Schemes and Strategies to Propagate and Analyze Uncertainties in Computational Fluid Dynamics Applications

Gianluca Geraci

THÈSE

présentée à

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

Par **Gianluca GERACI**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPECIALITÉ: **Mathématiques Appliquées et Calcul Scientifique**

Schemes and Strategies to Propagate and Analyze Uncertainties in Computational Fluid Dynamics Applications

Schémas et stratégies pour la propagation et l'analyse des incertitudes dans la simulation d'écoulements

Préparée à l'INRIA Bordeaux Sud-Ouest
Équipe BACCHUS

Directeur de thèse: Rémi Abgrall
Co-directeur de thèse: Pietro Marco Congedo

Soutenue le: 5 Décembre 2013

Après avis des rapporteurs:

• Bruno Després	Professeur, Université Pierre et Marie Curie
• Didier Lucor	Chargé de recherche, Université Pierre et Marie Curie

Devant la commission d'examen composeée de:

• Rémi Abgrall (*Examinateur*)	Professeur, Université Bordeaux 1 et INRIA
• Pietro Marco Congedo (*Examinateur*)	Chargé de recherche, INRIA
• Bruno Després (*Rapporteur*)	Professeur, Université Pierre et Marie Curie
• Didier Lucor (*Rapporteur*)	Chargé de recherche, Université Pierre et Marie Curie
• Bernhard Müller (*Président*)	Professeur, Norvegian Univ. of Science and Technology

– 2013 –

Ai miei genitori, Rita e Giacomo,
e alla memoria dei miei nonni,
Luigia, Lucia, Luigi e Carmelo.

Aknowlegments

I would like to thank my advisors, Prof. Rémi Abgrall and Dr. Pietro Marco Congedo, for their support during this thesis work. Working with both of them has been a privilege for me. They, each in a different way, constantly have encouraged me and, for sure, I have profited of only a small amount of their feedbacks. I am endlessly grateful to Prof. Abgrall for accepting me as one of his student offering me the possibility to engage a so stimulating and challenging topic. Moreover, the patience, the encouragement and the genuine trust that Dr. Congedo placed in me and in my work will be an example I will preserve for the rest of my life.

I thank also the two reviewers, Prof. Bruno Després and Dr. Didier Lucor, for the carefully reading of the manuscript and their invaluable comments. To they and to Prof. Bernhard Müller I would like to extend my sincere thanks for having accepted the duty to join the jury of my thesis defense.

I have shared three years with many colleagues at Inria-Bordeaux; all of them deserve a thanks, especially the ones in the Bacchus team. Among them, I would like to thank Dr. Julie Tryoen for her wise advice. It is also a pleasure to mention here Dr. Dante De Santis. We started together this PhD-journey and now, once more time, our paths seem that will cross again. We really shared the best and worst moments day by day, spending many hours discussing several CFD-related topics. Dr. Maria Giovanna Rodio gently introduced me to the two-phase flows world. Her ability to make work sessions jovial, within her expertise, greatly contributed to alleviate the effort we made to obtain more challenging test cases. Maria Giovanna, Dante and Pietro also contributed to recreate a pleasant family atmosphere in our (unfortunately rare) moments outside the office. Moreover, Dr. Mario Ricchiuto is acknowledged for his well-recognized welcoming role for our, Inria's Italian freshmen, as I was.

Part of this work has been conceived during my stay at the Stanford University. I am truly indebted with the Prof. Gianluca Iaccarino for hosting me in his UQ Lab. The insightful comments of Prof. Iaccarino were fundamental in developing the work related to high-order moments analysis. The stimulating discussions with him renovated in me the dormant taste for the engineering applications. During my visit at the Stanford University, I had the opportunity to meet many scientists involved in UQ and not only. Among the others, I would like to thank Dr. Catherine Gorlé, Dr. Per Petterson, Dr. Gary Tang, Dr. Mike Emory, Dr. John Axerio-Cilies, Dr. Paul Constantine, Dr. Riccardo Rossi, Dr. Filippo Coletti and Dr. Guido Lodato for welcoming me. Dr. Daniele Schiavazzi and his affable wife Giulia Vescovi deserve a very special thanks: they made me feel less far from home than I was during my stay there.

The genuine interest in my future of Prof. Alberto Guardone brought me in Bordeaux. I cannot forget it and for this, still today, I thank him.

My parents, Rita e Giacomo, and my brother Giorgio are the truly source of inspiration of my dedication and commitment. Dedication and commitment that have allowed me to achieve this goal.

Finally, Dora is acknowledged not only for her daily (endless) forbearance with me, but greatly for her very peculiar skill to introduce, neatly, randomness in my otherwise deterministic life.

Contents

Abstract

In this manuscript, three main contributions are illustrated concerning the propagation and the analysis of uncertainty for computational fluid dynamics (CFD) applications. First, two novel numerical schemes are proposed: one based on a collocation approach, and the other one based on a finite volume like representation in the stochastic space. In both the approaches, the key element is the introduction of a non-linear multiresolution representation in the stochastic space. The aim is twofold: reducing the dimensionality of the discrete solution and applying a time-dependent refinement/coarsening procedure in the combined physical/stochastic space.

Finally, an innovative strategy, based on variance-based analysis, is proposed for handling problems with a moderate large number of uncertainties in the context of the robust design optimization. Aiming to make more robust this novel optimization strategies, the common ANOVA-like approach is also extended to high-order central moments (up to fourth order). The new approach is more robust, with respect to the original variance-based one, since the analysis relies on new sensitivity indexes associated to a more complete statistic description.

Keywords: Uncertainty propagation, Multiresolution, Adaptivity, Finite Volume, Fluid Dynamics, Statistics, Sensitivity Analysis, Robust Optimization

Résumé

Ce manuscrit présente des contributions aux méthodes de propagation et d'analyse d'incertitude pour des applications en Mécanique des Fluides Numérique. Dans un premier temps, deux schémas numériques innovantes sont présentées: une approche de type "Collocation", et une autre qui est basée sur une représentation de type "Volumes Finis" dans l'espace stochastique. Dans les deux, l'élément clé est donné par l'introduction d'une représentation de type "Multirésolution" dans l'espace stochastique. L'objective est à la fois de réduire le nombre de dimensions et d'appliquer un algorithme d'adaptation de maillage qui puisse être utilisé dans l'espace couplé physique/stochastique pour des problèmes non-stationnaires. Pour finir, une stratégie d'optimisation robuste est proposée, qui est basée sur une analyse de décomposition de la variance et des moments statistiques d'ordre plus élevé. Dans ce cas, l'objectif est de traiter des problèmes avec un grand nombre d'incertitudes.

Keywords: Propagation de l'incertitude, Multirésolution, Adaptativité, Volumes Finis, Dynamique des Fluides, Statistiques, Analyse de sensibilité, Optimisation robuste

List of papers and related contributions

Journal Papers

The present manuscript is based on the following Journal papers. The policy adopted for the papers has been the so-called *co-first authorship*. The names are reported in alphabetical order for all the papers. The only exceptions is ***P6***, in which the first two authors are in *co-first authorship*. The remaining authors appear in order of contribution.

P1 – R. Abgrall, P. M. Congedo, and G. Geraci, *A One-Time Truncate and Encode Multiresolution Stochastic Framework*. Journal of Computational Physics, 257: 19-56, 2014.

P2 – R. Abgrall, P. M. Congedo, and G. Geraci, *Toward a Unified Multiresolution Scheme in the Combined Physical/Stochastic Space for Stochastic Differential Equations.* Mathematics and Computers in Simulation (under review), 2012.

P3 – R. Abgrall, P. M. Congedo, G. Geraci and G. Iaccarino, *A novel weakly-intrusive non-linear multiresolution framework for uncertainty quantification in computational fluid dynamics.* Journal of Scientific Computing (under review), 2013.

P4 – R. Abgrall, P. M. Congedo, G. Geraci and G. Iaccarino, *An adaptive multiresolution semi-intrusive scheme for UQ in compressible fluid problems.* International Journal of Numerical Methods in Fluids (under review), 2013.

P5 – R. Abgrall, P. M. Congedo, G. Geraci and M.G. Rodio, *Stochastic Discrete Equation Method (sDEM) for two-phase flows.* Journal of Computational Physics (under review), 2013.

P6 – P. M. Congedo, G. Geraci, R. Abgrall, V. Pediroda, and L. Parussini, *TSI metamodels-based multi-objective robust optimization.* Engineering Computations, Vol. 30 Iss: 8, 2013.

P7 – R. Abgrall, P. M. Congedo, G. Geraci, and G. Iaccarino, *Computation and decomposition of high-order statistics.* Reliability Engineering & System Safety (under review), 2013.

International Conferences/Proceedings

Although not included among the papers of the thesis, the following conference proceedings, produced during this thesis work, are closely related to the papers listed above.

C1 – R. Abgrall, P. M. Congedo, and G. Geraci, *A high-order adaptive semi-intrusive finite volume scheme for stochastic partial differential equations.* Proceedings of the 13th Meeting on Applied Scientific Computing and Tools, MASCOT 2013, IMACS Series in Computational and Applied Mathematics, San Lorenzo de El Escorial, Spain, 26th-30th August 2013.

C2 – R. Abgrall, P. M. Congedo, D. De Santis and G. Geraci, *Stochastic analysis and robust optimization for a converging shock wave experimental setting.* Proceedings of the 13th Meeting on Applied Scientific Computing and Tools, MASCOT 2013, IMACS Series in Computational and Applied Mathematics, San Lorenzo de El Escorial, Spain, 26th-30th August 2013.

C3 – P. M. Congedo, D. De Santis and G. Geraci *On the predictive estimation of converging shock waves.* Congrès Français de Mécanique, CFM 2013, Bordeaux, France, 26th-30th August 2013.

C4 – P. M. Congedo, G. Geraci, R. Abgrall and G. Iaccarino, *Multi-objective design optimization using high-order statistics for CFD applications.* 10th International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, EUROGEN 2013, Las Palmas de Gran Canaria, Spain, October 7th-9th 2013.

C5 – R. Abgrall, P. M. Congedo, G. Geraci, *A high-order non-linear multiresolution scheme for stochastic PDEs.* European Workshop on High Order Nonlinear Numerical Methods for Evolutionary PDEs: Theory and Applications (HONOM 2013), Bordeaux, France, 18th-22nd March, 2013.

C6 – R. Abgrall, P. M. Congedo, G. Geraci and G. Iaccarino, *High-order statistics-based robust design optimization.* 11th International Conference on Structural Safety & Reliability, Columbia University, New York, USA, 16th-20th June 2013.

C7 – M. Ricchiuto, P. M. Congedo, G. Geraci and R. Abgrall, *Uncertainty propagation in shallow water long wave runup simulations.* 1st International Conference on Frontiers in Computational Physics: Modeling the Earth System, Boulder, USA, 16th-20th December 2012.

C8 – R. Abgrall, P. M. Congedo, G. Geraci and G. Iaccarino, *Adaptive strategy in multiresolution framework for Uncertainty Quantification.* Prooceedings of the 25th Summer Program 2012, Studying Turbulence Using Numerical Simulation Databases - XIV, NASA Center For Turbulence Research, Stanford University, USA, 2012.

C9 – R. Abgrall, P. M. Congedo, and G. Geraci, *Semi-intrusive and Multi-Resolution schemes for UQ.* SIAM 2012 Conference on UQ, Raleigh, North Carolina, USA, 2nd-6th April 2012.

C10 – R. Abgrall, P. M. Congedo, and G. Geraci, *Strategies for UQ and robust optimization.* Invited conference for UQ Seminars - NASA Center For Turbulence Research - Stanford University, USA, November 2011.

C11 – R. Abgrall, P. M. Congedo, and G. Geraci, *An adaptive Multiresolution Inspired Scheme for Solving the Stochastic Differential Equations.* Proceedings of the 11th Meeting on Applied Scientific Computing and Tools, MASCOT 2011, IMACS Series in Computational and Applied Mathematics, R.M. Spitaleri (Eds), IAC-CNR Rome, Italy, 19th-21st October 2011.

C12 – P. M. Congedo, G. Geraci, R. Abgrall, V. Pediroda, and L. Parussini, *Efficient ANOVA decomposition and metamodels-based multi-objective robust optimization.* Proceedings of the International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and

Social Problems EUROGEN 2011, C. Poloni, D. Quagliarella, J. Périaux, N. Gauger and K. Giannakoglou (Eds), pp. 493-504, Capua, Italy, 14th-16th September 2011.

C13 – R. Abgrall, P.M. Congedo, S. Galéra and G. Geraci, *Semi-intrusive stochastic method in high dimensions for aerospace applications.*, 4th European Conference for Aerospace Sciences, EUCASS 2011, Saint Petersburg, Russia, 4th-8th July 2011.

Inria Research Reports

Several research reports have been produced during this thesis work. Many of them are directly related to Journal papers or Prooceedings of International conferences, however three of them are related to supplementary investigative studies. These latter are listed in the following to make easier the reference into the manuscript.

RR1 – R. Abgrall, P. M. Congedo, G. Geraci and G. Iaccarino, *Non polynomial expansion for stochastic problems with non classical pdfs.* INRIA Research Report RR-8191, 2012.

RR2 – R. Abgrall, P. M. Congedo, and G. Geraci, *Numerical Investigation on the Total Sensitivity Index influence in the solution of stochastic partial differential equations.* INRIA Research Report RR-7911, 2011.

RR3 – R. Abgrall, P.M. Congedo, G. Geraci, *On the use of the Sparse Grid techniques coupled with Polynomial Chaos*, INRIA Research Report RR-7579, 2011.

Résumé étendu

Ce travail est focalisé sur les méthodes de quantification des incertitudes dans le contexte de la dynamique des fluides computationnelle. Ce manuscrit est divisé en deux parties. La première partie décrit les éléments principaux pour le développement d'une approche intrusif multi-résolution de la propagation des incertitudes. La deuxième partie du manuscrit inclut une collection de sept articles, qui contiennent les détails du travail développé dans le cadre de cette thèse.

Depuis les années 60, le développement des ordinateurs programmables a permis de coupler, à la théorie et les expériences, le calcul scientifique comme un outil supplémentaire de l'investigation scientifique. Actuellement, le calcul scientifique joue un rôle prédominant à la fois dans la recherche scientifique et dans la conception en ingénierie. Dans plusieurs cas, il n'est pas possible de refaire des expériences pour des systèmes complexes à gérer, ou dans des conditions particulières et potentiellement catastrophique (par exemple l'explosion d'un réacteur nucléaire ou l'effondrement d'un barrage).

L'impact du calcul scientifique a été considérablement augmenté au cours des dernières années grce au développement d'algorithmes numériques plus efficaces et la construction d'ordinateurs de plus en plus puissants. Cependant, la simulation numérique présente encore beaucoup de limitations pour être vraiment prédictive. Cette limitation vient de l'impossibilité de représenter, avec un système déterministe, un système physique générique. L'approche classique du calcul scientifique est basé sur la modélisation d'un système physique à l'aide d'un système d'équations qui décrivent le comportement physique. Ce système d'équations doit alors être traduit dans un modèle numérique avec des conditions d'entré. Un schéma de ce processus est montré dans la figure 1 dans lequel l'input y se traduit, à travers le modèle, dans le sortie $u(y)$.

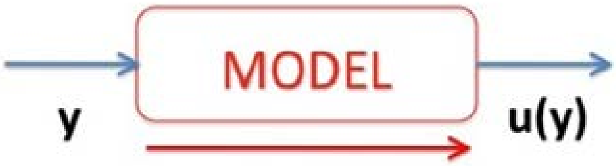


Figure 1: Schéma d'un modèle numérique classique avec une entrée déterministe. Figure reproduite à partir de [45].

Cependant, ce type d'approche est insuffisant pour la description des systèmes physiques. La modélisation de phénomènes complexes est inévitablement affectée par les limitations dans la connaissance exacte d'un système et dans la conception du modèle. Par exemple, l'aérodynamique d'un avion en vol transsonique est soumise à l'effet de nombreux facteurs qui ne peuvent pas être déterminées avec une précision absolue. Par exemple, il est impossible de connatre exactement les propriétés physiques et chimiques du fluide, la vitesse de l'avion, sa géométrie, etc. Il est clair que, en général, la prise en compte de paramètres qui sont caractérisés au niveau statistique, est indispensable. Ainsi, il est nécessaire un changement de perspectif dans la simulation numérique, pour prendre en compte la caractérisation statistique des input du système (voir la figure 2).

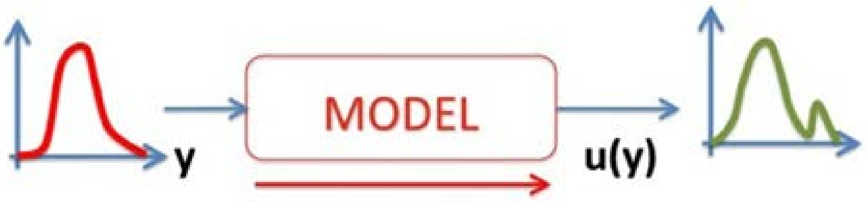


Figure 2: Schéma d'un modèle numérique général soumis à un input stochastique avec la caractérisation stochastique de la sortie. Figure reproduite à partir [45].

L'ensemble des activités, dédiées à l'analyse de l'effet de la présence d'incertitudes dans les simulations numériques, peut être appelé comme "quantification des incertitudes". Il y a trois étapes fondamentales à considérer:

- Caractérisation des sources d'incertitude;
- Propagation de l'incertitude;
- Caractérisation statistique de la sortie de la simulation.

Le travail présenté dans cette thèse a contribué à la quantification des incertitudes à travers deux actions. Le premier objectif était de développer des algorithmes efficaces pour la propagation des incertitudes pour des problèmes spécifiques à la dynamique des fluides. La deuxième contribution a été liée à la conception de stratégies pour l'optimisation robuste, avec un nombre modéré de paramètres incertains.

Le travail présenté dans cette thèse a été développé dans le but de contribuer à la quantification de l'incertitude dans le domaine de la dynamique des fluides. Dans ce contexte, la présence des incertitudes est très importante. Les problèmes liés à la dynamique des fluides sont caractérisés par des phénomènes non-stationnaires, multi-échelle et non-linéaires. Dans ce travail, un accent particulier a été mis sur les phénomènes liés aux fluides compressibles dans lequel l'apparition d'ondes de choc, peut provoquer le développement de régions relativement minces avec de gradients élevés de la solution. En particulier, on a considéré des lois conservation. Ces lois expriment la conservation de plusieurs quantités physiques à l'aide d'équations aux dérivées partielles de type hyperbolique. En particulier, les solutions d'équations de conservation de type hyperbolique, peuvent présenter des discontinuités, même avec des conditions initiales régulières.

La propagation des incertitudes, historiquement, a été traitée avec des méthodes basées sur l'échantillonnage (techniques de type Monte Carlo). Dans ce type de méthode, un code numérique est appelé plusieurs fois pour de différents jeux de paramètres. A la fin, les calculs sont post-traités pour pouvoir caractériser statistiquement la quantité d'intérêt qui est une sortie du calcul numérique. Ces familles de méthodes, bien que facile à appliquer et flexibles, peuvent être chères en terme de cot de calcul. Actuellement, deux types de méthodes sont utilisés: non-intrusive et intrusive. En ce qui concerne la propagation non-intrusive, dont les méthodes d'échantillonnage sont les archétypes, la caractérisation statistique de la solution est réalisée en lançant plusieurs fois un code de calcul pour des jeux de paramètres différents. Ces méthodes sont indispensables quand on ne peut pas accéder aux sources d'un code numérique, car le code est vu comme une bote noire. Pour résoudre le problème de propagation des incertitudes avec un cot de calcul plus faible, des techniques intrusives ont été récemment introduites. Ces techniques intrusives sont basées sur la modification du code numérique. Parmi les méthodes plus connues, on peut mentionner les techniques de chaos polynômial. Dans ce type de techniques, la solution est développée en série polynomiale et, en utilisant une projection de Galerkin, les coefficients sont calculés en résolvant un système d'équations (généralement) couplé. Dans cette famille de techniques, le nombre d'équations à résoudre est plus grand que celui du système initial et est lié essentiellement à la troncature de l'expansion choisie pour représenter les paramètres d'input. En même temps, cette famille de techniques peut présenter une perte de la convergence spectrale de la solution en présence de discontinuité.

Afin de résoudre à moindre cot le problème de la propagation des incertitudes dans le cas d'écoulements compressibles en présence de discontinuité, Abgrall et Congedo ont proposé une approche dite semi-intrusive (SI), dans laquelle, la solution est représentée, dans l'espace des paramètres, à travers une reconstruction de type volumes finis. Cette technique permet d'obtenir un schéma, dans lequel le nombre d'équations à résoudre ne change pas mais, qui permet également une grande flexibilité en terme de distributions de probabilités qui caractérisent les entrées du système. Cette généralité du schéma permet l'utilisation théorique de paramètres et des distributions arbitraires, par exemple à partir de mesures expérimentales.

La méthode semi-intrusif proposé par Abgrall et Congedo, même si permet de résoudre certaines limites des approches intrusives classique, reste très couteuse quand on 'augmente le nombre de paramètres incertains. Ce type de problème est connu dans la littérature comme "curse of dimensionality". Le but de ce travail de thèse a consisté dans la modification du schéma semi-intrusive pour obtenir un schéma numérique plus efficace. En particulier, on préserve les caractéristiques positives mentionnées précédemment, mais en plus cela permet la résolution des problèmes avec un nombre d'incertitudes modéré. En particulier, la contribution fondamentale de ce travail a été l'intégration d'une représentation multi-résolution dans le système semi-intrusive original.

Pour cette raison, l'approche multi-résolution proposé par Harten a été modifié pour l'utiliser dans le cadre de la quantification des incertitudes et utilisé. En particulier, une approche multi-résolution non-linéaire a été employée. Les opérateurs de reconstruction adoptés ont été formulées pour être dépendants des données. Dans ce contexte, une reconstruction d'ordre élevé et des techniques ENO/WENO ont été introduites. Premièrement, un schéma basé sur une représentation multi-résolution par points, a été employé pour construire une méthode basée sur une technique similaire à la collocation. Cette méthode, appelée spatial-TE, a été conçue pour conserver les propriétés de l'approche semi-intrusive, bien qu'elle soit différente en raison de l'application d'une reconstruction de la solution de type lagrangienne dans l'espace des paramètres. Le développement de la méthode sTE a permis de développer et de vérifier les algorithmes de cette méthode intrusive basée sur la multi-résolution dans le contexte d'équations aux dérivées partielles, en présence de paramètres et de solutions aléatoires discontinues.

