pages 12
- [206] J. Morice and S. Jaouen. Perturbations Linéaires d’Écoulements Monodimensionnels à Géométries Plates, Cylindriques et Sphériques. Technical Report 6040, CEA, 2003. (in French). pages 123, 130
- [207] I. Müller and T. Ruggeri. *Rational Extended Thermodynamics, 2nd ed.* Springer. Tracts in Natural Philosophy, Volume 37, 1998. Springer-Verlag, New York. pages 4, 6, 7, 8, 9, 36, 66, 71
- [208] H.N. Najim. Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics. *Annu. Rev. Fluid Mech.*, 41:35–52, 2009. pages 40
- [209] Harald Niederreiter. Point sets and sequences with small discrepancy. *Monatshefte für Mathematik*, 104(4):273–337, Dec 1987. pages 77
- [210] Harald Niederreiter. *Random Number Generation and quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992. pages 76, 77
- [211] T. N’Kaoua. Solution of the nonlinear radiative transfer equations by a fully implicit matrix monte carlo method coupled with the rosseland diffusion equation via domain decomposition. *SIAM J. Sci. Stat. Comput.*, 12(3):505–520, March 1991. pages 270
- [212] F. Nobile, R. Tempone, and C. Webster. A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008. pages 84, 90
- [213] Erich Novak and Klaus Ritter. The Curse of Dimension and a Universal Method for Numerical Integration. In *Multivariate Approximation and Splines*, pages 177–187, 1998. pages 245
- [214] Edgar Olbrant, Cory D. Hauck, and Martin Frank. A realizability-preserving discontinuous galerkin method for the m1 model of radiative transfer. *Journal of Computational Physics*, 231(17):5612 – 5639, 2012. pages 163
- [215] H. N. Najm O.P. Le Maître, O. M. Knio and R. G. Ghanem. A Stochastic Projection Method for Fluid Flow I: Basic Formulation. *J. Comp. Phys.*, 173:481–511, 2001. pages 31, 52
- [216] K.O. Ott and D.A. Meneley. Accuracy of the Quasistatic Treatment of Spatial Reactor Kinetics. *Nucl. Sci. Engng*, 36(3):402–411, 1969. pages 211, 212, 213

- [217] N. E. Owen, P. Challenor, P. P. Menon, and S. Bennani. Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):403–435, 2017. pages 102
- [218] M. Paffrath and U. Wever. Adapted Polynomial Chaos Expansion for Failure Detection. *J. Comp. Phys.*, 226(1):263–281, 2007. pages 123
- [219] G. C. Papanicolaou. Asymptotic Analysis of Transport Processes. *Bulletin of the American Mathematical Society*, 81(2), 1975. pages 165, 177, 183, 226
- [220] S. K. Park and K. W. Miller. Random number generators: Good ones are hard to find. *Commun. ACM*, 31(10):1192–1201, October 1988. pages 184
- [221] Cyril Patricot. *Couplages multi-physiques : évaluation des impacts méthodologiques lors de simulations de couplages neutronique/thermique/mécanique*. PhD thesis, 2016. Thèse de doctorat dirigée par Allaire, Grégoire Mathématiques appliquées Paris Saclay 2016. pages 211, 212, 213
- [222] I. Pazsit. Front matter. In IMRE PAZSIT and LENARD PAL, editors, *Neutron Fluctuations*, pages iii –. Elsevier, Amsterdam, 2008. pages 191
- [223] B. Perthame. *Transport Equations in Biology*. Birkhauser Verlag, Basel Boston Berlin, 2000. pages 163, 255
- [224] Per Pettersson, Gianluca Iaccarino, and year=2012 Jan Nordström. A roe variable based chaos method for the euler equations under uncertainty. pages 56
- [225] R. Peyret. *Spectral Methods for Incompressible Viscous Flow*. Applied Mathematical Sciences. Springer New York, 2002. pages 90
- [226] Teddy Pichard. *Mathematical modelling for dose depositon in phototherapy*. Theses, Université de Bordeaux, November 2016. pages 12, 163
- [227] Teddy Pichard, Graham W. Alldredge, Stéphane Brull, Bruno Dubroca, and M. Frank. An approximation of the m_2 closure: Application to radiotherapy dose simulation. *J. Sci. Comput.*, 71(1):71–108, 2017. pages 163
- [228] Teddy Pichard, Denise Aregba-Driollet, Stphane Brull, Bruno Dubroca, and Martin Frank. Relaxation schemes for the m_1 model with space-dependent flux: Application to radiotherapy dose calculation. *Communications in Computational Physics*, 19(1):168191, 2016. pages 12, 163
- [229] Christopher Poëtte. *Fragmented landscape : impact on atmospheric flow and tree stability*. Theses, Université de Bordeaux, December 2016. pages 14
- [230] Christopher Poëtte, Barry Gardiner, Sylvain Dupont, Ian Harman, Margi Böhm, John Finnigan, Dale Hughes, and Yves Brunet. The impact of landscape fragmentation on atmospheric flow: a wind-tunnel study. *Boundary-Layer Meteorology*, 163(3):393–421, 2017. pages 14
- [231] G. Poëtte. *Perturbations linéaires tridimensionnelles d'un écoulement de base monodimensionnel en coordonnées eulériennes dans le formalisme Lagrange+Projection*. Mémoire de master 2 recherche, Institut Élie Cartan, Nancy, 2006. pages 19, 123
- [232] G. Poëtte. *Propagation d'Incertitudes pour les Systèmes de Lois de Conservation, Méthodes Spectrales Stochastiques*. Phd thesis, Université Pierre et Marie Curie, Institut Jean Le Rond D'Alembert, 2009. pages 12, 17, 19, 26, 54, 56, 57, 64, 67, 71, 123, 128, 129
- [233] G. Poëtte. Extension du Chaos Polynomial pour la propagation d'incertitudes. *Revue Chocs Avancées*, 10(46), 2016. pages 19, 30
- [234] G. Poëtte. Propagation d'incertitudes par Chaos Polynomial. *Revue Chocs, Maitrise des incertitudes*, 48(01):24–35, 2018. pages 19, 30

- [235] G. Poëtte. Spectral convergence of the generalized Polynomial Chaos reduced model obtained from the uncertain linear Boltzmann equation. *Preprint submitted to Mathematics and Computers in Simulation*, 2019. pages 230
- [236] G. Poëtte, B. Després, and D. Lucor. Uncertainty Quantification for Systems of Conservation Laws. *J. Comp. Phys.*, 228(7):2443–2467, 2009. pages 12, 17, 19, 21, 57, 64, 67, 69, 71
- [237] G. Poëtte, B. Després, and D. Lucor. Treatment of Uncertain Interfaces in Compressible Flows. *Comp. Meth. Appl. Math. Engrg.*, 200:284–308, 2010. pages 12, 19, 56, 71, 123, 129, 130, 132
- [238] G. Poëtte and D. Lucor. Non Intrusive Iterative Stochastic Spectral Representation with Application to Compressible Gas Dynamics. *J. of Comput. Phys.*, 2011. DOI information: 10.1016/j.jcp.2011.12.038. pages 19, 42, 74, 104, 106, 107, 108, 112, 121, 138, 246
- [239] G. Poëtte, D. Lucor, and B. Després. Uncertainty Propagation for Systems of Conservation Laws, High-Order Stochastic Spectral Methods. In Springer, editor, *International Conference On Spectral Analysis and High Order Methods*, 2009. pages 19, 54, 60, 71
- [240] Gaël Poëtte. A comparative study of generalized Polynomial Chaos based Approximations: integration vs. regression vs. collocation vs. kriging. *International Journal for Uncertainty Quantification*, 2018. pages 17, 19
- [241] Gaël Poëtte. A gPC-intrusive Monte-Carlo scheme for the resolution of the uncertain linear Boltzmann equation. *Journal of Computational Physics*, 385:135 – 162, 2019. pages 15, 18, 19, 20, 222, 225, 229, 230, 231, 289, 290
- [242] Gaël Poëtte, Alexandre Biroilleau, and Didier Lucor. Iterative polynomial approximation adapting to arbitrary probability distribution. *SIAM J. Numerical Analysis*, 53(3):1559–1584, 2015. pages 19, 42, 74, 104, 107, 110, 112, 114, 117, 121, 246
- [243] Gaël Poëtte, Didier Lucor, and Hervé Jourdren. A stochastic surrogate model approach applied to calibration of unstable fluid flow experiments. *Comptes Rendus Mathematique*, 350(5):319 – 324, 2012. pages 12, 17, 19, 34, 123, 124, 129, 130, 132, 138, 288
- [244] Poggi, Thorembey, and Rodriguez. Velocity Measurements in turbulent gaseous mixtures induced by Richtmyer-Meshkov instability. *Physics of Fluids*, 10:11, 1998. pages 123, 131, 132, 134, 136, 137