Le second résultat obtenu pendant cette thèse a été le développement d'une nouvelle méthode adaptative semi-intrusive (adaptive-SI), avec une représentation de type volumes finis et avec une approche multi-résolution. L'effet de l'introduction de la multi-résolution est double. Le premier effet est de réduire la dimensionalité de l'espace de représentation de fonctions discrètes dans l'espace stochastique. Cette réduction permet un gain, en terme d'efficacité de calcul, en diminuant le nombre d'évaluations explicites du modèle pour représenter les variables dans l'espace stochastique. En même temps, la multi-résolution est la base pour représenter les fonctions exploitant la localité de la base. L'approche multi-résolution est un outil naturel pour étudier la régularité locale d'une fonction et peut être utilisée pour définir une procédure d'adaptation. Le deuxième avantage, lié à l'introduction de la multi-résolution, est de pouvoir définir une procédure générale de raffinement et de-raffinement du maillage dans l'espace stochastique dans le cas de problèmes instationnaires. La caractéristique principale de l'approche proposée est la possibilité d'adapter la discrétisation de l'espace stochastique, en fonction des coordonnées physiques et spatiales. Ce type d'adaptation est particulièrement adapté pour des applications dans lesquelles des phénomènes multi-échelle avec des gradients et/ou des discontinuités qui se propagent dans l'espace peuvent apparatre, comme c'est le cas pour la dynamique des fluides.

Les deux schémas proposés, sTE et aSI, ont été appliqués pour modéliser des problèmes à difficulté croissante. Les premiers cas ont été des fonctions-test qui ne dépend que de l'espace des paramètres. D'autres problèmes ont été des systèmes d'équations différentielles. Dans le cas de l'approche sTE, les équations différentielles ordinaires (tels que le système non-linéaire de Kraichnan-Orszag) et partielles ont été résolues. En plus, le schéma sTE a été également appliqué au système d'équations qui régit la propagation des ondes élastiques dans un matériel hétérogène. Tous les problèmes résolus ont été formulés dans un espace de paramètre 1D. Pour le schéma aSI, des équations aux dérivées partielles ont été résolues: équation d'advection linéaire, équation de Burgers et le système d'Euler avec d'espaces de paramètres 1D, 2D et 3D (dans le cas de l'équation d'advection). Dans tous les cas testés, les schémas proposés ont été comparés avec des approches non-adaptatives pour montrer l'augmentation de l'efficacité et une réduction de cot de calcul quand on utilise une approche multi-résolution.

La deuxième contribution de cette travail de thèse est la conception d'une méthode d'optimisation robuste en présence d'un nombre modéré d'incertitudes. En particulier, une stratégie basée sur la réduction de la taille de l'espace stochastique à travers une analyse stochastique des indices de sensibilité ANOVA, a été proposé. La procédure est constituée de deux étapes. Dans la première étape, une surface de réponse est construite pour les indices de sensibilité, pour chaque variable stochastiques du problème, dans l'espace d'optimisation. Ensuite, pendant le processus d'optimisation basé sur un algorithme génétique, la surface de réponse est utilisée pour fournir des indices de sensibilité pour un individu de l'espace d'optimisation. En utilisant un critère de réduction pour réduire la dimension stochastique, pour chaque individu de l'espace d'optimisation, en connaissant les indices de sensibilité, il est possible de réduire le cot globale de l'analyse stochastique. L'évaluation de la fonction objectif, est une fonction des statistiques associées à l'individu. L'évaluation des statistiques peut donc être effectuée sur une espace stochastique de taille réduite, ce qui nécessite, pour la même précision, un nombre d'évaluations inférieur. Cette procédure a été proposée en utilisant un critère qui était lié à la contribution de chaque incertitude à la variance globale. La stratégie a été appliquée avec succès à un problème d'optimisation robuste multi-objectif d'une turbine à gaz réel.

Afin d'améliorer la qualité du critère de réduction de la dimension stochastique, de différents critères de réduction ont été étudiés et proposés. En particulier, l'approche de décomposition de la variance a été étendue au cas de moments centraux de troisième et de quatrième ordre. On a pu vérifier que l'effet de réduction de la dimension sur la base de la variance pourrait négativement affecter le calcul de la queue de probabilité. On a donc proposé des critères plus robustes pour la réduction dimensionnelle basés sur les moments d'ordre supérieur. Pour le futur, l'objectif est d'intégrer ces nouveaux critères de réduction dimensionnelle dans le processus d'optimisation robuste.

Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man. Every careful measurement in science is given with the probable error, which is a technical term, conveying a precise meaning. It means: that amount of error which is just as likely to be greater than the actual error as to be less. It is a characteristic of those matters in which something is known with exceptional accuracy that, in them, every observer admits that he is likely to be wrong, and knows about how much wrong he is likely to be.

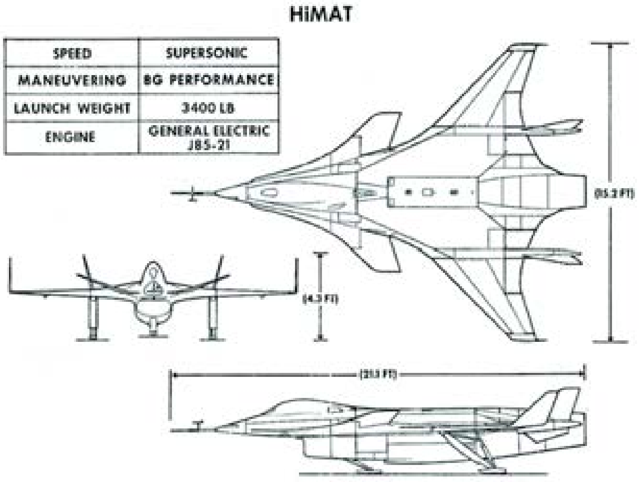
Bertrand Russell, *The Scientific Outlook*, 1931.

Introduction and motivation

Since ancient times, the primary method to design has been the so-called *trial & error* approach. Moving from theoretical observations for preliminary design, during and after the production of a new system, a number of tests were conducted to understand its characteristics and responses. Around 1700 the calculus was introduced and the era of the mathematical modeling of physics slowly began the main way of analyzing a complex system. However, only around the 1960s programmable digital computers started to widespread in academic and industrial organizations. The *scientific computing* era finally took place. This initial stage, of the modern *scientific computing*, also coincided with the growth in aerospace and military applications. One of the angular stone of the development of the scientific computing in its modern view is the paper *Computer Experiments in Computational Fluid Dynamics* [37] where, for the first time, the *scientific computing* has been recognized as the third pillar of science, along with theory and experiments. Computational fluid dynamics (CFD) played a central role already in the first phase of the development of the *scientific computing*. One of the first success, in the computational aerodynamics context, was the design of the aircraft called Highly Maneuverable Aircraft Technologies (HiMAT) (see figure 3) designed by the National Aeronautics and Space Administration (NASA). This aircraft has been designed, in the late '70, to test concepts of high maneuverability for the next generation of fighter planes [14]. Once built, the wind tunnel tests showed that it would have unacceptable drag at speed near the speed of sound. The cost of redesigning using wind tunnel tests would have been around $150,000. Redesigning entirely the wing, by means of computer simulations, cost has been of $6,000.

The impact of the *scientific computing* has also increased with an astonishing speed. Nowadays is well established for safety and reliability applications or to reduce the time (and the cost) of products design. For instance, specialized mathematical model can be formulated and solved to obtain data instead of carrying out expensive or impractical experimental tests. Actually, in many advanced technological sector, as aerospace, military systems, automotive, nuclear devices *etc.*, *scientific computing* plays an even increasing supporting role. This role can be to support experimental applications, or to analyze complex systems which experimental set-up would be impractical or even impossible. For instance a failure of a dam or the explosion of a nuclear reactor. However, the *scientific computing* cannot be truly *predictive* without a coupling between theory and experiments; this coupling procedure is called *validation* in [77]. The question of the credibility of the *scientific computing*, under these circumstances, becomes fundamental.

Even if in the last four decades a strong effort has been devoted to develop efficient, faster and more accurate numerical algorithms, the challenging question of the confidence of the scientific computing is still demanding a final answer. As pointed out in [65], only in more recent years the problem of the credibility of the scientific computing has received the attention it deserves. Again the CFD has been chosen



SPEED	SUPERSONIC
MANEUVERING	8G PERFORMANCE
LAUNCH WEIGHT	3400 LB
ENGINE	GENERAL ELECTRIC J85-21

Figure 3: HiMAT experimental vehicle designed in the late '70 to test high-maneuvering conditions. Its wing has been entirely designed by numerical simulation after a not fully satisfactory initial design. Figure reproduced from *BOEING FRONTIERS*, May 2007.

as subject of a research study. In 1986 the NASA funded a study to evaluate the maturity of the CFD computational framework, for different physical models and increasing complexity engineering systems. The result was quite surprising showing that the degree of maturity, *i.e.* development of the CFD techniques, could vary enormously changing the kind and the complexity of the applications.

To understand the actual level of confidence of scientists in the *scientific computing* context and the need for a systematic investigation of the quality of the computational results, Oberkampf and Roy in their book [65] wrote

> Claims of high maturity in CFD for complex systems, whether from commercial software companies or any other organization, are, we believe, unfounded. Companies and agencies that sell programs primarily based on colorful graphics and flashy video animations have no skin in the game. We also claim that this is the case in essentially all fields of scientific computing.

To rigorously assess the quality and accuracy of computational results, the framework of Verification & Validation has been developed and nowadays it is receiving an always increasing attention by the numerical community. One of the core aspect of the Verification & Validation framework is the emerging field of the Uncertainty Quantification which is the context of this thesis. The Verification & Validation (V&V) framework is the subject of the next section.

Verification & Validation: from the general context to the numerical paradigm

The context of this research work is the *scientific computing* and, in particular, the CFD. However, from an historical perspective, the *scientific computing* borrowed (with some obvious adaptations) fundamental concepts already well established in the engineering field. For instance, in a paper published in 2004 Bahill and Henderson presented consolidated definitions of V&V for both requirements and systems [81]. As natural, an engineering system is the (complex) physical counterpart of a (simplified) physical system represented by mathematical models. The numerical solution of mathematical models is the subject of the present thesis, hence it is interesting to consider the evolution of the basic concepts, related to the (engineering) systems, to make evident their correct declination in the *scientific computing* context. Quoting Bahill and Henderson [81]

> Verifying a system: building the system right, ensuring that the system complies with the system requirements and conforms to its design.
>
> Validating a system: building the right system, making sure that the system does what it is supposed to do in its intended environment. Validation determines the correctness and completeness of the end product, and ensures that the system will satisfy the actual needs of the stakeholders.

The two definitions deal with the need of achieving certain requirements (Verification) and the correct context of employment of a system (Validation). The importance of both procedures is made evident from the analysis of famous failure of complex systems for which, the authors in [81], collect the causes in three categories: requirements development (RD), verification (VER) and validation (VAL) of the system. The figure 4, extracted from [81], reports the table with the analysis of these famous failures.

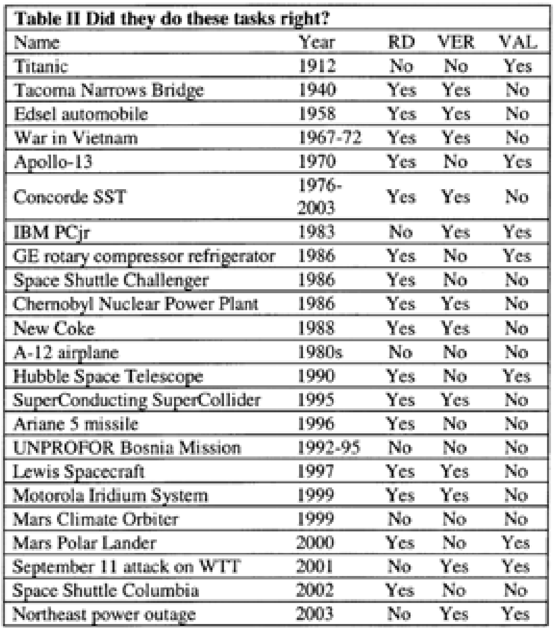


Table II Did they do these tasks right?

Name	Year	RD	VER	VAL
Titanic	1912	No	No	Yes
Tacoma Narrows Bridge	1940	Yes	Yes	No
Edsel automobile	1958	Yes	Yes	No
War in Vietnam	1967-72	Yes	Yes	No
Apollo-13	1970	Yes	No	Yes
Concorde SST	1976-2003	Yes	Yes	No
IBM PCjr	1983	No	Yes	Yes
GE rotary compressor refrigerator	1986	Yes	No	Yes
Space Shuttle Challenger	1986	Yes	No	No
Chernobyl Nuclear Power Plant	1986	Yes	Yes	No
New Coke	1988	Yes	Yes	No
A-12 airplane	1980s	No	No	No
Hubble Space Telescope	1990	Yes	No	Yes
SuperConducting SuperCollider	1995	Yes	Yes	No
Ariane 5 missile	1996	Yes	No	No
UNPROFOR Bosnia Mission	1992-95	No	No	No
Lewis Spacecraft	1997	Yes	Yes	No
Motorola Iridium System	1999	Yes	Yes	No
Mars Climate Orbiter	1999	No	No	No
Mars Polar Lander	2000	Yes	No	Yes
September 11 attack on WTT	2001	No	Yes	Yes
Space Shuttle Columbia	2002	Yes	No	No
Northeast power outage	2003	No	Yes	Yes

Figure 4: Famous failures of complex systems and their causes. Table reproduced from [81].

Some systems, reported in figure 4, failed for multiple causes, but several among them failed for only one of the aspects considered. This behavior clearly confirms the complementarity of the procedures of V&V for complex systems and also motivated the more recent effort in the Validation technique with respect to the more consolidated framework of Verification. Verification still remains a core activity for design, but needs to be supported by the other Validation activities.

Once recognized the importance of the V&V framework, how this methodology has been adopted in the scientific computing community? The first effort, in term of both methodology and terminology, has been realized by the community of operations research. This community in 1960s faced out to the development of the general concepts having in mind the application in systems involving many, intrinsically hard to model, *phenomena* such as the interaction between the human behavior and physical systems or computed controlled systems. The novelty has been, for the first time, the key role credited to the computerized model. The seminal paper has been published in 1979 by Schlesinger [73]

> Model verification: substantiation that a computerized model represents a conceptual model within specified limits of accuracy.
>
> Model validation: substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.

Due to the nature of the systems the definition appears, nowadays, vague. However, the changing in the scientific thinking took place and, after the beginning of the massive employment of computed-controlled system during 1970s, the Institute of Electrical and Electronics Engineers (IEEE), between 1984 and 1991, and then the US Department of Defense (DoD) in 1994 led to more refined definitions

> Verification: the process of determining that a model implementation accurately represents the developers conceptual description of the model.
>
> Validation: the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

In the last definitions key aspects finally emerge: the model implementation, its accuracy and the real world environment. For the purpose of this presentation, the last concepts naturally lead to the definitions adopted by the CFD community after six years (1992-1998) of work and debate committed by the American Institute of Aeronautics and Astronautics (AIAA):

> Verification: the process of determining that a model implementation accurately represents the developers conceptual description of the model and the solution to the model.

The previous definition adds the missing element: the accuracy must regard the numerical solution of the model. The Validation definition adopted by the AIAA has been borrowed from the DoD definition, but its application is strictly different. The AIAA guide *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations* [11] explicitly refers to the experimental data to represent the real world. Moreover, the self-explaining concept of *prediction* has been formally defined in [11]:

> Prediction: use of a computational model to foretell the state of a physical system under conditions for which the computational model has not been validated.

The definition of *prediction* contains all the elements which make clear the motivation for a research work as like the one conducted in this thesis: the final aim of the *scientific computing* should be to foretell (without knowing the results) the state of a physical system in conditions different from the set of Validation test cases. Validation procedures alone cannot furnish the evidence of the predictive capability of a numerical procedure, but they represent a reproducible evidence that a prescribed accuracy can be reached for specified problems. The contrast, even if sometimes ignored, with the common calibration techniques should be evident. Even if the ubiquity of calibration always induces an impact on the confidence related to the predictive capability of the numerical simulations, without doubt the calibration procedure should not be intended as a 'technique' ables to hide the weakness of a model: *[calibration] is like drawing the bulls-eye afterwards, whereas in prediction the target is there in advance* [56].

Less formally, the following questions help to clarify the previous concepts when dealing with a physical system represented by its mathematical form [45]:

> Verification: are we solving the equations correctly? (it is an exercise in mathematics)
>
> Validation: are we solving the correct equations? (it is an exercise in physics)

The range of interest of the two procedures is quite different; while the Verification procedure deals with the solution of a prescribed set of equations by numerical algorithm, only Validation can assure that the simplified representation (mathematical model) is adequate to this purpose. Therefore, it follows that the adequate model cannot exist without fixing a purpose. In this respect, the fluid dynamics context is quite different from other physical (at least of the classical physics) branches. The fluid flow problems are intrinsically rich of multiscale *phenomena* (small turbulent scales, shock waves *etc.*) and many different behaviors can occur in apparently very similar situations. For instance, the laminar and turbulent regimes could take place for the same phenomenon, the same object immersed in a flow for instance, changing only (even not too much) the free stream velocity in a wind tunnel experiment or the properties of the fluid, *i.e.* its viscosity. Moreover, the assumptions made to model a fluid flow phenomenon are quite important to circumvent the range of application of a model and, consequently, to determine the adequacy of its numerical counterpart. For instance, the same fluid, for example air, could be modeled with an increasing level of complexity from a thermodynamics point-of-view, ranging from ideal gas hypothesis to the inclusions of real gas effects or also ionization and dissociations of molecules. In many cases fluid flow simulations are dependent on models and lack of knowledge. A relevant class of fluid flow problems, where the lack of knowledge is always present, are the turbulent flows. Quoting Pope [69]

> In any turbulent flow there are, unavoidably, perturbations in initial conditions, boundary conditions and material properties.

As it is evident, in any turbulent flow, only a statistical description is of interest. This aspect, among many others, strongly introduces the importance of the randomness in fluid dynamics and its numerical counterpart, the CFD. More generally, it is not forced to state that, virtually, any kind of fluid flow is subject to randomness (boundary conditions, initial conditions, thermodynamics properties, chemical composition, manufacturing tolerances *etc*). While this should be true even for many other fields, the intrinsically non-linear nature (shock waves, complex thermodynamics behaviors *etc.*) of the fluid dynamics makes this aspect crucial. Numerical fluid dynamics simulations are nowadays demanded to face out this complex situation interacting, more deeply, with experimental settings.

This brief introduction has been conceived for introducing some highlights of the UQ framework, particularly for the CFD. The knowledge of many different *phenomena*, as said, is intrinsically non deterministic, while naturally computer simulations, provided necessary input data, are deterministic. Even the experimental data, whose constitute the comparison set for Validation purpose, cannot be stated as deterministic measure, but very often they can be viewed as known only in a statistical sense. The complete management of the randomness from its characterization (in the input data) to its propagation (in the computer model) and its analysis, in term of outputs, is the subject of UQ. UQ itself represents a broad field undergoing an outstanding evolution in recent years involving competence flowing from different areas. Iaccarino sketched [45] the concurring areas in the figure 5.

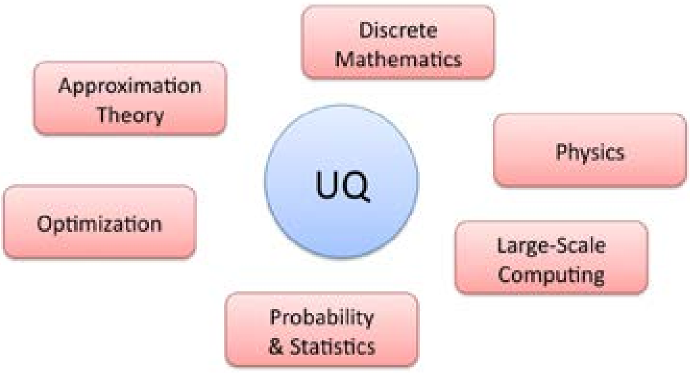


Figure 5: Areas concurring to the Uncertainty Quantification analysis (from [45]).

The description of the UQ, in its general sense, is the subject of the next session.

Uncertainty quantification

The first concept related to the UQ procedure is its object of analysis: the uncertainty. It could seem obvious, but an uncertainty it is not an error and this difference very often can generate misunderstandings in the *scientific computing* context as well as in other fields. The AIAA [11] defined also the concepts of error and uncertainty: the error is a recognizable deficiency of the model or algorithm employed, while the uncertainty is a potential deficiency due to a lack of knowledge. These definitions are hard to decline in the context of *scientific computing* because there is no separation between what involves the mathematics and what involves the physics. Iaccarino in [44] identified the error as an entity associated to the translation of the mathematical formulation into a numerical algorithm and its counterpart given by the numerical code. Uncertainties are instead related to the choice of the physical model, including the input parameters required. The errors can arise form, for instance, round-off, limited convergence of iterative algorithms, but also form implementation or usage mistakes. Examples of uncertainties are, for instance, the flight conditions of a cruising airplane, the chemical composition of the atmospheric layer surrounding a planet, the initial conditions relative to a shock tube problem, the thermodynamical properties of a fluid and so on. Uncertainty is then associated to the concept of non deterministic in a broad sense. Uncertainty is a concept related to the variability of the system, but the variability itself can bring within different information. If the variability present in the system is not related to a lack of knowledge, then it cannot be reduced and a probabilistic framework is adequate to represent it (in this case the uncertainty is referred as of a stochastic type). For instance, the input parameters of a physical experiment can be known only by a probability density function over an interval. Otherwise, if the variability is related to a potential lack of knowledge, it is called epistemic uncertainty. In the latter case the variability can arise from (strong) assumptions introduced during the derivation of the mathematical model and, in principle, it could be reduced improving the model. A paradigmatic example in CFD could be the level of complexity employed to represent the turbulence equations (Reynold-Averaged Navier-Stokes, Large Eddy Simulations, DNS). Again, it is important to remark that the quality of the model cannot be determined *a priori*, but only after that certain requirements are fixed. Despite the broad range of applications, the UQ is commonly based on three fundamental activities

- Characterization of the sources of uncertainty;
- Propagations of the uncertainties in the input data/model through the model itself;
- Analysis of the model outputs.