- [245] G. C. Pomraning. *The equations of radiation hydrodynamics*. Dover books on physics. Dover Publications, 1973. pages 164, 213, 215, 217, 218, 267, 268, 269
- [246] Anil K Prinja. Notes on the lumped backward master equation for the neutron extinction/survival probability. Technical report, Los Alamos National Laboratory (United States). Funding organisation: DOE/LANL (United States), 2012. pages 189, 191
- [247] R. Procassini, M. OBrien, and J. Taylor. Load balancing of parallel Monte Carlo transport calculations. *Mathematics and Computation, Supercomputing, Reactor Physics and Nuclear and Biological Applications*, 2005. pages 289
- [248] L. Saint Raymond. *The Boltzmann equation and its hydrodynamic limits*, volume 1971 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2009. pages 4, 6, 8
- [249] P. Reuss. *Précis de neutronique*. Collection Génie atomique. EDP Sciences, 2003. pages 186
- [250] Yann Richet. *Suppression du régime transitoire initial des simulations Monte-Carlo de criticit*. PhD thesis, 2006. These de doctorat dirigée par Carraro, Laurent Mathmatiques appliques Saint-Etienne, EMSE 2006. pages 19, 301, 304
- [251] S. Rjasanow and W. Wagner. *Stochastic Numerics for the Boltzmann Equation*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2006. pages 4

- [252] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. pages 185
- [253] P. T. Roy, N. El Moçayd, S. Ricci, J.-C. Jouhaud, N. Goutal, M. De Lozzo, and M. Rochoux. Comparison of polynomial chaos and gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows. *Stochastic Environmental Research and Risk Assessment*, 2017. submitted. pages 102
- [254] A. Saltelli and Sobol I.M. About the use of rank transformation in sensitivity analysis of model output. *Reliability Engineering and System Safety*, 50:225–239, 1995. pages 156
- [255] Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280 – 297, 2002. pages 230
- [256] G. Saporta. *Probabilités, Analyse de Données et Statistique*, 2e édition. Technip, 2006. pages 17, 19, 26, 31, 36, 76, 182, 188, 231, 232, 233, 237, 239, 295, 301, 304
- [257] L. Schlachter and F. Schneider. A hyperbolicity-preserving stochastic galerkin approximation for uncertain hyperbolic systems of equations. *arXiv preprint arXiv:1710.03587*, 2017. pages 72
- [258] R. Schöbi, B. Sudret, and S. Marelli. Rare Event Estimation Using Polynomial-Chaos Kriging. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(2):D4016002, 2017. pages 48, 84, 92, 93, 96, 102
- [259] W. Schoutens. *Stochastic Processes and Orthogonal Polynomials*. Springer-Verlag, New York, 2000. pages 42
- [260] D. Serre. *Systèmes Hyperboliques de Lois de Conservation, partie I*. Diderot, 1996. Paris. pages 14, 25, 51, 52, 68, 292, 293, 295
- [261] D. Serre. *Systèmes Hyperboliques de Lois de Conservation, partie II*. Diderot, 1996. Paris. pages 14, 25, 51, 52, 68
- [262] Roland Shöbi, Bruno Sudret, and Joe Wiart. Polynomial-chaos-based kriging. 2015. pages 48, 84, 92, 93, 96, 102
- [263] Winfried Sickel and Tino Ullrich. The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness. Technical report, Jenaer Schriften zur Math. und Inform., Math/Inf/14/06, 2006. pages 83
- [264] M. Siddhartha, C. Schwab, and J. Sukis. *Multi-level Monte Carlo Finite Volume Methods for Uncertainty Quantification in Nonlinear Systems of Balance Laws*, volume 92 of *Lecture Notes in Computational Science and Engineering*. Uncertainty Quantification in Computational Fluid Dynamics, 2013. pages 26, 68, 233
- [265] F. Simon, P. Guillen, P. Sagaut, and D. Lucor. A gPC based approach to uncertain transonic aerodynamics. *CMAME*, 199:1091–1099, 2010. pages 237
- [266] I. Sobol. Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe Modelirovaniye*, 2:112118, 1990. in Russian, translated in English in Sobol , I. (1993). Sensitivity analysis for nonlinear mathematical models. Mathematical Modeling & Computational Experiment (Engl. Transl.), 1993, 1, 407414. pages 18, 230
- [267] I.M. Sobol. Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics*, 16(5):236 – 242, 1976. pages 77
- [268] J. Spanier and E. M. Gelbard. *Monte Carlo Principles and Neutron Transport Problems*. Addison-Wesley, 1969. pages 76, 163, 168, 171, 186, 187, 257, 277
- [269] P. Spanos and R. G. Ghanem. Stochastic Finite Element Expansion for Random Media. *ASCE J. Eng. Mech.*, 115(5):1035–1053, 1989. pages 53, 133

- [270] B. Sudret. Global Sensitivity Analysis using Polynomial Chaos Expansion. *Rel. Engrg. Syst. Saf.*, 93:964–979, 2007. pages 84
- [271] B. Sudret. *Uncertainty Propagation and Sensitivity Analysis in Mechanical Models, Contribution to Structural Reliability and Stochastic Spectral Methods*. Habilitation à Diriger des Recherches, Université Blaise Pascal - Clermont II, 2007. pages 15, 28, 37, 84, 123
- [272] B. Sudret and A. Der Kiureghian. Stochastic Finite Element Methods and Reliability - A State of the Art Report. Technical Report UCB/SEMM-2000/08, Department of civil and environmental engineering, University of California, Berkeley, 2000. pages 28, 53
- [273] G. Szego. *Orthogonal Polynomials*, volume 23. American Mathematical Society, colloquim publications, 1939. pages 31, 42, 46, 82, 115
- [274] S. Tancogne and S. Jaouen. Perturbations Linéaires d'Écoulements Monodimensionnels à Symétrie Planes et Sphérique en Présence de Conduction de Chaleur Non-Linéaire. Technical Report DO 37, CEA, 2004. (in French). pages 9, 123, 130
- [275] M. Temporal and B. Canaud. ICF monomode calculations in the deceleration phase. Private communication. pages 123, 130
- [276] M. Temporal, S. Jaouen, and B. Canaud. Hydrodynamic instabilities in ablative tamped flows. *Phys. Plasmas*, 1:1–1, 2005. in preparation. pages 9, 123, 130
- [277] Lloyd N. Trefethen. Is Gauss Quadrature Better than Clenshaw-Curtis? *SIAM rev.*, 50(1):67 – 87, 2008. pages 82, 98
- [278] J. Tryoen, O. Le Maître, and A. Ern. Adaptive Anisotropic Stochastic Discretization Schemes for Uncertain Conservation Laws. *Proc. of ASME 2010, Third Joint US-European Fluids Engineering Summer Meeting*, 2010. pages 53, 57, 246
- [279] Rodolphe Turpault. Construction d'un modle m1-multigroupe pour les quations du transfert radiatif. *Comptes Rendus Mathematique*, 334(4):331 – 336, 2002. pages 12, 163
- [280] T. Ueki. Intergenerational Correlation in Monte-Carlo k-Eigenvalue Calculation. *Nucl. Sci. Engng.*, 141:101–110, 2002. pages 250
- [281] Xavier Valentin. *Analyse mathématique et numérique des modèles Pn pour la simulation de problèmes de transport de photons*. PhD thesis, Paris Saclay, 2015. Thèse de doctorat dirigée par Lafitte-Godillon, Pauline et Enaux, Cédric Mathématiques appliquées. pages 12, 54, 163
- [282] Xavier Valentin and Gaël Poëtte. To appear. Note interne, 2017. pages 271, 272, 276
- [283] J.G. van der Corput. Verteilungsfunktionen. I. Mitt. *Proc. Akad. Wet. Amsterdam*, 38:813–821, 1935. pages 77
- [284] Emmanuel Vazquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*. PhD thesis, Paris 11, 2005. Thèse de doctorat dirigée par Walter, Éric Sciences appliquées. pages 92
- [285] Daniel Verwaerde. Une approche non déterministe de la neutronique - modélisation. Technical report, CEA, 1993. pages 163, 189, 190, 191
- [286] Vetter and Sturtevant. Experiments on the Richtmyer-Meshkov instability of an air/SF6 interface. *Shock Waves*, 4:247–252, 1995. pages 123, 131, 132, 134, 136, 137, 138