Regarding the characterization of uncertainties both direct and inverse methods are at disposal. Among the direct methods, the experimental observations, theoretical arguments and also expert opinions are the principal; inference and calibration are, instead, the fundamental techniques to translate observed data in statistical input parameters. The propagation process is the ensemble of operations which allows to obtain the uncertain outputs prescribing, to the uncertain model, the uncertain inputs. In this work the focus is totally devoted to aleatory, *i.e.* stochastic uncertainties, so in this case the probability framework plays a special role not only for the characterization of the sources of uncertainties, but also for their propagation. The final UQ step involves the analysis of the ensemble of output quantities of interest in term of probability distributions, cumulative distributions or other statistic characterization of data. Moreover, also sensitivity analysis (SA) of the data could be of interest. In the Part II of this thesis examples, of how the SA analysis can be employed to improve the understanding of complex system, will be provided. However, even if they are related, it is important to underline the difference between UQ and SA. Sensitivity analysis investigates the connection between inputs and outputs of a model and, in particular, it makes possible to relate the variability of the outputs to the variability of the inputs. SA does not need input data and can be conducted on purely mathematical analysis, while the UQ, given a system, aims to quantify its output uncertainty. The meaning of SA in the design context appears clear: large variations of some (identified) parameters generate large variations of the outputs. In the UQ context, however, a large sensitivity of a parameter is not strictly connected to large uncertainties. In fact uncertainties related to input parameters, to which the system is greatly sensitive, could be so small to induce no uncertainty on the outputs at all. Moreover, it could be important determining how the uncertain structure of all the inputs maps the uncertain structure of all the outputs. In this case the SA is named global SA and its importance becomes crucial in the light of improving the quality of the model itself identifying the sources of a lack of knowledge. For instance, in the case of complex coupled physics *phenomena*, a global SA can indicate the relevant physical experiments to conduct to most reduce the epistemic uncertainty.

The present thesis illustrates some new numerical algorithms and techniques to propagate and analyze uncertainties in the stochastic framework with particular emphasis to fluid flow applications. From a mathematical point-of-view, it is universally recognized the importance of the hyperbolic problems for CFD applications; this class of problems embeds many fundamental characters of more complex situations like the propagation of waves and the possibility to handle (forming) discontinuity (as for example in transonic aerodynamics simulations). For these reasons, the work is mainly focused on this paradigmatic class of problems having in mind a virtually easy extension to more complex models.

Hyperbolic problems, and CFD applications in general, result to be very challenging for UQ analysis even in presence of efficient and well-established deterministic schemes. The reason is the high computational cost, even employing parallel algorithms, to perform very accurate deterministic simulations in presence of complex physics (turbulence, shock waves, unsteadiness *etc.*). As it will be clear in the next chapter, the UQ analysis requires, at least simplifying, the knowledge of the numerical solution in a set of points. If the solution in each points corresponds to a CFD simulations, it is easy to see that the UQ analysis can be very often prohibitive. This problem is closely connected with the cubature, *i.e.* the numerical integration of a function in a multi-dimensional space. It is also well known that the number of degree-of-freedoms needed for a such a numerical problem, employing a tensorization approach, increases exponentially with the number of dimensions. The number of dimensions, *i.e.* independent random parameters of the problem, plays a fundamental role and leads rapidly to intractable problems in real CFD application cases. Richard Ernest Bellman referred to the exponential growth of data in multidimensional spaces coining the term *curse of dimensionality* in [21]. Actually, the *curse of dimensionality* is an open problem for different fields, as for instance, numerical analysis, sampling, combinatorics, machine learning, data mining and databases; this issue is also common to virtually any kind of UQ propagation technique at the state-of-the-art and, only with the introduction of the adaptive strategies in recent years, it has been preliminary addressed. One of the key contributions of the present work is precisely the introduction of a multiresolution adaptive procedure, for a novel semi-*intrusive* UQ propagation approach, aiming to tackle the *curse of dimensionality*.

Thesis Outline

This manuscript is based on seven papers previously introduced. In different chapters, the central ideas, which drive this thesis work, are presented and discussed. It is important to remark that all the details of the algorithms and numerical examples are reported in the joint papers. In particular, in Chapter 1, the state-of-the-art of the UQ propagation techniques is introduced. Both *non-intrusive* and *intrusive* techniques are presented; particular attention is devoted to the introduction of the so-called the semi-*intrusive* scheme, which constitutes a pillar of the present work. Chapter 2 presents the introduction of the Harten's inspired multiresolution framework adopted, in both point-value and cell-average settings. Chapter 2 contains also the introduction of the sTE scheme, *i.e.* the point-value collocation based approach developed in this work. The cell-average approach is then formulated for yielding a time-dependent adaptive semi-*intrusive* scheme in the overall physical/stochastic space, based on the semi-*intrusive* scheme recalled in Chapter 1, *i.e.* the aSI scheme. The aSI scheme is fully presented in Chapter 3. Several numerical test cases are finally presented and discussed in Chapter 4 for the point-value and the cell-average schemes. The numerical test cases, reported in Chapter 4, contain only results not yet published and not reported in papers.

This manuscript does not contain a presentation of what is reported in papers ***P6*** and ***P7***. These papers deal with UQ propagation and analysis and the identification of the most influent uncertainties in problems with a moderate number of parameters. The two main contributions are illustrated. In paper ***P6***, a novel robust optimization strategy, constituting in a dimension reduction of the set of uncertainties by using variance ANOVA decomposition, is presented. This strategy is applied to the design of a turbine with complex real gas thermodynamics. Moreover, paper ***P7*** presents a potential extension of this work to high-order statistical moments. New sensitivity indexes, based on the identification of the conditional contributions related to high-order central moments of third and fourth order, are introduced. The importance of this analysis is highlighted on several numerical examples.

Part I: Propagation techniques/schemes for Uncertainty Quantification

CHAPTER 1

Uncertainty propagation: state of the art

In this chapter, after a brief introduction of the mathematical setting of the problem in §1.1, some techniques for the uncertainty propagation are presented in §1.2. Then, in section §1.3, the semi-intrusive (SI) approach, developed more recently by Abgrall and co-authors [3,4] (and ***C13***) is introduced. The semi-intrusive approach represents the core of the adaptive-SI scheme which is one of the final accomplishments achieved of this thesis.

In the following section, the mathematical framework on which the stochastic UQ analysis is commonly based, is described. Although, many test cases presented in the papers on which the thesis is based, are only simplified model equations, the focus of the work is devoted to Computational Fluid Dynamics (CFD) simulations. In the CFD context partial differential equations (PDEs) are of interest, so, despite the applicability of UQ technique to very general problems, its generic presentation is carried out having in mind PDEs.

1.1 Mathematical setting

Mathematical theory of probability furnishes the basis of the statistic and, in its modern (axiomatic) vision, following the work of Kolmogorov [48], extensively relies on the measure theory. The basics concepts for the measure theory are the measure and the measurable space. The measurable space can be defined giving a sample space Θ and a (non-empty) collection of its subsets Σ. The collection Σ is a $\sigma-$algebra (or $\sigma-$field) if

- $0 \in \Sigma$;
- $A \in \Sigma \Rightarrow \bar{A} \in \Sigma$ ($\bar{A}$ is the complement of A);
- $A_i \in \Sigma$ for all $i \in I \Rightarrow \bigcup_{i \in I} A_i \in \Sigma$.

The pair (Θ, Σ) is a measurable space. An example of $\sigma-$algebra, corresponding also to the smallest possible $\sigma-$algebra, is $\{0, \Theta\}$. A specific $\sigma-$algebra is the Borel $\sigma-$algebra in which the $\sigma-$algebra is generated by all the open sets of a topological set. This concept becomes functional to the probability theory if the sample space is the real line $\mathbb{R}$ (or one of its sub-interval). In this case the Borel $\sigma-$algebra becomes $\mathcal{B}(\mathbb{R})$ and form a measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the real line. A measure μ is a real-valued non-negative function defined over a set with the following properties

- $\mu(A) \geq 0$ for $A \in \Sigma$;
- Countable additivity: if $A_i \in \Sigma$ are disjoint sets then $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$.

A measure μ is called probability measure if $\mu(A) \in [0,1]$ and $\mu(\Theta) = 1$. A measurable space, equipped with a measure, it is called measure space (Θ, Σ, μ). Given a couple of measurable space (Θ, Σ) and (Θ', Σ') a measurable function $g : \Theta \to \Theta'$, for any event $A' \in \Sigma'$, is

$$g^{-1}(B) \overset{\text{def}}{=} \{x : g(x) \in A'\} \in \Sigma. \tag{1.1}$$

The measurable function g induces a probability measure $\mathbb{P} : \Sigma' \to [0,1]$ on (Θ', Σ')

$$\mathbb{P}(A' \in \Sigma') \overset{\text{def}}{=} \mu(g^{-1}(A')); \tag{1.2}$$

the probability induced by g it is also called its distribution. If the measurable function is $\xi : \Theta \to \mathbb{R}$, between the couple of measurable spaces (Θ, Σ) and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then ξ is a random variable. The random variables specify a set of events that happen with a corresponding probabilities; if there are some events $\{\theta\}$ that ξ maps to $B \in \mathbb{R}$ then the probability of B it the total probability of those event $\{\theta\}$. The distribution of a random variable ξ is then the probability measure $\mathbb{P}$ induced by the mapping ξ itself on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

$$\mathbb{P}(A) = \mu(\xi^{-1}(A)) = \mu(\theta : \xi(\theta) \in A). \tag{1.3}$$

As shorthand the induced probability $\mathbb{P}(A)$ it is common not written at all and replaced by $\mu(\xi \in A)$. Even if the probability induced by the random variable is called distribution, more useful concepts of distribution can be introduced. The cumulative distribution function F of a random variable ξ is defined as

$$F(x) = \mu(\xi < x), \tag{1.4}$$

from which follows $\mu([a,b]) = F(b) - F(a)$. Very often random variables can be known in term of probability density functions $p(\xi)$

$$F(b) - F(a) = \int_a^b \mathrm{d}\mu = \int_a^b p(\xi)\mathrm{d}\xi, \tag{1.5}$$

where $p(\xi) > 0$, it is a Lebesgue measurable function and F is absolutely continuous. All the concepts presented can be extended, in a straightforward manner, to random vectors $\boldsymbol{\xi} \subset \Xi \subset \mathbb{R}^d$ where all the components are random variables and the corresponding probability space Ξ can be obtained as product probability spaces of the probability spaces associated to each random variable, *i.e.* it is the support of the joint probability density function. In this thesis the case of independent inputs will be considered. This assumption reflects in the possibility to obtain the joint pdf $p(\boldsymbol{\xi})$ as a product of the probability distributions of each random variable

$$p(\boldsymbol{\xi}) = \prod_{i=1}^{d} p_i(\xi_i). \tag{1.6}$$

However, in many situations expansions in terms of independent variables may not be possible. Many limitations occur in a such situation, as for instance, the need to an analytical knowledge of the joint pdf; the numerical method designed in this thesis are not based on the independent distribution assumption, however all the numerical tests have been performed in this situation. Hereafter the independence of the random parameters is always implicitly stated.

The mathematical setting of probability, briefly sketched above, allows to generalize the problem statement for a generic PDE including the presence of randomness. In a general case the randomness can affect the PDE itself through, for instance, some coefficients not exactly known or even initial and boundary conditions. Moreover, the randomness can have a spatial variability and/or a time variability to which can correspond a variability in term of probability density function in space and time. In general, an unknown u, depending on the physical space $\mathbf{x} \in \Omega \subset \mathbb{R}^n$ and the time space $t \in T \subset \mathbb{R}^+$ and irrespective from the sources of uncertainties (if any), it will be itself a function of the random parameter $\boldsymbol{\xi}$ employed to describe the sources of uncertainties. In a case like this it is possible to generally state

$$\mathcal{L}(\mathbf{x}, t, \boldsymbol{\xi}; u(\mathbf{x}, t, \boldsymbol{\xi})) = \mathcal{S}(\mathbf{x}, t, \boldsymbol{\xi}) \tag{1.7}$$

where $\mathcal{L}$ is a differential/algebraic operator and $\mathcal{S}$ a source term, both defined on the domain $\Omega \times T \times \Xi$. The operator $\mathcal{L}$ can involve differentiations in space and time and can be non linear. Obviously, mathematical well-posed problems are obtained imposing proper boundary and initial conditions. One of the aims of the UQ analysis is to obtain a statistic characterization of the unknown u (or eventually other variables of interest referred in the following as outputs). Depending on the scope of the analysis, the statistical outcomes could be different. For instance, it could be necessary to compute the probability distribution of the output or, for instance for reliability and safety applications, the failure probability or the probability to exceed a certain level and so on. However, very often, a characterization in term of statistic moments is a first step. In all the numerical test cases presented in this work, except for those concerning explicitly the high-order moments, the quantities of interest, systematically computed on the output u, have been the expected value $\mathbb{E}(u, \mathbf{x}, t)$ and the second central moments, *i.e.* the variance $\mathrm{Var}(u, \mathbf{x}, t)$

$$\begin{aligned} \mathbb{E}(u, \mathbf{x}, t) &= \int_{\Xi} u(\mathbf{x}, t, \boldsymbol{\xi}) p(\boldsymbol{\xi}, t) \mathrm{d}\boldsymbol{\xi} \\ \mathrm{Var}(u, \mathbf{x}, t) &= \int_{\Xi} \left(u(\mathbf{x}, t, \boldsymbol{\xi}) - \mathbb{E}(u)\right)^2 p(\boldsymbol{\xi}, t) \mathrm{d}\boldsymbol{\xi}. \end{aligned} \tag{1.8}$$

In the following section a brief review of the state-of-the-art of the well established techniques for the propagation of uncertainties will be presented.

1.2 State of the art

In this section, a brief review of the numerical methods developed in recent years for UQ propagation is presented. Two alternative philosophies took place for UQ propagation, namely the *intrusive* and *non-intrusive* approach. Propagating uncertainties in the framework of the *scientific computing* gives, in principle, an alternative to the employment of multiple calls of the numerical model: the modification of the numerical code itself to obtain the propagation of the uncertainties. Unlike the experimental counterpart, in which the only possibility is to run experiments (realizations) of the physical system with different input data, when a mathematical model is at disposal another approach could be possible. In some situations the mathematical model itself can be modified to obtain, as outputs, statistical properties of interest, instead of only a realization of the system. In this latter case the input data are constituted by the representation of the input variables, for instance an interval with a probability distribution, and the approach is said to be intrusive, because it intrusively propagates uncertainties in the model requiring the modification of the model itself. From a computational point-of-view, the difference between the two approaches is evident. The *non-intrusive* approach requires only multiple runs of the numerical code, while an *intrusive* approach could require to reformulate entirely also the theoretical formulation of the problem other than requiring a more deep modification of the numerical code.

1.2.1 Sampling methods

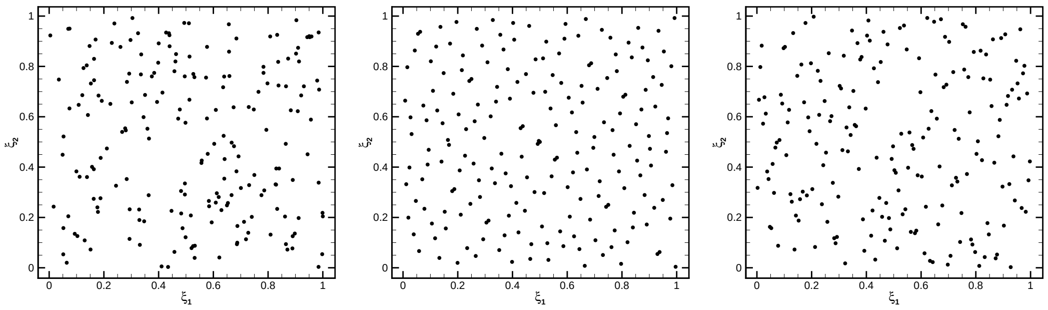
The archetypal method for the *non-intrusive* class is the Monte Carlo (MC) approach [64]. The principles and the application of the MC methods are very simple. A set of random points $\{\boldsymbol{\xi}_i\}_{i=1}^N$ in the stochastic space is generated in accord to the probability distribution of the uncertain parameters. Multiple runs of the code are performed, to obtain $u_i = u(\boldsymbol{\xi}_i)$ and the statistics are finally computed as

$$\begin{aligned}\mathbb{E}(u) &\simeq \frac{1}{N}\sum_{i=1}^N u_i \\ \mathrm{Var}(u) = \mathbb{E}(u^2) - (\mathbb{E}(u))^2 &\simeq \frac{1}{N}\sum_{i=1}^N u_i^2 - \left(\frac{1}{N}\sum_{i=1}^N u_i\right)^2,\end{aligned} \tag{1.9}$$

and so on for the higher moments. The MC method is universally applicable and is provable convergent for $N \to \infty$, however the rate of convergence is quite slow $\mathcal{O}(1/\sqrt{N})$ [23, 36, 64]. The advantage of MC, which makes it still attractive, is its independence from the number of uncertain parameters. The main reasons, for the slow convergence of the MC method, are a non well distributed set of points, showing clustered regions or holes in the stochastic space, and multiple occurrences for low probability samples. To overcome this issue many methods, derived from the MC, have been introduced. Among of them, a popular family of method is the so called quasi-MC (QMC) family. The QMC method differs from the original MC only for the points generation. In particular, the random generator of points[1] is substituted by a low-discrepancy sequence [23, 36]. A low discrepancy sequence is a sequence in which the number of points in a set is proportional to its measure. A well known low discrepancy sequence has been introduced by Sobol and it is named Sobol sequence [75]. The convergence rate can be improved to $\mathcal{O}(1/N)$ [23,36], however they cannot totally tackle the prohibitively cost of the MC approach. Another strategy, to force the distribution of random points covering better the stochastic space, is the so-called Latin Hypercube sampling (LHS) strategy. In the case of LHS the 'uniformity' of

[1] To be more precise in MC method pseudorandom generators are employed. A sequence is pseudorandom if it shares the statistical properties of a random sequence, but it is generated in a totally deterministic way.

distribution of the samples is obtained requiring that each equiprobable bin contains a realization. In the figure 1.1, the comparison between MC, QMC and LHS sampling is reported for two uniform random variables and a number of samples equal to 200.



(a) (b) (c)

Figure 1.1: Comparison between Monte Carlo (a), quasi-MC (b) and Latin Hypercube sampling (c) for two uniform random variables for a set of 200 samples.

In figure 1.1, the QMC distribution of the samples appears to be the most uniform, while both MC and LHS exhibit some holes and cluster regions. However, the LHS sampling, unlike MC, for construction assures the most uniform distribution of the samples along the range of each random variable.

1.2.2 Stochastic collocation approach

In the collocation methods, the aim is to satisfy the governing equations in a set of discrete points in the stochastic space. In this respect, the sampling methods are collocation methods. However, instead of employing a random distribution of points, the polynomial approximation theory can be adopted to locate the nodes and obtain higher accuracy. Two different approaches are commonly employed in literature [93]. The first is based on the Lagrange interpolation technique, while the second one relies on a pseudo-spectral expansion. In the following, for simplicity of exposure and clarity of notation, the model function $u = u(\boldsymbol{\xi})$ is assumed to be dependent only on the random vector $\boldsymbol{\xi}$, while the methods presented can be applied to solve, for instance PDEs, depending on physical space and time (recursively employing it at each physical-time location).

A well known result from the interpolation theory (in 1D) states (see for instance [71]) that given $N+1$ separate points ξ_i, then a single polynomial $\Pi_N \in \mathbb{P}_N$ exists satisfying $\Pi_N(\xi_i) = u(\xi_i) = u_i$ for all $i = 0, \dots, N$. The interpolation polynomial $\Pi_N(\xi)$ takes the form, named Lagrange form,

$$\Pi_N(\xi) = \sum_{k=0}^{N} u_k L_k(\xi), \tag{1.10}$$

where the Lagrange polynomials $L_k \in \mathbb{P}_N$ are defined as

$$L_k(\xi) = \prod_{\substack{j=0 \\ j \neq k}}^{N} \frac{\xi - \xi_j}{\xi_k - \xi_j}. \tag{1.11}$$

A relevant property of the Lagrange polynomial is $L_k(\xi_j) = \delta_{jk}$ where δ is the Kronecker symbol.

Once the Lagrange interpolation is at disposal, *i.e.* the N deterministic runs of the code have been performed and each single Lagrange polynomial $L_k(\xi)$ has been computed, the statistics can be evaluated directly on $\Pi_N(\xi)$. For instance, for the expected value

$$\mathbb{E}(u) = \int_\Xi u(\xi)p(\xi)\mathrm{d}\xi \simeq \int_\Xi \Pi_N(\xi)p(\xi)\mathrm{d}\xi = \sum_{k=0}^{N} u_k \int_\Xi L_k(\xi)p(\xi)\mathrm{d}\xi, \tag{1.12}$$

where the integral terms have the role of weights in the discrete sum. Even if the application of the method can seem straightforward several drawbacks limit the direct application of the technique as presented in (1.12). The first problem is relative to the choice of the nodes; in multidimensional problems even some aspects of the theory related to the Lagrange interpolation remain unclear. Moreover, the weights can be obtained only if the analytical expression for the Lagrange polynomial is at disposal; the polynomial can be obtained numerically by inverting a Vandermonde-type matrix, but the procedure results to be cumbersome. Aiming to overcome these issues, choosing the nodes as a set of cubature points has been proposed in literature (see for instance [19, 94]). In this case, the integrals in (1.12) become the weights of the cubature (multiplied by the pdf evaluated in ξ_k) thanks to the properties of the Lagrange polynomials. Furthermore, the direct knowledge of the Lagrange basis is not more required and the method reduces to be a truly sampling one.

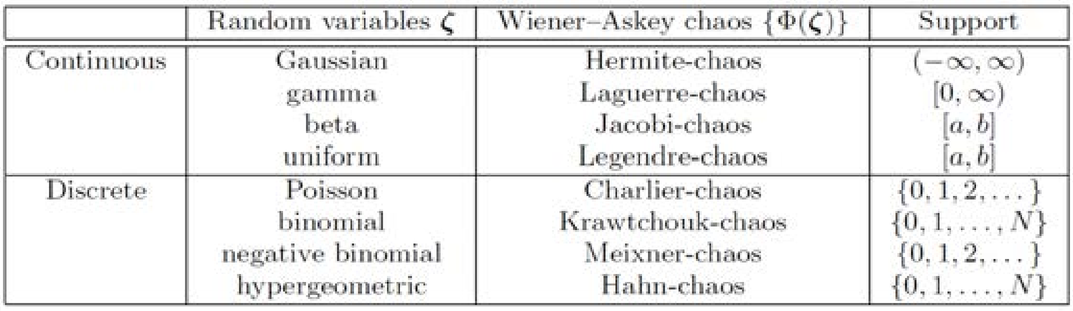
If reduced to a sampling technique, the stochastic collocation approach loses its capability to reproduce the random output over the entire stochastic space. If some characterizations of the output are needed, as for instance its distribution, the knowledge of the model in a discrete set of points could not be enough. A pseudo-spectral approach can be introduced as proposed in [92]. The pseudo-spectral approach relies on the homogeneous chaos theory proposed by Wiener, for Gaussian random variables, in the seminal work [89] and successively generalized, for different measures, by Xiu and Karniadakis in [96]. According to the theorem of Cameron and Martin [24], Hermite-chaos provides a means of expanding any second-order, *i.e.* with finite variance, random process in terms of orthogonal polynomials with a convergence in L_2 sense. However, if the random inputs are not Gaussian, the optimal exponential convergence can be lost, so in [96] the Wiener-Askey polynomial chaos expansion has been proposed to deal with more general random inputs. In the following this technique is referred as generalized Polynomial Chaos (gPC) expansion (*non-intrusive* in this context).

The general expansion for a second order random output $u(\boldsymbol{\xi}) \in L_2(\Xi)$ can be obtained as

$$u(\boldsymbol{\xi}) = \sum_{k=0}^{\infty} \beta_k \Psi_k(\boldsymbol{\xi}), \tag{1.13}$$

where $\{\Psi_k(\boldsymbol{\xi})\}$ represents a Wiener-Askey polynomial chaos basis (see the figure 1.2). Having reduced the attention to independent input random variables, each multi-dimensional polynomial $\Psi(\boldsymbol{\xi})$, of total degree n_0, involves tensorization of (corresponding) one-dimensional ones $\phi(\xi)$ by multi-index $\mathbf{m} = (m_1, \ldots, m_d)$, which determines the degree of the approximation along each separate dimension

$$\Psi(\boldsymbol{\xi}) = \prod_{i=1}^{d} \phi_{m_i}(\xi_i). \tag{1.14}$$



	Random variables $\boldsymbol{\zeta}$	Wiener–Askey chaos $\{\Phi(\boldsymbol{\zeta})\}$	Support
Continuous	Gaussian	Hermite-chaos	$(-\infty, \infty)$
	gamma	Laguerre-chaos	$[0, \infty)$
	beta	Jacobi-chaos	$[a, b]$
	uniform	Legendre-chaos	$[a, b]$
Discrete	Poisson	Charlier-chaos	$\{0, 1, 2, \ldots\}$
	binomial	Krawtchouk-chaos	$\{0, 1, \ldots, N\}$
	negative binomial	Meixner-chaos	$\{0, 1, 2, \ldots\}$
	hypergeometric	Hahn-chaos	$\{0, 1, \ldots, N\}$

Figure 1.2: Wiener-Askey scheme for different distributions of random variables. Table reproduced from [96].

The set of polynomials $\{\Psi_k(\boldsymbol{\xi})\}$ forms an Hilbert basis on $L_2(\boldsymbol{\xi}, p(\boldsymbol{\xi}))$ and the inner product in Ξ can take advantage from the orthogonality of the terms with respect the weighting function which, in this context, is easily identified as the joint pdf $p(\boldsymbol{\xi})$ of the random input variables $\boldsymbol{\xi}$, having

$$\langle \Psi_i, \Psi_j \rangle = \int_{\Xi} \Psi_i(\boldsymbol{\xi}) \Psi_j(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} = \delta_{ij} \left\langle \Psi_i^2 \right\rangle. \tag{1.15}$$

In practical applications, the (1.13) needs to be truncated after a finite term P

$$u(\boldsymbol{\xi}) \simeq \sum_{k=0}^{P} \beta_k \Psi_k(\boldsymbol{\xi}), \tag{1.16}$$

depending from the number of stochastic dimensions d and the maximum polynomial order n_0 to achieve

$$P + 1 = \frac{(n_0 + d)!}{n_0!\, d!}. \tag{1.17}$$

Thus $\mathcal{W}_P$, the subspace of $L_2(\boldsymbol{\xi}, p(\boldsymbol{\xi}))$ of the orthonormal polynomials of total degree n_0, has cardinality $P + 1$. Unlike the Lagrange interpolation approach, the random output $u(\boldsymbol{\xi})$ is fully characterized, over the entire stochastic space Ξ, once the coefficients β_k are computed. In the Lagrange interpolation case, instead, the approximation and the approximation space are implicitly prescribed by the selected points, *i.e.* the Lagrange polynomials depend on the collocation points. In the gPC case the polynomial basis, and the approximation space $\mathcal{W}_P$, are fixed *a priori* irrespective of the collocation points. The random output is then approximated by an orthogonal projection onto $\mathcal{W}_P$ by the (1.16) where

$$\langle u(\boldsymbol{\xi}), \Psi_k \rangle = \beta_k \langle \Psi_k, \Psi_k \rangle \tag{1.18}$$

is the relation useful to compute the coefficients β_k of the expansion. The problem reduces to compute $P+1$ multi-dimensional integrals $\int_\Xi u(\boldsymbol{\xi})\Psi_k(\boldsymbol{\xi})p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}$, while the normalization factor $\langle \Psi_k, \Psi_k \rangle$, depending on the basis employed, can be computed analytically.

Giving the set of coefficients β_k and the polynomial basis, the statistics can be easily computed. For instance, the expected value and the variance can be obtained as

$$\begin{aligned} \mathbb{E}(u) &\simeq \beta_0 \\ \mathrm{Var}(u) &\simeq \sum_{k=1}^{P} \beta_k^2 \left\langle \Psi_k^2(\boldsymbol{\xi}) \right\rangle . \end{aligned} \tag{1.19}$$

More complicated expressions can be employed to compute high-order central moments; the expressions relative to the third and fourth central moments are explicitly obtained in the paper ***P7***. These expressions are reduced, in ***P7***, taking into account the contributions always equal to zero.

Even if in principle a least square approximation can be employed to compute the coefficients β_k of the expansion, this possibility does not exploit the orthogonality of the basis and in general it is not a preferable choice [27] with respect to the spectral projection approach. To avoid the slow convergence related to the sampling methods, the most efficient techniques to compute the coefficients β_k are related to the cubatures, *i.e.* multi-dimensional extension of 1D integration rules.

Remarks on quadrature/cubature *formulæ*

In the context of *non-intrusive* spectral projection techniques, a natural choice is constituted by Gauss type quadratures [27, 51]. The Gauss quadrature *formulæ* are high-order approximations for integrals in which a strictly positive weight function is present. The Gauss quadrature achieves the highest possible degree of exactness integrating polynomial functions of degree less than $2N$ giving N points. Gauss quadrature *formulæ* are closely related to the orthogonal polynomial family for the weight function, which is identified as the joint pdf $p(\boldsymbol{\xi})$. Both Gauss quadrature polynomial basis and the gPC basis are the same and they are identified in function of the polynomial basis, for instance Gauss-Legendre, Gauss-Hermite and so on (see also the figure 1.2). The nodes ξ_i and the weights w_i can be computed by solving an eigenvalue problem; in particular the nodes ξ_i for a $N-$point rule are the zeros of the polynomial $\phi(\xi)$ of order $N+1$ [45] and

$$w_k = \int_\Xi \phi_k(\xi)p(\xi)\mathrm{d}\xi. \tag{1.20}$$

Finally, the approximation by Gauss quadrature, irrespective of the measure, can be written as

$$\int_\Xi u(\xi)p(\xi)\mathrm{d}\xi \simeq \sum_{i=1}^{N} u(\xi_i)w_i. \tag{1.21}$$

The advantage related to the Gauss quadrature is the highest possible accuracy for polynomial integrands, however the maximum polynomial accuracy could not be a fundamental requirement if the random output is not polynomial. This consideration opens the way to other attractive possibilities as the so-called nested quadratures. A nested quadrature family is one in which high-order accuracy *formulæ*, *i.e.* *formulæ* with more nodes, can be obtained adding the missing points to the previous low-accuracy *formulæ*. Examples of this quadrature families are the Cleanshaw-Curtiss and Fejèr rules, or the Newton-Cotes *formulæ* on equally spaced points. This latter family is employed in the work ***P4*** to build nested sequences of meshes for multiresolution purpose (see also §3.2 for the importance of the choice of the quadrature formula for efficient adaptive schemes).

The remaining aspect, related to the quadrature techniques for the spectral projection purpose, is obtaining multi-dimensional cubature *formulæ*. Techniques able to achieve this goal in a data-independent context, *i.e.* without employing the knowledge of the integrand function as for instance in adaptive techniques, are the full tensorization and sparse grids [74]. The full tensorization approach involves the straightforward tensorization of 1D quadrature *formulæ*. This approach constitutes the most robust one to extend 1D quadrature *formulæ*, but it has the disadvantage to be infeasible for high-dimensional problems, exhibiting exponential increase of the nodes with the dimensions, and to be non-optimal in the sense that the same exactness can be obtained with a lower number of nodes. On the contrary, in the sparse grids approach, the multi-dimensional tensorial product is constructed as a linear combination of tensor product interpolants, each of them with relatively few points in its respective points set. This greatly reduces the total number of nodes, while retaining the accuracy and convergence of the 1D interpolants. The sparse grid technique has been explored at the beginning of this research work and some results, of its coupling with the gPC method, can be found in ***RR3***. Moreover, the detailed description of the sparse grid approach is also presented in ***RR3***. The sparse grid concept can be extended to build anisotropic grids; this possibility constitutes still an active research field. Anyway, some results can be found in [51].

The sparse grid technique can be seen historically as the first attempt to tackle the so-called *curse of dimensionality*. This aspect is one of the predominant issues regarding the UQ analysis and affects, in a broad sense, all the UQ propagation techniques. The *curse of dimensionality* also has been one of the fundamental drives for the motivation of the present thesis work; in the context of the semi-*intrusive* (SI) approach, proposed by Abgrall and co-authors [3, 5] (and ***C13***), the multiresolution framework is injected into it to reduce the computational associated with the exponential growth of the degree-of-freedoms of the original SI scheme. One of the aims of the introduction of the multiresolution is representing the discrete data, over the stochastic space, by means of a discrete space with lower cardinality. The original SI scheme will be presented in the §1.3, while in the following section the *intrusive* gPC approach is recalled.

1.2.3 Stochastic Galerkin approach

The gPC expansion can also be employed in a Galerkin fashion, *i.e.* in a weighted residual formalism, to obtain systems of governing equations with respect to the unknown coefficients β_k. Galerkin projection, in the context of UQ propagation, has been introduced by Ghanem and Spanos in [35]. The key difference of this technique, with respect to the others presented above, is that the Galerkin stochastic methods do not rely on the individual, *i.e.* decoupled, realizations of the numerical model. The main idea is to transfer the *randomness* into the polynomial basis obtaining a system of deterministic equations.

Recalling the (abstract) problem statement (1.7) for differential problems, the solution $u(\mathbf{x}, t, \boldsymbol{\xi})$ can be interpreted as a random process and expanded by the Wiener-Askey chaos approach (as already seen for the *non-intrusice* gPC)

$$u(\mathbf{x}, t, \boldsymbol{\xi}) = \sum_{k=0}^{P} u_i(\mathbf{x}, t)\Psi_k(\boldsymbol{\xi}). \tag{1.22}$$

The previous expansion can be injected into the governing equation obtaining

$$\mathcal{L}\left(\mathbf{x}, t, \boldsymbol{\xi}; \sum_{k=0}^{P} u_i(\mathbf{x}, t)\Psi_k(\boldsymbol{\xi})\right) = \mathcal{S}(\mathbf{x}, t; \boldsymbol{\xi}). \tag{1.23}$$

In general, unless $P \to \infty$, the residual associated to the previous discretization is not equal to zero, so, in weak form, it is necessary to require it being orthogonal to any test function $\phi(\boldsymbol{\xi})$ spanning Ξ. Due to the infinite dimensionality of Ξ, a finite approximation must be chosen for the test functions $\Psi(\boldsymbol{\xi})$; in the classical Bubnov-Galerkin context, the test functions coincide with the approximation basis $\{\Psi_k(\boldsymbol{\xi})\}$. The following set of coupled problems can be obtained

$$\left\langle \mathcal{L}\left(\mathbf{x}, t, \boldsymbol{\xi}; \sum_{k=0}^{P} u_i(\mathbf{x}, t)\Psi_k(\boldsymbol{\xi})\right), \Psi_i(\boldsymbol{\xi}) \right\rangle = \langle \mathcal{S}(\mathbf{x}, t; \boldsymbol{\xi}), \Psi_i(\boldsymbol{\xi})\rangle \quad \text{for} \quad i = 0, \ldots, P \tag{1.24}$$

from which, exploiting the orthogonality of the polynomial basis, a set of $P+1$ coupled equations for each mode $u_i(\mathbf{x}, t)$ can be obtained. One of the main drawbacks of the stochastic Galerkin approach is the size of the set of equations to solve, which results to be $P+1$ times larger than the corresponding deterministic problem. The occurrence of large system of equations is the source of an increase in the computational cost and also poses problems in the solution strategies to adopt. To tackle the increasing of the computational cost strategies, decoupling each spectral mode $u_i(\mathbf{x}, t)$ results to be very effective [51]. However in several cases, the approach is not feasible due to theoretical difficulties. Moreover, a larger system (even linear) cannot be solved with direct resolution method requiring iterative techniques. Other issues are related to the projection of non-linearities on the expansion basis [51]. The stochastic Galerkin method, as presented in this brief sketch, represents nowadays *de facto* the standard *intrusive* technique [29,62,95]. However, as pointed out in [51], as for Fourier expansion where Gibbs *phenomena* can occur, also the gPC expansion can exhibit a lack of suitability due to its high regularity if the functional to approximate admits a bifurcation or a discontinuity. The consequences can be both a slow convergence in the infinite case or the presence of parasitic oscillations in the finite case. Moreover, for problems involving the time-integration of dynamical systems with random frequencies, the broadening of the solution spectrum can cause the loss of the spectral convergence after a finite time. This problem has been discussed, for instance in [33,88].

Both bifurcations, discontinuities or in general steep variations with random inputs are common *phenomena* in fluid dynamics applications among other fields. Forming shocks, energy cascades or also combustion problems, for instance, can lead to these behaviors and constitute challenging applications for UQ propagation. The presence of this kind of issues for non-smooth problems has been one of the key aspects which motivated the introduction of the semi-*intrusive* scheme in the seminal work of Abgrall [3]. The semi-*intrusive* method will be introduced in §1.3, while in the next section a brief review of the techniques adopted to improve the stochastic Galerkin method in presence of non-smooth solution will be presented.

1.2.4 Some cures proposed for non-smooth problems

The introduction of suitable techniques, which can be able to improve the efficiency and accuracy of the UQ propagation schemes in presence of non-smooth functions in the stochastic space, is a very recent and actual field of research. The book [51] constitutes a systematic introduction to some possible cures for non-smooth solutions in stochastic space for stochastic Galerkin methods. Basically, in this text both Haar wavelet basis and multiwavelets have been introduced and analyzed. The first attempt to obtain a multiscale scheme, for UQ propagation, has been [50]. A wavelet-based expansion has been developed introducing piecewise functions, the Haar wavelets, in stochastic Galerkin methods. This method is robust, but it exhibits lower rate of convergence, with respect to a global spectral approximation, for smooth problems. The same authors of [50] refined the approach proposing to generalize the strategy to arbitrary polynomial order expansions [49]. This improved strategy results in a *h*-*p* refinement techniques where, the level of resolution gives the *h*-refinement, while the polynomial order guarantees the *p*-refinement. A related concept of *h*-*p* refinement has been proposed by Wan and Karniadakis to obtain a Multi-Element generalized PC (ME-gPC) scheme [88]. In this context, the *h*-refinement is due to the size of the random elements, while the *p*-refinement is performed by a local gPC approximation of arbitrary order. An *a posteriori* heuristic criterion is employed in ME-gPC to guide the partitioning. An original *hp*-refinement has also been introduced by Lucor and co-authors (see for instance [61]). Closely related to the concept of *h*-refinement by local basis refinement, methods based on of hierarchical sparse grids are proposed in [10, 63]. A different approach has been proposed in [60, 67] to directly tackle the presence of Gibbs oscillations in the neighboring of a discontinuity, aiming to avoid unphysical values. The novel stochastic Galerkin method has been introduced relying on entropy variables, which are related related to the main variables through the entropy of the system. This approach shows an interesting alternative, to bound unphysical oscillations, unlike adapting (in both senses) in the stochastic space.

In [83], to tackle the presence of discontinuities in both physical and stochastic space, a finite-volume approach in the physical space has been coupled with a stochastic Galerkin method in the stochastic space using a piecewise polynomial basis and an original Roe-type solver [85]. Despite improvements, with respect to the standard Galerkin approach, the method is still very expensive from a computational cost point-of-view. However, the same authors, in successive works [84, 86], introduced the adaptivity anisotropic feature formulated in the multiresolution context. This scheme represents the state-of-the-art for the stochastic Galerkin method. Anyway, some issues are still present; the Roe-type stochastic solver requires the explicit knowledge (*a priori*) of the eigenstructure of the stochastic hyperbolic system, which derives from the Galerkin projection. If this is not an issue for equations like Burgers and Euler, it can be unfeasible for general conservation laws. Another limitation is related to the spatial discretization of the method. Only the first order can be obtained incurring in both theoretical and practical difficulties generalizing the approach. This limitation, from a practical point-of-view, makes the high-order convergence in the stochastic space of reduced utility because the overall error appears rapidly dominated by the physical spatial accuracy. More recently, adopting the multiwavelet basis, Petterson and co-authors in [66] obtained an intrusive Galerkin method for two-phase flows. Unlike the work [86], in [66] the problem is solved with a hybrid method coupling the continuous phase region with the discontinuous phase region through a numerical interface. The non-smooth region is solved with the HLL-flux and MUSCL-reconstruction in space; finite-difference operators in summation-by-parts form are used for the high-order spatial discretization.

In the *non-intrusive* context, the issue related to discontinuous surfaces in the stochastic space is also present. Many solutions have been proposed in several different frameworks. For instance, Chantrasmi and Iaccarino in [25] proposed a multidimensional approach based on Pade-Legengre approximation for CFD applications in presence of shock waves. A new iterative formulation, improving the convergence of standard stochastic collocation approach, has been presented by Poëtte and Lucor in [68]. The authors demonstrated the capability of the method to achieve a better convergence with no additional cost, *i.e.* the additional operations with respect the standard spectral method are all preformed in the post-processing phase. The method has been successfully applied to Euler system of equations in [59, 68]. More recently, in the context of the simplex approach [90], Witteveen and Iaccarino introduced the concept of sub-cell resolution for problems in which the discontinuities in the random space are directly related to their physical counterparts [91]. The presence of a sparsity character of the solution, *i.e.* only few coefficients in the PC basis are really non-null, has been employed by Doostan and Owhadi in [31] to obtain a non-adapted sampling method. If the assumptions of mild dimensionality and sparsity are still valid, and in the presence of sharp gradients and/or discontinuities, an adaptive important sampling strategy can be introduced to increase the efficiency of this techniques [72]. A direct comparison between the iterative spectral approach, the sub-cell resolution technique within the simplex method and the adaptive important sampling for compressing sensing has been proposed for some test cases in [59].

1.2.5 Remarks on other (less general) intrusive approaches

Other less general techniques exist for *intrusive* UQ propagation. A popular non-sampling method is the so-called perturbation method in which the random field is approximated by a Taylor expansion. However, the related system of equations becomes cumbersome beyond second-order and also the approach is limited to narrow magnitude of uncertainties, both for input and output variables. Some engineering applications of the method can be found in [47]. Similarly to perturbation problems are the operator based methods. This family of methods is based on the expansion of the stochastic operators, for instance by Neumann expansion [97] or weighted integral methods [30]. They share with the perturbation method the limited applicability in term of magnitude of uncertainties. A direct averaging of the governing equations can lead to a system of equations in which the unknowns are the moments of the solution. However this approach poses issues related to the closure of the problem, *i.e.* information about higher-moments are very often required. Despite its limited applicability, in [98] some application examples are presented.