- [287] P. Vos. Time dependent polynomial chaos. Master of science thesis, Delft University of technology, Faculty of Aerospace engineering, 2006. pages 243
- [288] Jörg Waldvogel. Fast construction of the fejér and clenshaw–curtis quadrature rules. *BIT Numerical Mathematics*, 46(1):195–202, 2006. pages 82

- [289] Clément Walter. *Using Poisson Processes for rare event estimations*. PhD thesis, Université de Paris VII, November 2016. pages 83
- [290] X. Wan and G. E. Karniadakis. Stochastic Heat Transfer Enhancement in a Grooved Channel. *J. Fluid Mech.*, 565:255–278, 2006. pages 31
- [291] X. Wan and G.E. Karniadakis. Beyond Wiener-Askey Expansions: Handling Arbitrary PDFs. *SIAM J. Sci. Comp.*, 27(1-3), 2006. pages 38, 40, 42, 105, 114
- [292] X. Wan and G.E. Karniadakis. Long-Term Behavior of Polynomial Chaos in Stochastic Flow Simulations. *Comp. Meth. Appl. Mech. Engrg.*, 195:5582–5596, 2006. pages 138
- [293] X. Wan and G.E. Karniadakis. Long-Term Behaviour of Polynomial Chaos in Stochastic Flow Simulations. *Comput. Meth. Appl. Mech. Engrg.*, 216(5582-5596), 2006. pages 40
- [294] X. Wan and G.E. Karniadakis. Multi-Element generalized Polynomial Chaos for Arbitrary Probability Measures. *SIAM J. Sci. Comp.*, 28(3):901–928, 2006. pages 31, 40, 49, 246
- [295] N. Wiener. The Homogeneous Chaos. *Amer. J. Math.*, 60:897–936, 1938. pages v, 30, 31, 32, 33, 34, 35, 47, 130, 160, 288
- [296] B. R. Wienke, TR Hill, and PP Whalen. Multigroup particle transport in a moving material. *Journal of Computational Physics*, 72(1):177–201, 1987. pages 214, 215, 218, 220, 289
- [297] B.R. Wienke. Transport equations in moving material part i: Neutrons and photons. *Progress in Nuclear Energy*, 46(1):13–55, 2005. pages 214, 215, 216, 218, 289
- [298] D. Winske, B. J. Albright, D. S. Lemons, and W. Daughton. Quiet monte carlo method for the simulation of hydrodynamics, radiation, transport, and plasma. In *IEEE Conference Record - Abstracts. 2002 IEEE International Conference on Plasma Science (Cat. No.02CH37340)*, pages 124–, May 2002. pages 203, 290
- [299] J. A. S. Witteveen and H. Bijl. Using Polynomial Chaos for Uncertainty Quantification in Problems with Non Linearities. *47th AIAA Aerospace Sciences Meeting and Exhibit*, AIAA 2006-2066, 2006. pages 40, 54
- [300] J. A. S. Witteveen and H. Bijl. An Unsteady Adaptive Stochastic Finite Elements Formulation for Rigid-Body Fluid-Structure Interaction. *Comp. and Struct.*, 2008. pages 53
- [301] Allan B. Wollaber. Four decades of implicit monte carlo. *Journal of Computational and Theoretical Transport*, 45(1-2):1–70, 2016. pages 201, 203, 210, 269, 270, 271, 272, 273, 274, 276, 284
- [302] D. Xiu and J.S. Hesthaven. High-Order Collocation Methods for Differential Equations with Random Inputs. *J. Sci. Comput.*, 27(3):1118–1139, 2005. pages 84, 90
- [303] D. Xiu and G. E. Karniadakis. The Wiener-Askey Polynomial Chaos for Stochastic Differential Equations. *SIAM J. Sci. Comp.*, 24(2):619–644, 2002. pages 40
- [304] D. Xiu and G.E. Karniadakis. Modeling Uncertainty in Steady State Diffusion Problems via generalized Polynomial Chaos. *Comp. Meth. Appl. Mech. Engrg.*, 191:4927–4948, 2002. pages 40
- [305] D. Xiu and G.E. Karniadakis. The Wiener-Askey Polynomial Chaos for Stochastic Differential Equations. *SIAM J. Sci. Comp.*, 24:619–644, 2002. pages 38, 40, 42, 105
- [306] D. Xiu and G.E. Karniadakis. Modeling Uncertainty in Flow Simulations via generalized Polynomial Chaos. *Comp. Meth. Appl. Mech. Engrg.*, 187:137–167, 2003. pages 40
- [307] D. Xiu, D. Lucor, and G. E. Karniadakis. Stochastic Modeling of Flow-Structure Interactions. In *Computational Fluid and Solid Mechanics, Elsevier, Proceedings of the 1st MIT conference, Cambridge, Massachusetts, K.J. Bathe (Ed.)*, volume 2, pages 1420–1423, 2001. pages 31