1.3 Semi-intrusive scheme (a hierarchical interpretation)

In this section, the semi-intrusive (SI) method proposed by Abgrall and co-workers in [3, 4] (and ***C13***) is presented. The SI scheme has been further investigated during this thesis research work since its original development in [3]. The SI scheme constitutes the building block of the overall scheme developed during this thesis, the so-called adaptive-SI (aSI) scheme, which will be presented in the Chapter 3. The aSI scheme can be seen essentially as a semi-intrusive scheme based on a multiresolution approximation space for the basis of the function in the random space. However, the original formulation of the scheme presented in [3] does not make possible a direct connection within the multiresolution framework developed during this thesis work. Despite the idea of a general finite volume (FV) reconstruction in the stochastic space is retained, in the aSI scheme the original SI method needs to be, at least, reformulated in a hierarchical way. The hierarchical re-arrangement of the scheme translates in a change of focus from the physical to the stochastic space. The numerical algorithm is not performed for all the physical locations with a fixed stochastic coordinate, but instead for (some) stochastic locations for fixed physical coordinates. In this respect the aSI scheme is certainly more intrusive than its SI counterpart and it is closely related to the deterministic formulation because, in general, it is necessary to enforce a decoupling between the values along the physical coordinates. All these aspects will be discussed in more detail in the Chapter 3, while in this section the hierarchical reinterpretation of the SI scheme is presented. In this thesis, the focus has been devoted to time-dependent problems with a particular emphasis on hyperbolic problems. It is well known that this class of problems is of fundamental importance, among other fields, for fluid dynamics applications. Moreover, this class of problems, both for linear and non-linear equations, can deal with discontinuous solutions. Several numerical methods are nowadays at disposal for hyperbolic systems of conservation laws [14, 32, 43, 53, 55, 80, 82]. In this thesis, aiming to develop a general stochastic scheme, without loss of generality, the focus has been restricted to FV scheme. The reasons are the well-established theory and numerical knowledge of this class of methods, their widespread use in the CFD context (nowadays they are the standard CFD technique). Moreover, the Monotone Upstream-centered Scheme for Conservation Laws (MUSCL) [82] is a FV approach allowing an easy extension of standard Godunov methods to second the order of accuracy. A very robust MUSCL method is the so-called MUSCL-Hancock method (MHM) [52]. The following introduction of the (hierarchical recasted) SI scheme will be presented for the MHM employed in the thesis papers ***P3*** and ***P4***.

The deterministic MHM is first introduced in §1.3.1 and, afterwards it is formulated and extended in the stochastic space with the SI approach in §1.3.2. Anyway, a standard MUSCL method [87] is instead employed (as deterministic basis for the aSI) in the paper ***P5*** in which the Discrete Equation Method (DEM) [7,8] formulation is adopted for multiphase flow simulations.

1.3.1 MUSCL-Hancock deterministic numerical formulation

The MUSCL-Hancock method consists, as the classical MUSCL approach, of two fundamental steps: a predictor and a corrector step. However, the main difference among the two schemes is that the MHM does not require the solution of a Riemann problem at each interface during the predictor stage. On the contrary, in the predictor stage the only fluxes required at the interfaces can be computed analytically evaluating the flux function relative to an opportune extrapolated value of the conservative variable (or vector of variables). The following derivation is made for a scalar 1D conservation law because the extension to system of conservation laws is straightforward, while the multidimensional extension can be carried out with a dimensional splitting approach (on Cartesian grids). The interested reader can find many details in a reference book as [82].

A general 1D scalar conservation law reads

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f(u(x,t))}{\partial x} = 0, \tag{1.25}$$

where $x \in \Omega \subset \mathbb{R}$ is the physical space and $t \in T \subset \mathbb{R}^+$ is the time space. The physical space is divided in a set of non-overlapping cells $\mathcal{C}_i$ with $\Omega = \bigcup_i \mathcal{C}_i$. The classical first order Godunov scheme, applied to (1.25), is obtained introducing the so-called cell-average $\bar{u}_i$ on each cell $\mathcal{C}_i$:

$$\bar{u}_i(t) = \frac{1}{|\mathcal{C}_i|} \int_{\mathcal{C}_i} u(x,t)\mathrm{d}x, \tag{1.26}$$

where $|\mathcal{C}_i|$ indicates the volume of the cell. The Godunov method is only first order accurate in space due to the constant approximation of the solution $u(x,t)$ over each spatial cell. Following the idea of Van Leer, high-order schemes can be constructed employing non-constant data. If a piecewise linear approximation is used for the solution $u(x,t)$, on the cell $|\mathcal{C}_i|$ it is possible to write:

$$u(x,t_n) = \bar{u}_i^n + \sigma_i^n(x - x_i) \quad \text{with} \quad x_{i_L} \leq x \leq x_{i_R}, \tag{1.27}$$

with σ_i^n the so-called slope in the cell $\mathcal{C}_i = [x_{i_L}, x_{i_R}]$. Of course, the choice of $\sigma_i^n = 0$ leads to the Godunov scheme; the approach is always conservative because the value of the slope σ_i^n does not affect the cell average value $\bar{u}_i^n$. To obtain a second-order accurate method, a nonzero slope σ_i^n must be computed in a way in which it approximates $\partial u(x,t)/\partial x$ over the ith cell. A slope limiter should be introduced near the discontinuities to avoid oscillations. In this work, both the Roe's superbee limiter and the van Leer limiter are employed. The superbee limiter in its limited slope form is

$$\begin{cases} \sigma_i^n = \mathrm{maxmod}\left(\sigma_{(1)}^n, \sigma_{(2)}^n\right) \\ \sigma_{(1)}^n = \mathrm{minmod}\left(\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathcal{C}_i|}\right), 2\left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathcal{C}_i|}\right)\right) \\ \sigma_{(2)}^n = \mathrm{minmod}\left(2\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathcal{C}_i|}\right), \left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathcal{C}_i|}\right)\right), \end{cases} \tag{1.28}$$

where the minmod and maxmod functions are defined as follows

$$\mathrm{minmod}(a,b) = \begin{cases} a & \text{if} \quad |a| < |b| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |a| > |b| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab <= 0 \end{cases}$$

$$\text{maxmod}(a,b) = \begin{cases} a & \text{if} \quad |a| > |b| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |a| < |b| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab <= 0. \end{cases}$$

The van Leer limiter, in the form of slope limiter, is defined as (see Toro [82] for further details)

$$\sigma_i^n = \begin{cases} \min\left(\frac{2R}{1+R}, \frac{2}{1+R}\right) \frac{\bar{u}_{i+1}^n - \bar{u}_{i-1}^n}{2|\mathcal{C}_i|} & \text{if} \quad R > 0 \\ 0 & \text{if} \quad R \leq 0, \end{cases} \tag{1.29}$$

where R is the ratio between successive slopes $R = (\bar{u}_i^n - \bar{u}_{i-1}^n)/(\bar{u}_{i+1}^n - \bar{u}_i^n)$. Anyway, other approaches are also possible as the so-called flux-limiter formulation (see for instance [53, 55]).

At the interfaces, the conservative variable can be reconstructed both employing the right or the left cell. To avoid the solution of a Riemann problem at the interface, the MHM is based on a prediction step totally interior to each cell; after reconstructing the slope σ_i^n, the two values at the interfaces of the cell are obtained. The values extrapolated at the interfaces are employed to evaluate the net flux into the ith cell evolving their value for half time step ($\Delta t/2$). Finally, the updated value for the ith cell average $\bar{u}_i^{n+1}$ is obtained solving the Riemann problems at the interfaces employing the evolved values. In conclusion, the fully discrete second order MHM, for computing the cell averaged solution $\bar{u}_i^{n+1}$, consists of the following three steps:

- Step 1 - For each cell $\mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}\}$, the solution at the interface is computed according to

$$\begin{cases} u_{\ell_L}^n = \bar{u}_\ell^n - \sigma_\ell^n \frac{|\mathcal{C}_\ell|}{2} \\ u_{\ell_R}^n = \bar{u}_\ell^n + \sigma_\ell^n \frac{|\mathcal{C}_\ell|}{2} \end{cases} \tag{1.30}$$

- Step 2 - On each cell $\mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}\}$, the solution is evolved of a half time step employing the flux function $f = f(u)$:

$$\begin{cases} u_{\ell_R}^{\Uparrow} = u_{\ell_R}^n + \frac{1}{2}\frac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u_{\ell_L}^n) - f(u_{\ell_R}^n)\right) \\ u_{\ell_L}^{\Uparrow} = u_{\ell_L}^n + \frac{1}{2}\frac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u_{\ell_L}^n) - f(u_{\ell_R}^n)\right) \end{cases} \tag{1.31}$$

- Step 3 - The cell-averaged value on the cell $\mathcal{C}_i$ evolves following

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{|\mathcal{C}_i|}\left(\mathcal{F}^{\text{RM}}\left(u_{i-1_R}^{\Uparrow}, u_{i_L}^{\Uparrow}\right) - \mathcal{F}^{\text{RM}}\left(u_{i_R}^{\Uparrow}, u_{i+1_L}^{\Uparrow}\right)\right). \tag{1.32}$$

The symbol $\mathcal{F}^{\text{RM}}$ is employed to indicate the flux evaluated at the interface, after the solution of the Riemann problem defined by two constant states based on the evolved extrapolated values. Both exact or approximated Riemann solvers can be employed. For instance, in this thesis exact Riemann solvers are employed in the papers ***P3*** and ***P4*** for the advection and Burgers equations, while a Roe-Pike method with the Harten-Hyman entropy fix following [82] is employed in the paper ***P4*** and a DEM [7, 8] solver is employed in ***P5*** for multiphase flows.

The time advancing formula is limited to a stencil of only three cells $\mathcal{C}_{i-1}$, $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$, but the computation of the slopes for the cells $\mathcal{C}_{i-1}$ and $\mathcal{C}_{i+1}$ requires (see (1.28) and (1.29)) also to know the solutions on the two surrounding cells $\mathcal{C}_{i-2}$ and $\mathcal{C}_{i+2}$. The average solution $\bar{u}_i^{n+1}$, on each cell $\mathcal{C}_i$ at time $t_{n+1} = t_n + \Delta t$, can be computed knowing only the solution on the augmented stencil $\left\{\bar{u}_{i-2}^n, \bar{u}_{i-1}^n, \bar{u}_i^n, \bar{u}_{i+1}^n, \bar{u}_{i+2}^n\right\}$. In the following, the notation $\bar{u}_i^{n+1} = \mathrm{MHM}\left(\bar{u}_{i-2}^n, \bar{u}_{i-1}^n, \bar{u}_i^n, \bar{u}_{i+1}^n, \bar{u}_{i+2}^n, \Delta t\right)$ is used to identify the ensemble of the operations described above. The aim is to evaluate the updated value in time of a certain cell $\bar{u}_i^{n+1}$, knowing the solution at the previous time step (only in the augmented stencil).

1.3.2 Semi-intrusive formulation for the MHM

The fundamental ingredient of the (hierarchical) SI scheme is, at deterministic level, to make evident the connection between a stencil of cell average values at time n and the updated value at time $n+1$ depending on it. This has been accomplished in the previous section. The remaining step is to equip the problem with further dimensions, the stochastic coordinates, in which a finite volume like representation can be obtained. It should be clear that the equation (1.25) must be intended, in this context, dependent on the vector of random variables $\boldsymbol{\xi}$

$$\frac{\partial u(x,t,\boldsymbol{\xi})}{\partial t} + \frac{\partial f(u(x,t,\boldsymbol{\xi}))}{\partial x} = 0, \tag{1.33}$$

as already introduced in (1.7). The finite volume approximation of the stochastic space Ξ requiring to introduce a tessellation, constituted by N_ξ cells Ξ_j with $j = 1, \dots, N_\xi$, with the following properties

$$\begin{aligned} &\bigcup_{j=1}^{N_\xi} \Xi_j = \Xi \\ &\mu(\Xi_i \cap \Xi_j) = 0 \quad \text{for} \quad i \neq j, \end{aligned} \tag{1.34}$$

where the first condition expresses the fulfillment of the stochastic space Ξ, while the second one requires the mutual independence between cells. From a topological point-of-view, the latter condition means that the cells must be disjoint. The measure $\mu(\Xi_j) > 0$ represents the probability measure as already introduced at the beginning of the chapter. Mimicking the deterministic FV approach, a generalized cell average operator is introduced $\mathbb{E}\left(\bullet \,|\, \Xi_j\right)$; the linear operator $\mathbb{E}\left(\bullet \,|\, \Xi_j\right)$ represents the conditional expected value (of a random function) with respect to the cell Ξ_j:

$$\mathbb{E}(\bullet \,|\, \Xi_j) = \frac{1}{\mu(\Xi_j)} \int_{\Xi_j} \bullet(x,\xi,t)\, p(\xi,t)\, \mathrm{d}\xi. \tag{1.35}$$

If the conditional expected value operator is applied to the step three of the MHM scheme (1.32), the following scheme is obtained:

$$\begin{aligned} \mathbb{E}\left(u_i^{n+1} \,|\, \Xi_j\right) &= \mathbb{E}\left(u_i^n \,|\, \Xi_j\right) \\ &- \frac{\Delta t}{|\mathcal{C}_i|} \left(\mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i-1_R}^{\Uparrow}, u_{i_L}^{\Uparrow}\right) \,|\, \Xi_j\right) - \mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i_R}^{\Uparrow}, u_{i+1_L}^{\Uparrow}\right) \,|\, \Xi_j\right)\right). \end{aligned} \tag{1.36}$$

The evaluation of the updated conditional expected value $\mathbb{E}\left(u_i^{n+1}\,|\,\Xi_j\right)$ is possible knowing its value at the previous time step $\mathbb{E}\left(u_i^n\,|\,\Xi_j\right)$ and the conditional expected value of the fluxes. The value $\mathbb{E}\left(u_i^n\,|\,\Xi_j\right)$ is always at disposal both for initial value problems, in which the initial condition is analytically known, or in steady problems, in which an iterative procedure can be initialized from an analytically known solution. The computation of the conditional expected values of the fluxes requires more attention. Basically, a cubature rule needs to be applied to evaluate the integrals; as it has been already seen, this reduces to the computation of the flux functions $\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right)$ and $\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right)$ in a set of quadrature points. The key element, to evaluate the set of flux values in the quadrature points, is the introduction of a reconstruction of the solution $u(x, t_n, \boldsymbol{\xi})$ knowing only its conditional expected values $\mathbb{E}\left(u_i^n\,|\,\Xi_j\right)$. In particular, a polynomial $\mathcal{P}_j(\boldsymbol{\xi})$ can be introduced on the cell Ξ_j to reconstruct, in a conservative fashion, the solution over the cell. When $\mathcal{P}_j(\boldsymbol{\xi})$ is at disposal its value can be evaluated in the quadrature points and their values injected into the Step 1 (1.30) and Step 2 (1.31) of the MHM. At this point, a multiresolution representation over Ξ of the unknown function $u(x, t, \boldsymbol{\xi})$ can be employed and this aspect has been one of the key features introduced during this thesis work. For this reason, at this level, it is not useful to introduce further details on the quadrature rule to employ or on the conservative interpolation procedure for the polynomial reconstruction (they can be of any kind). However, these aspects are closely related with the multiresolution framework and they will be introduced in the Chapter 3.

1.4 Closing remarks

In this section, a brief review of the state-of-the-art has been presented. The main limitations of the actual strategies for UQ simulations are related to the discontinuous responses and the possibility to handle whatever form of pdf, as for instance when experimental data are at disposal. Despite the great flexibility of the *non-intrusive* techniques for UQ propagation, allowing a virtually infinite range of application, the potentially efficiency gain of the *intrusive* methods demands for an improvement of the actual algorithms. The present thesis work moves in this direction. In particular, actually, the established class of *intrusive* stochastic Galerkin approach remains limited to specific applications for the theoretical and algorithmic effort required for each specific application. Even if the recent introduction of the MR setting partially solved the issues related to the presence of discontinuous responses, the main limitation related to the growing number of equations is still present. In this sense, the SI scheme has been developed having in mind the need for an easy adaptation to general applications without requiring to solve a more complex system of equations. Moreover, the possibility to handle very general pdf, as discontinuous time varying pdf, has not been faced at all in spectral methods. All these aspects motivated the introduction of the SI. However, as all the other *intrusive* techniques, its numerical cost reflects the increase of the degree-of-freedoms with the growth of the stochastic dimensions. In this respect, one of the main goals of this thesis is to develop an adaptive version, of the original SI scheme, ables to increase the computational efficiency. Having in mind the fluid dynamics applications, the multiresolution framework appears a natural choice due to its capability to represent multi-scale *phenomena*. The multiresolution framework is introduced in the following chapter.

CHAPTER 2

From the Harten multiresolution framework to the Truncate and Encode approach

Historically, one of the most basic and powerful tool, for a great variety of applications in many different scientific fields, is the Fourier analysis. The drawback, related to the non-compact support of the basis functions (sines and cosines in the Fourier analysis, but also Wiener-Askey polynomials in gPC theory), is the oscillatory behavior in the truncated series for functions with isolated singularities. The wavelet basis can be employed to analyze square-integrable functions as in the Fourier analysis [28], but they perform better. The reason is due to the localized support of the wavelet basis. The wavelet decomposition features a multiresolution/multiscale property that can be exploited in many ways in the numerical analysis. The building blocks of the theory remains a scaling function, also called mother wavelet, and a dilation relation. The orthonormal wavelet basis is composed by dilating and translating the mother function. In some sense, the construction of such orthonormal basis is equivalent to searching for the solution of the dilating equation [78]. However, conceptual difficulties arise in extending wavelets to bounded domains and general geometries [26], and in obtaining adaptive (data-dependent) representations[1]. A generalized multiresolution (MR) framework has been obtained by the contribution of many researchers from the seminal idea of Ami Harten in the '90. The Harten's framework combines ideas from multigrid methods, numerical schemes for conservation laws, hierarchical bases in finite element and the theory of wavelets resulting in a more general framework, which is able to represent discrete multiscale data. Within this MR framework, the dilation equation loses its central role and the connection between different (adjacent) resolution levels is associated to the decimation and the prediction operators. Moreover, these two operators rely on a discretization and a reconstruction operator.

In this chapter, the multiresolution framework of Harten is briefly introduced following [6, 38–40] focusing on UQ analysis in the section §2.1. Depending on the discretization procedure adopted, two different frameworks can be proposed: the point-value and the cell-average frameworks. The extension to the point-value setting is presented in §2.2, while the cell-average framework is described in §2.3. The MR framework can be considered a rearrangement of (discrete) information [16] corresponding to different resolution levels. If a procedure for the identification of the non-significant data is introduced, the original representation can be reduced in its dimensionality. This kind of procedure, if related to numerical algorithms, can allow

[1] In more recent years, the so-called lifting scheme emerged as a tool for the construction of more general wavelets, which are not related to translation and dilation of a single function (see [79]).

an increasing of efficiency in terms of computational cost (only the significant data are processed) and memory requirement (only significant data are stored). The high-quantity of degrees-of-freedoms in the combined physical/stochastic space for time-dependent problems makes this approach very attractive for designing intrusive UQ schemes.

In the following, some basic ingredients are presented, while further details are reported in the papers ***P1***, ***P2***, ***P3*** for the point-value setting and ***P4*** and ***P5*** for the cell-average framework. However, some extensions have not yet been published and are reported here for the first time[2]: the introduction of the Weighted Essentially Non-Oscillatory (WENO) reconstruction in the point-value setting, presented in §2.2.1, and the conservative reconstruction in two and three stochastic dimensions, for the cell-average framework, reported in section §2.3.1.

2.1 Multiresolution framework for stochastic problems

The building blocks of the framework are the operators of discretization $\mathcal{D}_k$, which allows the transfer of information from the continuous to the discrete space, and the reconstruction operator $\mathcal{R}_k$ which performs the inverse operation. Using both operators of discretization $\mathcal{D}_k$ and reconstruction $\mathcal{R}_k$, the discrete operators of decimation D_k^{k-1} and prediction P_{k-1}^{k}, which operate between consecutive levels of resolutions k (higher resolution) and $k-1$ (lower resolution), can be defined.

Let us consider a function $f \in \mathcal{F}$, where $\mathcal{F}$ is a proper space of functions. A set of discrete operators of discretization $\{\mathcal{D}_k\}_{k=0}^{L}$, each of them defined on a vectorial space of finite dimension J_k, is defined as

$$\mathcal{D}_k : \quad \mathcal{F} \to V_k \quad \text{with} \quad \dim(V_{k+1}) > \dim(V_k) = J_k. \tag{2.1}$$

The sequence $\{\mathcal{D}_k\}_{k=0}^{L}$ is nested according to the following properties:

- $\mathcal{D}_k$ is onto
- the null space of each level includes the null space associated to the previous resolution level $\mathcal{N}(\mathcal{D}_k) \subset \mathcal{N}(\mathcal{D}_{k+1})$.

These properties reflect in the following relation between discretization operators

$$\mathcal{D}_{k+1}(f) = 0 \Rightarrow \mathcal{D}_k(f) = 0 \quad \forall f \in \mathcal{F}. \tag{2.2}$$

Thanks to the onto property of each operator $\mathcal{D}_k$, the reconstruction operator $\mathcal{R}_k$ can be defined as follows

$$\mathcal{R}_k : \quad V_k \to \mathcal{F}. \tag{2.3}$$

The reconstruction operator is not required to be linear, in the sense of having a data-dependent basis; this point makes the Harten's multiresolution more general with respect to the wavelet framework [34].

Both operators $\mathcal{D}_k$ and $\mathcal{R}_k$ need to satisfy the following consistency relationship

$$(\mathcal{D}_k\mathcal{R}_k)(v) = v \quad \forall v \in V_k, \tag{2.4}$$

[2] WENO reconstruction procedure has been presented at the HONOM 2013 workshop on high-order methods (see also ***C5***).

thus implying $\mathcal{D}_k\mathcal{R}_k = \mathrm{I}_k$, where I_k is the identity operator on V_k.

In the case of nested sequences, whose elements are defined in (2.1), the decimation operator D_k^{k-1} can be defined as a linear mapping between V_k onto V_{k-1}:

$$\mathrm{D}_k^{k-1}: \quad V_k \to V_{k-1}, \tag{2.5}$$

where

$$\mathrm{D}_k^{k-1} v^k = \mathcal{D}_{k-1} f \in V_{k-1} \quad \forall v^k = \mathcal{D}_k f \in V_k. \tag{2.6}$$

The decimation operator is used to generate recursively the set of discrete data from the highest resolution level ($k = L$) to the lowest one ($k = 0$) $\{v^k\}_{k=0}^{L-1}$

$$v^{k-1} = \mathrm{D}_k^{k-1} v^k \quad \forall k = L, L-1, \dots, 1. \tag{2.7}$$

Inversely, the prediction P_{k-1}^k allows to approximate the set of data v^k from v^{k-1}

$$v^k = \mathcal{D}_k f \approx \mathcal{D}_k(\mathcal{R}_{k-1} v^{k-1}). \tag{2.8}$$

This leads to the definition of the prediction operator P_{k-1}^k between discrete data on successive resolution levels as

$$\mathrm{P}_{k-1}^k \overset{\text{def}}{=} \mathcal{D}_k\mathcal{R}_{k-1}: \quad V^{k-1} \to V^k. \tag{2.9}$$

The role of the operators $\mathcal{D}_k$, $\mathcal{R}_k$, D_k^{k-1} and P_{k-1}^k in transferring information between the discrete levels and the continuous space is schematically represented in the figure 2.1.

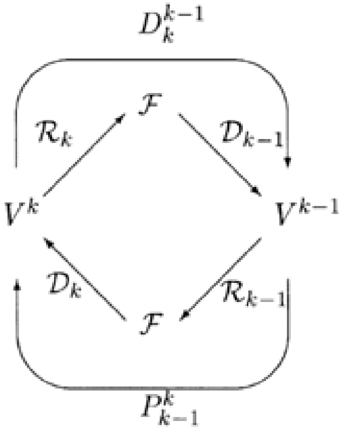


Figure 2.1: Sketch of the role played by the operators $\mathcal{D}_k$, $\mathcal{R}_k$, D_k^{k-1} and P_{k-1}^k between the discrete and continuous spaces. Figure reproduced from [12].

A consistency properties can be found between the discrete operators, *i.e.* $\mathrm{D}_k^{k-1}\mathrm{P}_{k-1}^k = \mathrm{I}_k$, following from

$$v^{k-1} = \mathrm{D}_k^{k-1} v^k = \mathrm{D}_k^{k-1}\mathcal{D}_k f = \mathrm{D}_k^{k-1}\mathcal{D}_k\mathcal{R}_{k-1} v^{k-1} = \mathrm{D}_k^{k-1}\mathrm{P}_{k-1}^k v^{k-1}. \tag{2.10}$$

Now, an error prediction in the MR framework, e^k, can be defined as

$$e^k \overset{\text{def}}{=} v^k - \mathrm{P}_{k-1}^k v^{k-1} = (\mathrm{I}_k - \mathrm{P}_{k-1}^k \mathrm{D}_k^{k-1}) v^k. \tag{2.11}$$

The prediction error satisfies (from the consistency property (2.10))

$$\mathrm{D}_k^{k-1} e^k = \mathrm{D}_k^{k-1}(v^k - \mathrm{P}_{k-1}^k v^{k-1}) = v^{k-1} - v^{k-1} = 0, \tag{2.12}$$

then it is in the null space of the decimation operator $e^k \in \mathcal{N}(\mathrm{D}_k^{k-1})$. Remembering the definition (2.5), and applying the rank theorem, it is possible to find that

$$\begin{aligned} \dim(V_k) &= \dim(\mathcal{N}(\mathrm{D}_k^{k-1})) + \dim(V_{k-1}) \\ &\to \dim(\mathcal{N}(\mathrm{D}_k^{k-1})) = \dim(V_k) - \dim(V_{k-1}) = J_k - J_{k-1}. \end{aligned} \tag{2.13}$$

The linear $J_k - J_{k-1}$ independent coordinates of e^k are called wavelets or details d^k. Two operators can be defined to link the prediction error to the details, E^k and G^k, as follows

$$e^k \overset{\text{def}}{=} E^k d^k, \quad d^k \overset{\text{def}}{=} G^k e^k \quad \text{with} \quad E^k G^k : V^k \to \mathcal{N}(\mathrm{D}_k^{k-1}). \tag{2.14}$$

A multiresolution representation of data can be defined using the operators defined above to perform the *encoding* and the *decoding* procedures. The *encoding* moves from the highest resolution level to the lowest one applying recursively (for all $k = L, \ldots, 1$) the decimation operator and computing the details

$$\begin{cases} v^{k-1} = \mathrm{D}_k^{k-1} v^k \\ \quad d^k = G_k(\mathrm{I}_k - \mathrm{P}_{k-1}^k \mathrm{D}_k^{k-1}) v^k. \end{cases} \tag{2.15}$$

The *encoding* procedure results in a hierarchical representation, *i.e.* decomposition, of a function on nested resolution levels. The *decoding* procedure is the dual procedure with respect to the *encoding*: it recursively moves from the lowest resolution level v^0 together with the prediction error e^k, as function of the details d^k for all the levels $k = 1, \ldots, L$

$$v^k = \mathrm{P}_{k-1}^k v^{k-1} + E_k d^k. \tag{2.16}$$

The *decoding* procedure allows to obtain the finest resolution level knowing only the discrete data on the coarsest level and the details d^k. The one-to-one correspondence between the highest resolution level v^L and the sequence of the details d^k in addition to the lowest resolution level v^0 turns directly in the possibility to define a multiresolution representation v_{MR}:

$$v_{\mathrm{MR}} \overset{\text{def}}{=} \{v^0, d^1, \ldots, d^L\}. \tag{2.17}$$

The v_{MR} represents an alternative and equivalent form of the original function, but the details d^k also contain information relative to the presence of different scales. Introducing an operator of data truncation, the non-significant information, under a certain tolerance, can be eliminated and the dimensionality, *i.e.* number of non null details of the multiresolution representation v_{MR}, can be greatly reduced. This compact representation can be either employed as a compact representation in the signal/image representation schemes [17] or injected into a numerical scheme involving multi-scale *phenomena* as in the seminal work of Harten [39,40].

2.1.1 Note on the Truncation and stability requirements

The most important applications of the multiresolution framework are connected to its compression capabilities. The main idea is to eliminate all the redundant information keeping only the significant ones. It is remarkable that the compression process, despite the algorithm used to obtain it, always generates an approximation of the original data. With particular reference to what has been presented in the previous sections, if a function f, discretized on the finest resolution level $v^L = \mathcal{D}_L f$, is represented in its multiresolution form $v^L = v_{MR}$ the aim of the truncation procedure is to produce an approximated multiresolution representation $\hat{v}^L = \hat{v}_{MR}$. The truncation procedure should yield, after the application of the *encoding* algorithm on $\hat{v}_{MR}$, a data set $\hat{v}^L$ which is close, in some norms and under certain accuracy requirements, to the original data v^L. Different truncation procedures can be designed to achieve the correct reproduction of the data obtaining the desired level of compression. In this work, despite the possibility to use more exotic (with respect to the pure classical numerical scheme) procedures as the quantization (see for instance [41]), the truncation based on the elimination of the wavelets d^k under a prescribed tolerance is addressed. The problem statement is the following: given a sequence of scale coefficients or wavelets for a fixed level d^k and assigned a level dependent tolerance criterion ε_k, generating $\hat{d}^k = \left\{\hat{d}^k_j\right\}_{j=1}^{J_k - J_{k-1}}$ in accord to

$$\hat{d}^k_j = \mathrm{tr}(d^k_j, \varepsilon_k) = \begin{cases} 0 & |d^k_j| \leq \varepsilon_k \\ d^k_j & \text{otherwise}, \end{cases} \tag{2.18}$$

where $\mathrm{tr}(d^k_j, \varepsilon_k)$ indicates the truncation of d^k_j with respect to the local resolution threshold ε_k.

Different choices are described in literature for the threshold ε_k: a level independent choice $\varepsilon_k = \varepsilon$ or a dependent criterion $\varepsilon_k = \varepsilon / 2^{L-k}$. Since the original work of Harten, the question for the stability of the MR representation of the data has been analyzed. Harten proposed [38] to modify the *encoding* procedure to preserve a condition as follows

$$||v^L - \hat{v}^L|| \leq C\varepsilon, \tag{2.19}$$

with a constant C and measured in some norms as the L^1 and L^∞.

Classically the effectiveness of a MR approach is measured in term of its compression capability, *i.e.* the number of activated *wavelets* N_w with respect to the dimension of the discrete space at the finest level $\dim(V_L) = J_L$. The compression ratio μ_{cr} is usually defined as follows:

$$\mu_{\mathrm{cr}} = \frac{J_L}{N_w + J_0}. \tag{2.20}$$

This ratio is useful to measure the gain in terms of memory requirements.

However, in the Truncate and Encode (TE) approach presented in this work, the algorithm does not require the knowledge of the solution at the finest level. This is a totally original feature of the present approach, with respect to all the other MR schemes in literature. This leads to another ratio, an evaluation ratio τ, measuring the number of evaluations N_{eval} of the model with respect to the dimension of the full discrete space J_L

$$\tau = \frac{J_L}{N_{\mathrm{eval}}}. \tag{2.21}$$

In Chapter 4, results on both the compression ratio μ_{cr} and the evaluation ratio τ are provided for different applications of the TE algorithm in the point-value framework.

In a general MR framework, another aspect is crucial: the stability of the algorithm, *i.e.* when a function is compressed obtaining $\hat{v}_{MR}$, the decoding procedure on $\hat{v}_{MR}$ should recover the exact original v^L. This aspect becomes crucial for non-linear, *i.e.* data-dependent MR schemes. In the general framework presented here, the situation is slightly different: the algorithm searches for the smallest possible (in terms of evaluations required) representation $\hat{v}^L$ moving, recursively, from the level v^0. Once obtained, the discrete function $\hat{v}^L$ is not more manipulated, *i.e.* compressed; there is not a stability problem, concerning the duality between *encoding* and *decoding* procedures, however it remains crucial the question related to the convergence $||v^L - \hat{v}^L||$ in some norms. The interested reader can find a detailed analysis on the stability of the Harten framework with the relative error bounds in [16], while the convergence property of the TE algorithm, depending on the regularity of the function to reproduce, and on the order of the polynomial reconstruction employed to build the predictor operator, can be obtained. The derivation of these properties is explicitly made in the paper ***P1*** and numerically demonstrated, also in Chapter 4, for the so-called steady functions, *i.e.* functions depending only on the random space. In the following sections the MR framework, as described in this section, is discussed focused on UQ applications.

2.1.2 Limitation of the *pure* Harten's framework for UQ propagation

Two main properties of MR are of great interest in this thesis work. First, the compact MR representation of the functions permits to handle data with a great dimensionality (in term of the cardinality in the discrete space). Moreover, an adaptive time-dependent procedure can be designed exploiting the capability of the prediction error e^k to carry information relative to the (local) regularity of the functions. This quantity can be analyzed in order to guide the topological refinement in the stochastic space when the computation of the object function is performed. In particular, focusing on the reconstruction of an unknown function, the refinement is reached moving from the coarsest level towards the finest one, by comparing two resolution levels at once in terms of prediction error e^k. In this case, it is possible to identify the regions where the attained resolution is not sufficient to guarantee a prescribed tolerance criterion. A fundamental difference exists between this approach and the classical MR framework: in the latter, a MR representation can be obtained only if the finest resolution is available. This procedure can be directly applied in image compressing applications, where an image in high resolution is first translated in its MR representation and then, after the application of the truncation operator, compressed. However, in the case of initial value problems, as in the solution of time evolving PDEs, the procedure just described is useful only for the initial condition [39]. The initial solution should evolve in time and a proper MR scheme must exploit the scale decomposition of the function in order to update only the significant information. However, this approach employs a procedure to guess the movement of the significant points from a time n to a time $n+1$; generally a CFL based criterion is used for deterministic problems. A similar approach would fail if directly adopted for time-evolving problems in presence of random parameters; there is no possibility to obtain a pure evolution equation, with a CFL-like criterion, in the only stochastic space. Let us consider, for instance, a stochastic ODE. There is no way to predict the movement of significant points in the stochastic space. The reason is that each ODE (for a fixed random vector) is independent from the others. Moreover, the MR representation could vary due to the dependence of the pdf in time, or the presence of a bifurcation behavior with respect to the random parameters could produce an unexpected/unpredictable change in the function regularity. One of the main point to address is to build a procedure, in order to produce a MR representation of a time-dependent function at time $n+1$, without guessing its multiscale form at the previous time step n, but identifying it in *real-time*. In this sense, a Truncate and Encode (TE) algorithm permits to obtain a (already truncated) MR representation of a function, which is obtained by moving progressively from the coarsest resolution level to the finest one. The first step is to design the TE procedure in both the frameworks (point-values and cell-averages). This is sketched in §2.2 and §2.3. The TE procedure can be also coupled with a collocation-like approach to obtain the so-called spatial-TE (sTE) algorithm, as it is demonstrated in the papers ***P1***, ***P2*** and ***P3*** (presented in §2.2.2). This method is the fundamental step to bring a MR approach into the semi-*intrusive* scheme [4], where a direct link between the MR representation, the deterministic scheme and its stochastic extension is needed. It yields the adaptive-SI (aSI) scheme presented in the following chapter. The remaining part of the present chapter will be devoted to the development of the TE algorithm within a discussion of the proper reconstruction procedure to obtain the operator $\mathcal{R}_k$ in both the frameworks (point-value and cell-average) and in a linear and a non-linear fashion. As discussed in the previous chapter, a fundamental role in UQ propagation is played by the pdf of the input parameters. Further comments on the pdf are reported in the following section.

2.1.3 Remarks on the presence of non-uniform pdf

The general framework sketched above does not consider explicitly the presence of a probability measure. In UQ propagation problems, the presence of a distribution of the random parameters should be taken into account. Two different cases can occur: bounded or unbounded pdf. In the case of bounded pdf both regular, *i.e.* uniform, meshes based on the Lebesgue or probability measure can be built. In the first case the mesh is always uniform irrespective of the probability distribution of the parameters, while in the latter case the uniform tessellation is obtained directly on the probability measure. In this situation, each interval has the same probability, but the Lebesgue measure is not constant. If the probability distribution is not bounded, as for instance in the case of Gaussian distribution, the tessellation can be performed only on the probability measure [3]. Dealing with uniform meshes is conceptually and practically easier than handling non-uniform tessellation, so when possible, the uniform partition of the stochastic space should be employed. In this sense, even if the schemes developed here are designed also to deal with unbounded pdf, only bounded probability distributions are employed. However, non-standard and even discontinuous pdf are analyzed in the papers ***P1*** and ***P2*** where the point-value setting is adopted. In this case, the framework described above in an abstract way, is applied to the product between the function and the pdf of the random parameters hence represented on a regular (in the Lebesgue sense) mesh on the random space. In the cell-average framework, the connection between the probability measure and the MR representation is closer. The discretization operator $\mathcal{D}_k$ (see below) is based on a weighting function for the averaging step that, in this context, it is not difficult to identify with the pdf. More general averaging procedures are also possible, as the hat-based discretization [18, 34], but they are not investigated during this work because less adapted for UQ purpose (see [34], in particular, for a generalization of the discretization procedure based on averaging). In the original SI scheme [3], Abgrall employs an uniform tessellation on the measure probability, however the scheme is developed for fixed meshes in times. When dealing with the adaptation, the pdf is involved in the refinement procedure because each interval in the stochastic space can be divided again according to the Lebesgue or the probability measure. In the case of uniform distribution of the parameters obviously the two measures are equal (see the papers ***P3***, ***P4*** and ***P5***).

2.2 Point-value setting

The point-value setting represents, since its introduction by Harten in [39], a very flexible setting to build numerical schemes with a multiresolution rearrangement of the information. Conceptually, the point-value setting is the most natural one with only discretized data, *i.e.* data known in a set of finite points.

The appropriate functional space for this framework is the space of the bounded functions $f \in \mathcal{F} = \mathcal{B}(\Xi)$ with

$$f : \Xi \subset \mathbb{R}^d \to \mathbb{R}, \tag{2.22}$$

where Ξ must be also intended bounded with respect to its probability measure $\mathrm{d}\mu$. On the domain Ξ, let us suppose to generate nested sequences of points, also referred as a mesh of resolution level k, $\mathcal{G}^k = \left\{\boldsymbol{\xi}_j^k\right\}$ where $\boldsymbol{\xi}_j^k \in \Xi$. The sequence is nested if the following condition is satisfied

$$\mathcal{G}^{k-1} = \mathcal{G}^k \cap \mathcal{G}^{k-1}, \tag{2.23}$$

that allows the possibility to increase (decrease) the resolution level only adding (removing) a finite set of points for a fixed level k. For instance, in a 1D stochastic space, the situation is sketched in the figure 2.2. Hereafter, the exposition is made for a 1D stochastic problem for simplicity of exposure.

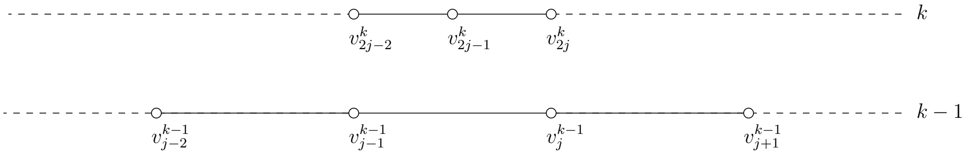


Figure 2.2: Example of 1D stochastic nested meshes for the point-value setting.

The nested property of the meshes directly turns into the nested character of the discretization operator

$$(\mathcal{D}_k f)_j = f(\xi_j^k) = v_j^k, \tag{2.24}$$

from which the discretization operator D_k^{k-1} is obtained directly removing from v^k all the components $v_j^k = f(\xi_j^k)$ where $\xi_j^k \in \mathcal{G}^k \setminus \mathcal{G}^{k-1}$.

The reconstruction operator $\mathcal{R}_k$ can be associated to the polynomial interpolation $\mathcal{P}_j^k$ on a fixed stencil $\mathcal{S}_j^k$ relative to the interval $[\xi_{j-1}^k, \xi_j^k]$. More details, about the selection of the stencil and the construction of the polynomial $\mathcal{P}_j^k$, are reported in the following sections. In this case, the prediction operator P_{k-1}^k can be defined as:

$$(\mathrm{P}_{k-1}^k v^{k-1})_{2j-1} = (\mathcal{D}_k \mathcal{P}_j^{k-1})_{2j-1} = \mathcal{P}_j^{k-1}(\xi_{2j-1}^k). \tag{2.25}$$

In this setting, the error e^k is equal to zero for all the points $\xi_j^k \in \mathcal{G}^{k-1}$, while the number of non-redundant, *i.e.* linear independent, coordinates d^k is equal to $\mathrm{card}(\mathcal{G}^k \setminus \mathcal{G}^{k-1})$, where the wavelets are defined as follows

$$d_j^k = v_{2j-1}^k - (\mathrm{P}_{k-1}^k v^{k-1})_{2j-1} \quad \forall \boldsymbol{\xi}_{2j-1}^k \in \mathcal{G}^k \setminus \mathcal{G}^{k-1}. \tag{2.26}$$

The components of the error e^k can be employed as an indicator for the enrichment of the discrete space following the TE algorithm. This algorithm is employed in the papers ***P1***, ***P2*** and ***P3*** where its formal definition can be found. In the Algorithm 1 the conceptual sketch of the strategy is provided.

Algorithm 1: Truncate and Encode algorithm for the point-value setting.

```
while $k < L$ do
    Encoding $(v^{k-1}, v^k) \rightarrow d^k$
    Truncation $(d^k, \varepsilon_k) \rightarrow \hat{d}^k$
    for $\xi_j^{k+1} \in \mathcal{G}^{k+1}$ do
        $v_j^{k+1} = \begin{cases} \textit{Evaluation} \rightarrow f(\xi_j^{k+1}) & \text{if} \quad e_\ell^k(\xi_j^{k+1}) > \varepsilon_k \\ \textit{Decoding}(v^k, \hat{d}^k) & \text{otherwise} \end{cases}$
    end
end
```

The error vector e^k contains components equal to zero, if a point belongs to both the resolution levels compared $\xi_j^k \in \mathcal{G}^k \cap \mathcal{G}^{k-1}$, or equal to the wavelet. In the Algorithm 1, the component associated to the interval to which ξ_j^{k+1} belongs is indicated as $e_\ell^k(\xi_j^{k+1})$. More formally there is a unique non-null component $0 \neq e_\ell^k \in \mathcal{I} \ni \xi_j^{k+1}$ with $\mathcal{I} \subset \Xi$ for $\xi_j^{k+1} \in \mathcal{G}^{k+1} \setminus \mathcal{G}^k$. Obviously, the *decoding* procedure relies on the reconstruction operator $\mathcal{R}_k$ (locally the polynomial $\mathcal{P}_j$) through the predictor P_{k-1}^k.

This MR point-value setting shows a great flexibility in terms of reconstruction operators $\mathcal{R}_k$. In this work, this flexibility is exploited and both linear and non-linear families of polynomial reconstructions are built to obtain P_{k-1}^k, and hence $\mathcal{R}_k$. Well established techniques for this scope are the essentially non oscillatory ENO or the WENO reconstruction procedures [42, 46, 57], extended in [6, 9, 58] for virtually any kind of meshes. The next section presents the techniques employed in the present work to perform the polynomial reconstruction in the point-value setting.

2.2.1 Non-linear polynomial reconstruction

The reconstruction operator $\mathcal{R}_k$, useful to perform the MR analysis, can be obtained as the union of all the polynomial interpolants $\mathcal{P}_j^k$ on each jth interval on the mesh $\mathcal{G}^k$. The strategy to obtain $\mathcal{P}_j^k$ is here described, for linear and non-linear cases, in the 1D context. During this work, the reconstruction procedure, for point-values setting, is not extended to multiple dimensions, but the general procedures to obtain efficiently Lagrangian interpolants, even in the non-linear case, are described in [6, 9].

The generic stencil $\mathcal{S}$ for a polynomial interpolation of order $r > 0$ is

$$\mathcal{S} = \mathcal{S}(r, s) = \{-s, -s+1, \dots, -s+r\}, \quad \text{with } r \geq s > 0. \tag{2.27}$$

On the stencil $\mathcal{S}$ it is possible to define a number of $N_{\mathcal{S}} = \mathrm{card}(\mathcal{S})$ Lagrange polynomials:

$$L_m(y) = \prod_{\substack{l=-s \\ l \neq m}}^{-s+r} \left(\frac{y-l}{m-l} \right) \quad \text{with} \quad L_m(i) = \delta_{i,m} \text{ and } i \in \mathcal{S}. \tag{2.28}$$

For each $\xi \in [\xi_{j-1}, \xi_j]$ the generic polynomial $\mathcal{P}_j$ (hereafter referred as q_j to distinguish from the cell-average setting) is defined as

$$q_j(\xi; f, r, s) = \sum_{m=-s}^{-s+r} v_{j+m} \, L_m \left(\frac{\xi - \xi_j}{h} \right), \tag{2.29}$$

where $q_j(\xi_l) = v_l = f(\xi_l)$. To each $q_j(\xi; f, r, s) \in [\xi_{j-1}, \xi_j]$ the topological stencil associated is $\mathcal{S}_j = \{\xi_{j-s}, \xi_{j-s+1}, \dots, \xi_{j-s+r}\}$ with $\mathrm{card}(\mathcal{S}_j) = r+1$. The degree of the polynomial reconstruction has direct consequences on the accuracy of the interpolation (see for instance [71]) and on the property of the MR reconstruction (see the paper ***P1*** where the convergence properties of the point-value TE scheme is obtained with respect to the polynomial reconstruction and the regularity of the function f).

If the stencil is fixed *a priori*, choosing both the degree r and the type of stencil s, the multiresolution framework is said to be linear (hence not data-dependent). It is well known (but further investigated in the paper ***P3***) that the error is minimized by selecting centered stencils. However, this general criterion is not sufficient for optimal polynomial interpolations for non smooth functions. This reflects also the difficulty to compute very small divided differences, which can be dominated by round-off errors.

In the MR context, a measure of the degradation of the interpolation is contained in the $r+1$ divided difference $f[\mathcal{S}_j, \xi]$. From this observation, Harten et al. introduced, in [42], the so-called Essentially non oscillatory (ENO) interpolation in the context of the numerical methods for conservation laws. The idea is to adapt the stencil, in presence of discontinuity, to avoid crossing it; the interpolation is carried out only using the regions of smoothness. Two different algorithms have been presented in [42]: a hierarchical selection and a non-hierarchical one. The non-hierarchical selection is demonstrated [17] to be able to detect even jump in the derivative of the functions. However, the non-hierarchical selection is, in the same paper, claimed to produce biased stencils away from discontinuity regions. For this reason, aiming to introduce the ENO technique in the MR context to gain in term of compression capabilities, the focus, in the present work, is on a hierarchical selection of the stencil employing the following algorithm [42]

Algorithm 2: Hierarchical selection of the stencil

```
$s_0 = j$
for $l = 0, \dots, r-2$ do
    if $|f[\xi_{s_l-2}, \dots, \xi_{s_l+l}]| < \left|f[\xi_{s_l-1}, \dots, \xi_{s_l+l+1}]\right|$ then
        $s_{l+1} = s_l$
    end
end
$s_j = s_{r-1}$
```

to obtain the stencil $\mathcal{S}_j^{\mathrm{ENO}} = \left\{\xi_{s_j-1}, \xi_{s_j}, \dots, \xi_{s_j+r-1}\right\}$ and where, the generic divided difference, is

$$f[\xi_0, \dots, \xi_n] = \sum_{j=0}^{n} \frac{f(\xi_j)}{\prod_{k\in\{0,\dots,n\}\setminus\{j\}} (\xi_j - \xi_k)}. \tag{2.30}$$

In papers ***P1*** and ***P2***, the linear reconstruction ($r = 1$) is adopted for the point-value setting. Paper ***P3*** presents the extension, to the high-order ($r = 3$) non-linear (ENO) procedures, to select the stencil obtaining better results with respect to both the low order and the linear schemes. The ENO interpolation consists in comparing r stencils each containing $r + 1$ points. The overall stencil (virtually) visited during the ENO evaluation contains $2r$ points that, at least in the smooth regions, could be used to produce high-order approximations. To exploit this possibility the resulting method, called Weighted-ENO, has been introduced in 1994 by Liu, Osher and Chan [57] in which they presented a third order accurate finite volume WENO scheme, and the generalization to arbitrary order in [46]. In the following the WENO technique is presented and extended to the TE scheme purpose, following [15].

WENO interpolation

The WENO interpolation techniques have been introduced to take advantage from all the information collected during the selection of the (best) stencil in a ENO algorithm (see for instance the Algorithm 2). Restricting the presentation to 1D case[3], and avoiding to explicitly reporting the kth resolution level, the ENO procedure to obtain the polynomial $\mathcal{P}_j(\xi)$ with $\xi \in [\xi_{j-1}, \xi_j]$ of order r selects the smoothest stencil among $\mathcal{S}_j(s) = \{\xi_{j-s}, \xi_{j-s+1}, \dots, \xi_{j-s+r}\}$ where $r \geq s > 0$. The union of the r stencils $\mathcal{S}_j(s)$ is the set $\mathcal{S}^{\cup} = \{\xi_{j-r}, \xi_{j-s+1}, \dots, \xi_{j-1+r}\}$ with cardinality $2r$. Despite the ENO algorithm employs information originating from $2r$ points, the maximum accuracy possible is limited by the cardinality $r + 1$ of each single stencil $\mathcal{S}_j(s)$ because, at the end of the procedure, only one of them is retained to build the polynomial $q_j(\xi)$. The seminal idea proposed in [57] was to build a convex combination of all the possible polynomials $q_j(\xi; f, r, s)$ obtained on the stencils $\mathcal{S}_j(s)$ in a way to recover the polynomial of the maximum possible order on the stencil $\mathcal{S}^{\cup}$ while, in the presence of a lack of regularity, reducing to the ENO selection. Basically, the combination of the polynomials is based on the smoothness of the function on each stencil in a way to obtain, practically, no influence from the stencil containing singularity points. It is

[3] This assumption is not so strong as it would appear. In the context of point-value MR, the multidimensional schemes are obtained through a recursive application of 1D algorithm via tensorial product (see for instance [12, 13, 22]).

important to note that both the original paper [57] and its generalization [46] handle cell-average interpolations as of interest for the solution of the conservation laws. In this thesis, however, the approach follows what has been proposed by Aràndiga, Belda and Mulet in [15] where the procedure is adapted to the Lagrangian interpolation and the (non-linear) weights, for the convex combinations of the polynomials, are obtained adapting also the smoothness criterion.

The formal definition of the WENO interpolant $q_j^{\mathrm{W}}(\xi)$ (omitting the explicitly dependence from the function and the polynomial order) is the convex combination of all the polynomial $q_j(\xi; s)$ defined over each stencil $\mathcal{S}_j(s)$

$$q_j^{\mathrm{W}}(\xi) = \sum_{s=1}^{r} \omega_s \, q_j(\xi; s) \quad \text{where} \quad \omega_s \geq 0 \quad \text{and} \quad \sum_{s=1}^{r} \omega_s = 1. \tag{2.31}$$

The first requirement consists in determining the optimal coefficients ω_s defined as the coefficients for which the convex combination (2.31) recovers the highest order ($2r$) of accuracy on $\mathcal{S}^{\cup}$. The optimal weights ω_s^{opt} are recovered as

$$\omega_s^{\mathrm{opt}} = \frac{1}{2^{2r-1}} \binom{2r}{2s-1}. \tag{2.32}$$

Note that the equation (2.32) contains a slightly change of notation with respect to its counterparts reported in [15] because in this work the weights are directly obtained as function of the stencil index s. The optimal weights for r, up to four, are reported in the table 2.1.

	$s=1$	$s=2$	$s=3$	$s=4$
$r=2$	1/2	1/2	-	-
$r=3$	3/16	10/16	3/16	-
$r=4$	1/16	7/16	7/16	1/16

Table 2.1: Optimal weights for the WENO interpolation (2.31) obtained by the (2.32).

Non linear weights can be obtained to emulate the ENO interpolation, *i.e.* the polynomial corresponding to a stencil containing a discontinuity should have a negligible influence in (2.31). In both [57] and [15], the non-linear weights are defined as follows

$$\omega_s = \frac{\alpha_s}{\sum_{i=1}^{r} \alpha_i}, \quad \text{where} \quad \alpha_s = \frac{\omega_s^{\mathrm{opt}}}{(\nu_h + \mathcal{SI}_s)^t}, \tag{2.33}$$

where the smoothest indicator $\mathcal{SI}_s = \mathcal{SI}_s(h)$ depends on the function to interpolate over the stencil $\mathcal{S}_j(s)$ and ν_h is introduced, in principle, only to avoid null denominators. However a proper dependence from h is of a crucial importance for the choice of ν_h. The following property holds (adapted from [15])

Proposition 1 *Let $\mathcal{SI}_s$ be the smoothness indicators of $f(\xi)$ on $\mathcal{S}_j(s)$ that verify $\mathcal{SI}_s = \mathcal{O}(h^2)$, if $f(\xi)$ is smooth at $\mathcal{S}_j(s)$; $\mathcal{SI}_s \nrightarrow 0$ for $h \to 0$ if $f(\xi)$ is not smooth at $\mathcal{S}_j(s)$ and*

$$\mathcal{SI}_p - \mathcal{SI}_q = \mathcal{O}^{m+2}, \quad \forall 1 \leq p, q \leq r,$$

for some $m \geq r-1$ whenever f is smooth at all the stencils $\mathcal{S}_j(s)$, $0 < s \leq r$. Moreover, choosing $\nu_h = h^2$ then $f\left(\frac{\xi_{j-1}+\xi_j}{2}\right) - q_j^{\mathrm{W}}\left(\frac{\xi_{j-1}+\xi_j}{2}\right) = \mathcal{O}(h^{2r})$ if f is smooth, otherwise

$f\left(\frac{\xi_{j-1}+\xi_j}{2}\right) - q_j^{\mathrm{W}}\left(\frac{\xi_{j-1}+\xi_j}{2}\right) = \mathcal{O}(h^{\min(2t,r+1)})$ *if f it is smooth at least at one of the stencils $\mathcal{S}_j(s)$.*

Proposition 1 drives the choice of the parameters for determining the non-linear weights (2.33): ν_h should be a function of the size of the mesh h and in particular $\nu_h = h^2$, while the exponent t influences the order of the interpolation in the case of non smooth functions. In particular, it appears convenient to choose $2t \geq r+1$ (and $\nu_h = h^2$) to guarantee an accuracy at least equal to the ENO interpolation (r) in the same case, while in the smooth region the order $2r$ can be (theoretically) achieved also with a fixed ν_h. In practice, in *standard* WENO applications, as the solution of systems of conservation laws on fixed regular meshes, the parameter ν_h is usually fixed as $10^{-5} \sim 10^{-6}$ to realize $\nu_h \approx h^2$ for meshes of the size $h \approx 10^{-3}$. However, in this context, the most robust choice is a function ν_h depending on h in order to obtain the highest accuracy even on the meshes of the coarser resolution levels.

The remaining term to compute is the smoothness indicator $\mathcal{SI}_s$ for each stencil. In [57], the authors proposed the following definition for cell-average applications

$$\mathcal{SI}_s = \sum_{l=1}^{r-1} \int_{\xi_{j-1/2}}^{\xi_{j+1/2}} h^{2l-1} (q_j^{(l)}(\xi; s))^2 \mathrm{d}\xi, \tag{2.34}$$

while in [15] the previous definition is adapted to the point-value case. Only shifting the interval it is possible to write

$$\mathcal{SI}_s = \sum_{l=1}^{r-1} \int_{\xi_{j-1}}^{\xi_j} h^{2l-1} (q_j^{(l)}(\xi; s))^2 \mathrm{d}\xi, \tag{2.35}$$

where the term h^{2l-1} is retained to remove the $h-$dependence in the lth derivative of q_j. The integration in the equation (2.35) can be approximated by the mid-point rule or the trapezoidal rule obtaining quite equivalent versions of the index. In this thesis, a comparison between the three indicators proposed in [15] has been performed, but they show to be equivalent in the TE algorithm framework. After some manipulations, in the case of $r=3$ adopted in this thesis, the smoothness indicator yields

$$\mathcal{SI}_s = (\delta_1(s) + \delta_2(s) + \delta_3(s))^2 + \frac{13}{3}\left(\delta_2(s) - \frac{3}{2}\delta_3(s)\right)^2 + \frac{781}{20}\delta_3^2(s), \tag{2.36}$$

where δ_1, δ_2 and δ_3 depend on the values of the function f on the points belonging to each stencil, hence they vary between the stencils $\mathcal{S}_j(s)$. In the following, the expressions as functions of the stencil $\mathcal{S}_j = \{\xi_{j-s}, \xi_{j-s+1}, \dots, \xi_{j-s+r}\}$ are reported:

$$s = 3 \rightarrow \begin{cases} \delta_1 = -\frac{1}{3}v_{j-3} + \frac{3}{2}v_{j-2} - 3v_{j-1} + \frac{11}{6}v_j \\ \delta_2 = -\frac{1}{2}v_{j-3} + 2v_{j-2} - \frac{5}{2}v_{j-1} + v_j \\ \delta_3 = -\frac{1}{6}v_{j-3} + \frac{1}{2}v_{j-2} - \frac{1}{2}v_{j-1} + \frac{1}{6}v_j, \end{cases}$$

$$s = 2 \rightarrow \begin{cases} \delta_1 = \frac{1}{6}v_{j-2} - v_{j-1} + \frac{1}{2}v_j + \frac{1}{3}v_{j+1} \\ \delta_2 = \frac{1}{2}v_{j-1} - v_j + \frac{1}{2}v_{j+1} \\ \delta_3 = -\frac{1}{6}v_{j-2} + \frac{1}{2}v_{j-1} - \frac{1}{2}v_j + \frac{1}{6}v_{j+1}, \end{cases}$$

and

$$s = 1 \rightarrow \begin{cases} \delta_1 = -\frac{1}{3}v_{j-1} - \frac{1}{2}v_j + v_{j+1} - \frac{1}{6}v_{j+2} \\ \delta_2 = \frac{1}{2}v_{j-1} - v_j + \frac{1}{2}v_{j+1} \\ \delta_3 = -\frac{1}{6}v_{j-1} + \frac{1}{2}v_j - \frac{1}{2}v_{j+1} + \frac{1}{6}v_{j+2}, \end{cases}$$

Numerical results, concerning the WENO interpolation introduced in the point-value resolution scheme, are not present in any of the papers which constitutes this manuscript. For this reason some numerical results to highlight the difference between linear and non-linear (ENO and WENO) reconstructions, in the point-value framework, are reported in Chapter 4 of this manuscript. The point-value MR framework, presented in §2.2, is the fundamental brick of the spatial-TE (sTE) scheme introduced to solve problems where the solution depends also from the physical and time space. The sTE algorithm is depicted in the following section summarizing what is done in ***P2*** and ***P3***.

2.2.2 Introducing the spatial-TE algorithm for stochastic partial differential equations

The TE algorithm, as shown in the previous sections, allows to reconstruct with a low computational cost, *i.e.* a less number of functional evaluations, a function over the stochastic space on a mesh with a certain finest resolution level, moving from the coarsest to the finest resolution. When dealing with stochastic PDEs, as introduced in (1.7), the solution of the problem depends also on the physical space Ω and the time space T. Conceptually, the stochastic space can be viewed as a supplementary dimension and the solution, at each time step, could be represented on a coupled physical/stochastic space. However, this approach becomes a pure multidimensional approach even if $\Omega \subset \mathbb{R}$ and $\Xi \subset \mathbb{R}$. In paper ***P3***, this approach is defined as *strong* coupling. During this thesis, the focus is concentrated to a *weak* coupling approach: each space, physical or stochastic, remains discretized by its own representation. Despite its conceptual simplicity, the *weak* coupling yields a very flexible approach allowing to introduce a MR representation in problems where the spatial formulation has already been obtained. In this sense, the *weak* coupling shares the same vision of the original semi-*intrusive* scheme of Abgrall [3]. To make clear the idea, let us imagine a tessellation of the physical space $\Omega \supset \mathcal{T} = \bigcup_{i=1}^{N_x} \mathcal{T}_i$ in which N_x is the number of elements. The overall space $\Omega \times \Xi$ can be approximated as

$$\Omega \times \Xi \simeq \left(\bigcup_{i=1}^{N_x} \mathcal{T}_i \right) \times \Xi = \bigcup_{i=1}^{N_x} \mathcal{T}_i \times \Xi. \tag{2.37}$$

The discrete solution $u(\mathbf{x}, t, \boldsymbol{\xi})$ can be approximated as

$$u(\mathbf{x}, t, \boldsymbol{\xi}) \simeq \bigcup_{i=1}^{N_x} u(\mathbf{x}_i, t, \boldsymbol{\xi}). \tag{2.38}$$

If the time space $T = [0, t_F]$ is also discretized (let us suppose N_t constant intervals $\Delta t = t_F/N_t$ for a sake of simplicity of exposure) and $t_n = n\Delta t$ for $n = 0, \ldots, N_t$, the semi-discrete (time-physical space) solution reads

$$u(\mathbf{x}, t_n, \boldsymbol{\xi}) = \bigcup_{i=1}^{N_x} u(\mathbf{x}_i, t_n, \boldsymbol{\xi}). \tag{2.39}$$

Fixing a spatial coordinate $\mathbf{x}_i$ and a time step t_n, the semi-discrete solution is a collection of function defined over Ξ; all the functions $u(\mathbf{x}_i, t_n, \boldsymbol{\xi})$ are discretized on resolution level L. Hence, the TE algorithm described above can be applied on each $u(\mathbf{x}_i, t_n, \boldsymbol{\xi})$ to obtain their compressed representation. As it is evident, the approximation procedure just described is not influenced by the spatial discretization. For instance, in ***P2*** a finite-element discretization is employed, while in ***P3*** a finite-volume tessellation is adopted for the space Ω. For the stochastic space Ξ, the MR basis, defined on different discrete resolution spaces V^k, is the approximation space.

From an algorithmic point-of-view, the TE procedure plays two fundamental roles: first it drives the overall algorithm to locate points in the stochastic space where the solution should be computed; secondly the MR representation, obtained through the TE algorithm, consists in a reduced set of points which represents the independent expansion coefficients for each solution $u(\mathbf{x}_i, t_n, \boldsymbol{\xi})$ on the space of the piecewise polynomials of degree r. The approximation space, in the case of the non-linear MR setting, will be dependent on the physical space and the time, as well as from the stochastic space. In the point-value framework, the approximation of $u(\mathbf{x}_i, t_n, \boldsymbol{\xi})$ is fully determined knowing the value of $\{u(\mathbf{x}_i, t_n, \boldsymbol{\xi}_j)\}$ on a set of N_ξ collocation points in the stochastic space. As already shown, each function is represented on a finest resolution level through a hierarchical representation on the lower levels. Main difference, with respect to the pure TE algorithm, is that in the sTE, each $u(\mathbf{x}_i, t_n, \boldsymbol{\xi}_j^k)$ depends (explicitly) on $u(\mathbf{x}_i, t, \boldsymbol{\xi}_j^L)$ for $t < t_n$ when evaluating. Conceptually, the only additional step, with respect to the TE algorithm, consists in the introduction of the spatial discretization and time discretization, but at a fixed stochastic coordinate; this approach makes possible to employ the same theoretical and numerical framework adopted in the deterministic counterparts. To make things clear, let us consider a scalar random conservation laws discretized by a Godunov method in a node-centered approach on a regular mesh where each cell is $\mathcal{C}_i = [x_{i-1/2}, x_{i+1/2}]$ and $x_{i\pm1/2}$ are the interfaces. If a standard first order discretization is adopted in space and time, for a fixed random coordinate ξ_j, it is possible to write

$$u(\mathbf{x}_i, t_{n+1}, \xi_j) = u_i^{n+1}(\xi_j) = u_i^n(\xi_j) + \frac{\Delta t}{\Delta x}\Big(F_L(x_{i-\frac{1}{2}}, t_n, \xi_j) - F_R(x_{i+\frac{1}{2}}, t_n, \xi_j)\Big). \tag{2.40}$$

In the case of standard Godunov method, each numerical flux evaluated at the interfaces is function of only the cell average values on the cell sharing the same interface, *i.e.* $F_L(x_{i-\frac{1}{2}}, t_n, \xi_j) = F_L(u_{i-1}^n, u_i^n, \xi_j)$ and $F_R(x_{i+\frac{1}{2}}, t_n, \xi_j) = F_L(u_i^n, u_{i+1}^n, \xi_j)$. The value of $u_i^{n+1}(\xi_j)$ can be obtained as a series of operations providing the stencil of values $\{u_{i-1}^n, u_i^n, u_{i+1}^n\}$. For a general space discretization, if the time discretization is explicit, a proper stencil can be identified. In papers ***P3***, this stencil is called *physical vector* and indicated with $\mathbb{PV}$. The application with different $\mathbb{PV}$ can be found in ***P2***, ***P3***.

It important to note that the *physical vector* contains elements always belonging to different MR representations because each MR representation is performed at different spatial coordinates. In general, a procedure of *physical assembling* PhAs is needed (see for instance ***P2*** and ***P3***). Conceptually, the PhAs procedure is the searching for each element of $\mathbb{PV}$ in the relative MR structure. In general, the problem corresponds (turning to the notation employed for the TE algorithm) to find the value of v_j^k knowing, *i.e.* having stored, a different resolution level. Dealing with time-varying problems with traveling discontinuities or high gradient regions for each spatial location, the final resolution level can vary. However, a sequence of decimations $\mathcal{D}_k$ can be applied to obtain the value from an higher resolution level, or a sequences of prediction P_{k-1}^k can be employed to obtain the value from a lower resolution level. If the latest level obtained by the TE algorithm is retained, the sequences of prediction is exactly (under the tolerance prescribed via ε) the value searched because all the (truncated) details for the higher resolution levels are zero; this is not the case if the solution would be further compressed at the end of the TE algorithm. In this case, clearly, the truncated details are not all zero. More formally searching for the value v_j^k on a generic level k knowing $v^{\bar{k}}$, yields

$$v_j^k = \begin{cases} (\mathrm{D}_{k+1}^k \cdots \mathrm{D}_{\bar{k}-1}^{\bar{k}-2} \mathrm{D}_{\bar{k}}^{\bar{k}-1} v^{\bar{k}})_j & \text{with} \quad k < \bar{k} \\ (\mathrm{P}_{k-1}^k \mathrm{P}_{k-2}^{k-1} \cdots \mathrm{P}_{\bar{k}}^{\bar{k}+1} v^{\bar{k}})_j & \text{with} \quad k > \bar{k}. \end{cases} \tag{2.41}$$

The PhAs algorithm is then performed to obtain all the elements belonging to $\mathbb{PV}$ employing the (2.41). The final sTE algorithm is the recursive execution of the TE Algorithm 1 for all the time steps $n < N_t$ and for all the spatial points $i < N_x$ where, the *evaluation* step is the application of the deterministic scheme providing the *physical vector* $\mathbb{PV}$ obtained recurring to the PhAs procedure. The complete sTE algorithm described in this chapter is reported in **P3**.

2.3 Cell-average setting

In section §2.2, the case of discretization in a point-value setting has been presented. However, another approach to represent discrete data relies on cell-average framework. This approach recovers an important role for all the class of methods where integral quantities are a more natural way to deal with the discrete equations, as for instance, finite volume methods. In this case the function $f = f(\boldsymbol{\xi})$ is $f : \Xi \subset \mathbb{R}^d \to \mathbb{R}$ with d the number of dimensions of the stochastic problem, *i.e.* the number of uncertain parameters. The functional space $\mathcal{F}$ is defined as $\mathcal{F} = L^2(\Xi)$, because of the cell-average setting, and the need to have functions with finite variance in the UQ framework. In the cell-average setting, a weighting function should be adopted. In the UQ context, this weighting function is easily identified as the joint pdf of the random inputs. However, general averaging procedures, as the hat-based [18, 34], exist in literature. In particular in paper [34], generalizations of the discretization procedure, based on averaging, are reported. Choosing the pdf as the weighting function, the space Ξ is equipped with the measure

$$\mathrm{d}\mu(\boldsymbol{\xi}) = p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}. \tag{2.42}$$

Hence, a tessellation of the stochastic space Ξ, satisfying the classical non overlapping requirements, can be considered

$$\Xi = \bigcup_{j=1}^{N_\xi} \Xi_j, \quad \text{with} \quad \Xi_i \cap \Xi_j = 0 \quad \text{if} \quad i \neq j. \tag{2.43}$$

In this setting, the discretization operator on the k-th level can be defined over the j-th cell Ξ_j^k as

$$(\mathcal{D}_k f)_j \overset{\text{def}}{=} \frac{1}{\mu(\Xi_j^k)} \int_{\Xi_j^k} f(\boldsymbol{\xi}) \mathrm{d}\mu(\boldsymbol{\xi}) = v_j^k. \tag{2.44}$$

By an agglomeration (splitting) procedure with a generic mesh, even unstructured, it is always possible to obtain a less (higher) resolution level. In a general case, to each cell Ξ_j^k at the lower resolution level corresponds a certain number of cells at the higher resolution level. In order to preserve the nested character between levels, the following properties between meshes must hold:

$$\Xi_j^k = \sum_{l=1}^{\bar{l}_c} \Xi_{jl}^{k+1}. \tag{2.45}$$

In this case, the decimation operator could be obtained as follows (see figure 2.3 for a representation in the 1D case) thanks to the additivity of integrals

$$\begin{aligned}(\mathrm{D}_k^{k-1} v^k)_j = (\mathrm{D}_k^{k-1} \mathcal{D}_k f)_j = (\mathcal{D}_{k-1} f)_j &= \frac{1}{\mu(\Xi_j^{k-1})} \int_{\Xi_j^{k-1}} f(\boldsymbol{\xi}) \mathrm{d}\mu(\boldsymbol{\xi}) \\ &= \frac{1}{\mu(\Xi_j^{k-1})} \sum_{l=1}^{\bar{l}_c} \mu(\Xi_l^k)(\mathcal{D}_k f)_l.\end{aligned} \tag{2.46}$$

Note that in the general case of an arbitrary pdf $p(\boldsymbol{\xi})$, even the 1D case sketched in figure 2.3, the nested sequence produces nested relations with non constant coefficients even for the same level of resolution depending on the measure. To recover the counterparts of the physical space case, the splitting/agglomeration of each cell, based on a Lebesgue measure, should be replaced by a splitting based directly on the probability measure. The nested sequence of the meshes, even in this case, is totally independent on the function and can be generated *a priori* with the only requirement to know the probability distribution $p(\boldsymbol{\xi})$.

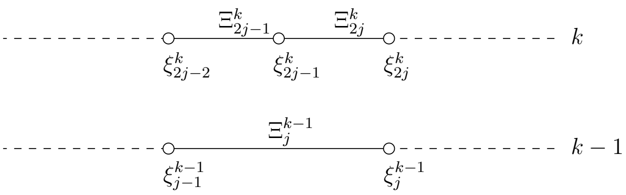


Figure 2.3: Example of 1D stochastic nested meshes for the cell-average setting decimation procedure.

However, in this work, uniform pdfs are employed for the numerical test cases, then the two *criteria*, the splitting/agglomeration based on the Lebesgue or probability measure, are the same. In particular, the corresponding discretization operator can be defined by the convolution of the function f with an Haar scaling function. This approach constitutes a direct link between the Harten framework and the wavelet one.

In the cell average framework, historically, the reconstruction operator $\mathcal{R}_k$ has been introduced by Harten in the 1D case invoking a primitive function. This approach can be extended quite easily, to multidimensional spaces, in the case of regular meshes, *i.e.* tensorial product meshes [16,22]. However, as pointed out in [12], the primitive function is only a design tool and, in practice, the reconstruction reduces to a relation between the cell-average values. In the case of unstructured meshes, the problem becomes very complicated. To avoid Gibbs-like *phenomena* the introduction of the ENO procedure for multidimensional unstructured grids is necessary. Quoting van Leer in [52]

> The ENO procedure is the only known non-oscillatory interpolation that allows a truly multidimensional extension, albeit very costly. The single paper about this subject is due to the French numerical analyst Rémi Abgrall [1].

In this work, the MR scheme is applied to regular mesh, even if not structured, as it will be evident in the following. However, the general procedure to obtain the conservative polynomial reconstruction, without employing the primitive function, has been described in [9] even for multidimensional problems on unstructured meshes [6].

The schemes adopted during this thesis are non-linear in the 1D case (see papers ***P4*** and ***P5***), while the linear case, with $r = 1$ has been used in the multidimensional 2D and 3D (stochastic space) case. In the following, the general reconstruction procedure is described, while the other details, as for instance the selection of the stencils, are reported in section §2.3.1. Fixed a polynomial degree of reconstruction r, a stencil $\mathcal{S}_j^k$ of cells with cardinality $\mathrm{card}(\mathcal{S}_j^k)$ can be fixed. On each stencil $\mathcal{S}_j^k$, a polynomial $\mathcal{P}_j^k(\boldsymbol{\xi}; f)$ of degree r can be constructed. The admissibility of this kind of stencils remains subject to a Vandermonde type condition (see [9] for further details). Here, supposing the stencils admissible, the conditions to satisfy, for the computation of the unique polynomial $\mathcal{P}_j^k$, attaining the conditional average $\mathcal{D}_k(f)_l$

$$\mathcal{D}_k(\mathcal{P}_j^k(\boldsymbol{\xi}; f))_l = \mathcal{D}_k(f)_l, \quad \forall l \in \mathcal{S}_j^k. \tag{2.47}$$

The reconstruction operator $\mathcal{R}_k$ is exactly equal to the union of all the polynomials $\mathcal{P}_j^k$ defined on all the cells Ξ_j^k. This makes possible, without introducing confusion, to change $\mathcal{R}_k$ with $\mathcal{P}_j^k$ when the cell Ξ_j^k is of interest.

The prediction operator P_{k-1}^k is obtained following its definition (2.9), using first the reconstruction procedure (2.47) for the level $k-1$, and then applying the discretization operator $\mathcal{D}_k(\mathcal{P}_j^{k-1})$ relative to the level k.

The remaining step is to define a relation between the error e^k and its, linear independent, coordinates d^k (the wavelets). In the general case (2.45), a number of $\bar{l}_c - 1$ linear dependent relations for the components of e^k must hold. In this work, the case of tensorial dyadic splitting ($\bar{l}_c = 2^d$) is addressed and for each cell the linear independent components of the error vector e^k, the wavelets d^k, are $\bar{l}_c - 1$, hence 1, 3, 7 respectively in 1D, 2D and 3D stochastic spaces.

To obtain a relation between the error component $e^k_{\bar{l}_c} = v^k_{\bar{l}_c} - (\mathrm{P}^k_{k-1}v^{k-1})_{\bar{l}_c}$ with respect to the $\bar{l}_c - 1$ linear independent wavelets, it is necessary to recall the following relation (referring to the splitting of the generic cell Ξ^{k-1}_j at level $k-1$ in $\bar{l}_c$ cells at level k)

$$\sum_{l=1}^{\bar{l}_c} \mu(\Xi^k_l)v^k_l = \mu(\Xi^{k-1}_j)v^{k-1}_j \rightarrow v^k_{\bar{l}_c} = \frac{1}{\mu(\Xi^k_{\bar{l}_c})}\left(\mu(\Xi^{k-1}_j)v^{k-1}_j - \sum_{l=1}^{\bar{l}_c-1}\mu(\Xi^k_l)v^k_l\right), \quad (2.48)$$

while the predicted value $(\mathrm{P}^k_{k-1}v^{k-1})_{\bar{l}_c}$ can be obtained as combination of the predicted values of the remaining $\bar{l}_c - 1$ cells at level k

$$(\mathrm{P}^k_{k-1}v^{k-1})_{\bar{l}_c} = \frac{1}{\mu(\Xi^k_{\bar{l}_c})}\left(\mu(\Xi^{k-1}_j)v^{k-1}_j - \sum_{l=1}^{\bar{l}_c-1}\mu(\Xi^k_l)(\mathrm{P}^k_{k-1}v^{k-1})_l\right). \quad (2.49)$$

The final expression for $e^k_{\bar{l}_c}$ is obtained as

$$e^k_{\bar{l}_c} = -\frac{1}{\mu(\Xi^k_{\bar{l}_c})}\left(\sum_{l=1}^{\bar{l}_c-1}\mu(\Xi^k_l)(v^k_l - (\mathrm{P}^k_{k-1}v^{k-1})_l\right) = -\sum_{l=1}^{\bar{l}_c-1}\frac{\mu(\Xi^k_l)}{\mu(\Xi^k_{\bar{l}_c})}d^k_l. \quad (2.50)$$

Again, for each cell, it is necessary to compute only $\bar{l}_c - 1$ details to know all the error components. From an implementation point-of-view, it is important to remark that in the present work the MR framework is introduced not only to represent in a compact way functions defined in the stochastic space, but also to refine the tessellation in the regions with the highest gradients or discontinuities. Hence, the TE algorithm is performed from the coarsest to the finest resolution level and, according to the numerical evaluation of the cell averages, the relation expressed by the equation (2.48) should be imposed. For this reason, the so-called Discretize Agglomerate Decimate (DAD) algorithm is introduced. The DAD algorithm can be summarized as follows

```
Algorithm 3: DAD algorithm
  Discretization:
      $v^k_l = \frac{1}{\mu(\Xi^k_l)}\int_{\Xi^k_l} f(\boldsymbol{\xi})\mathrm{d}\mu(\boldsymbol{\xi})$ for $l = 1,\dots,\bar{l}_c$
  Agglomeration:
      $\mu(\Xi^{k-1}_j) = \sum_{l=1}^{\bar{l}_c}\mu(\Xi^k_l)$
  Decimation:
      $(\mathrm{D}^{k-1}_k v^k)_j = v^{k-1}_j = \frac{1}{\mu(\Xi^{k-1}_j)}\sum_{l=1}^{\bar{l}_c}\mu(\Xi^k_l)v^k_l$
```

The DAD algorithm force the relation (2.48) to hold, preventing from round-off and approximation errors in the computations of the integral quantities. In practice, when a prediction operator should be computed (through the reconstruction operator) the cell values of the previous resolution level $k-1$ are obtained via the more accurate values (agglomerating the relative cells) at the finer resolution level k. The DAD algorithm permits to obtain the error vector in the null space of its resolution level with the advantage to express it by only its wavelets components.

Moving from the coarsest to the finest resolution level can be performed in two ways. The first possibility is to compute each resolution level entirely, *i.e.* evaluating the value in each cell having a local error greater of the threshold and predicting the values in the other cells. In this case, data structure evolves, towards the highest resolution levels, through levels constituted of equally spaced cells (in the measure chosen). Otherwise, only the cells to evaluate are retained in the evolving data structure, while the other cells are maintained at their last resolution level reached. In the figure 2.4, the two approaches are presented. Both strategies have advantage and disadvantages. The first one is the most expensive among the two, because the number of operations could be very high if the highest resolution level is very refined. Moreover this aspect could be prohibitive in a multi-dimensional context in which the number of cells would grow exponentially.

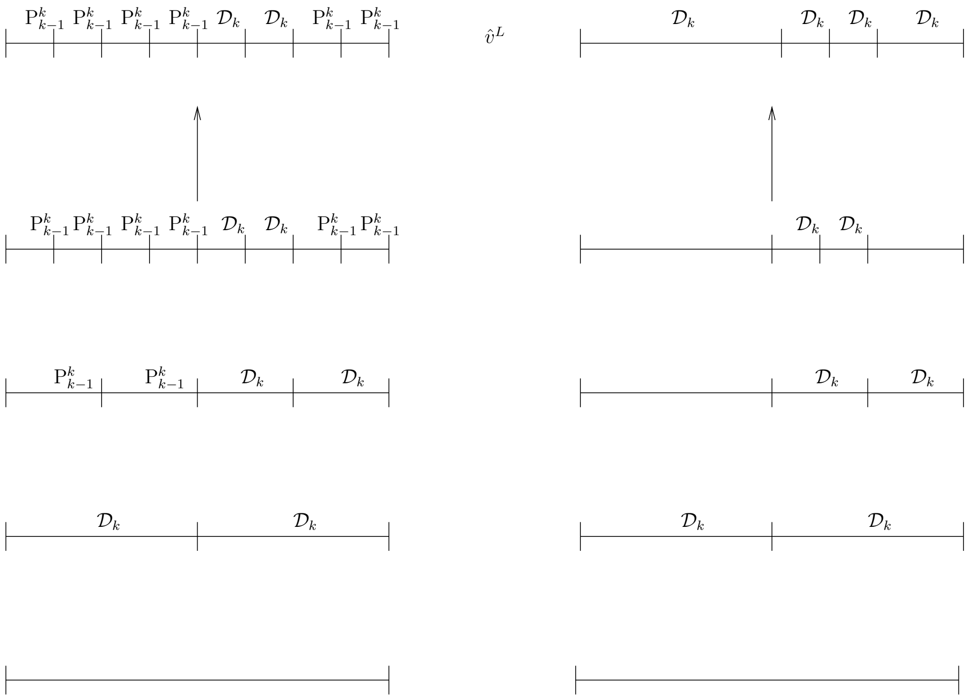


Figure 2.4: Two different level advancing techniques. On the left the approach with the entire resolution level (employing recursively the prediction operator P^k_{k-1} to obtain the discrete values in the cells where $\mathcal{D}_k$ is not applied) and on the right the approach based on the refinement dependent data-structure.

Concerning memory requirement, the effort can be drastically reduced storing the value of only the significant cells. For this reason, in this work, both in the 1D and the multi-D cases, the approach chosen is to employ an evolving data structure. However, an evolving data structure requires a major implementation effort and also makes more difficult the reconstruction step. This is due to the presence of non-uniform grids in 1D context while, in the case of multidimensional problems, the mesh becomes non conformal. In the 1D case, the reconstruction can be obtained via a generalization of the reconstruction procedure with non constant coefficients, while in the multidimensional case, in this work, the issue is solved choosing a proper stencil, for a linear reconstruction procedure, entirely contained in the mother cell.

Let us consider a generic cell $\Xi^k_j \in \mathcal{G}^k$. The TE algorithm in the cell-average setting can be obtained by the following Algorithm 4

Algorithm 4: TE algorithm for the cell-average setting

```
while $1 \leq k \leq L$ do
    for $\Xi_j^k \in \mathcal{G}^k$ do
        DAD
    end
        $\Longrightarrow v^{k-1}, v^k$
    if $k \neq L$ then
        Encoding $(v^{k-1}, v^k) \to d^k$
        Truncation $(d^k, \varepsilon_k) \to \hat{d}^k$
        Level advancement $(\mathcal{G}^k, \hat{d}^k) \to \mathcal{G}^{k+1}$
    end
end
```

As evident from the Algorithm 4, if the discrete structure is updated at each time step, then the sequences of meshes $\{\mathcal{G}^k\}$ is not known *a priori* and an advancement procedure must be introduced. This compares the local components of the error with the local threshold by dividing the cell (at the finer resolution level) if the accuracy criterion is not respected. The reconstruction operator $\mathcal{R}_k$, contained in the *encoding* step, is mandatory for comparing the predicted value and the exact one. The local polynomial reconstruction must be performed on non uniform meshes and the coefficients of the interpolation cannot be computed once at the beginning, but they vary at each resolution level for each different distribution of the cells in the level. Details of the reconstruction procedure, in all the cases considered in this work, are provided in the following section.

2.3.1 Further comments on the conservative reconstruction step

In this section, further details on the conservative reconstruction step are provided. The problem definition requires two steps: the identification of a proper stencil and the solution of the linear system obtaining the polynomial coefficients. Without loss of generality, the problem is presented as follows. Knowing the conditional expected value of a function $f(\boldsymbol{\xi})$ in a proper stencil $\mathcal{S}_j^k \subseteq \Xi_j^k$, at the kth resolution level, the polynomial $\mathcal{P}_j^k$ for which $\mathbb{E}\left(\mathcal{P}_j^k \,|\, \Xi_j\right) = \mathbb{E}\left(f(\boldsymbol{\xi}) \,|\, \Xi_j\right)$ is computed. A generic polynomial, obtained including all the contributions up to a rth order, contains $\frac{(r+d)!}{r!d!}$ coefficients, hence a stencil $\mathcal{S}_j^k$ with this cardinality should be chosen. In the case of the 1D stochastic space, see papers ***P4*** and ***P5***, the second order polynomial has been selected:

$$\mathcal{P}_j = a(\xi - \xi_j)^2 + b(\xi - \xi_j) + c, \tag{2.51}$$

where ξ_j is the coordinate of the center of the stochastic cell. In the case of centered reconstruction, the stencil is $\mathcal{S}_j^k = \{\Xi_{j-1}, \Xi_j, \Xi_{j+1}\}$. This kind of reconstruction is formally third order accurate in the stochastic space provided a smooth enough function. However, in the case of functions with high-gradients or discontinuities, to avoid an oscillatory behavior, the stencil needs to be biased avoiding the discontinuities region. Following [1, 2], such a stencil can be obtained selecting the one on which the polynomial $\mathcal{P}_j$ contains the lowest high order coefficient, *i.e.* $\min(|a|)$ in the (2.51). The problem reduces to collect a set of stencils on which the $\mathcal{P}_j$ must be recovered after selecting among the ones available. For the 1D case the choice is quite easy: the three candidate stencils are $\left\{\Xi^k_{j-2}, \Xi^k_{j-1}, \Xi^k_j\right\}$, $\left\{\Xi^k_{j-1}, \Xi^k_j, \Xi^k_{j+1}\right\}$ and $\left\{\Xi^k_j, \Xi^k_{j+1}, \Xi^k_{j+2}\right\}$. At the boundaries the collection of the admissible stencils needs to be properly reduced. In this work, to maintain the highest order of accuracy (at least for smooth problems), the stencil at the boundary is only biased. However, in principle, it is possible to reduce locally the order of interpolation when the number of cells available is reduced.

The coefficients for the polynomial $\mathcal{P}_j$ can be obtained solving the linear system obtained as

$$\mathbb{E}\left(\mathcal{P}^k_l \,|\, \Xi^k_l\right) = \mathbb{E}\left(f(\xi) \,|\, \Xi^k_l\right) = v^k_l \quad \text{for} \quad \Xi^k_l \in \mathcal{S}^k_j. \tag{2.52}$$

The implementation results very easy because

$$\mathbb{E}\left(\mathcal{P}^k_l \,|\, \Xi^k_l\right) = \mathbb{E}\left(a(\xi - \xi_j)^2 \,|\, \Xi^k_l\right) + \mathbb{E}\left(b((\xi - \xi_j)) \,|\, \Xi^k_l\right) + c \quad \text{for} \quad \Xi^k_l \in \mathcal{S}^k_j. \tag{2.53}$$

The resulting linear system $\mathtt{Ax} = \mathtt{b}$ contains a matrix $\mathtt{A}$ in which the elements are analytically known and are only function of the coordinate ξ_j and the boundaries of the cell Ξ^k_l. The structure of the inverse matrix $\mathtt{A}^{-1}$ is also analytically known as function of ξ_j. For each cell Ξ^k_j, fixing the coordinate ξ^k_j, the inverse matrix $\mathtt{A}^{-1}$ can be analytically evaluated as well as the vector of coefficients $\mathtt{x} = \{a, b, c\}^{\mathrm{T}}$.

The extension to multi-dimensional problems (2D and 3D cases) requires the definition of a proper stencil in a multi-dimensional space. In this work, only preliminary results are obtained in the multi-dimensional context and only for linear reconstruction $r = 1$. Even if the stochastic reconstruction should be pushed towards higher interpolations to gain in terms of compression capabilities of the MR scheme (and hence with an increase of computational efficiency), from an accuracy point-of-view the resulting scheme presented in Chapter 3 appears well balanced by achieving (formal) second order accuracy in space, time and stochastic space. Moreover, in the multi-dimensional case, the reconstruction is performed over a fixed stencil. This choice is motivated by the need to avoid difficult selection of stencils on non conformal meshes. In the multidimensional case two neighboring cells could be splitted or not independently; the situation in which the resulting mesh can be non conformal is not rare. However, considering that the polynomial reconstruction $\mathcal{P}^k_j$ is needed only to build the prediction operator P^k_{k-1}, it is easy to note that $\Xi^k_{jl} \in \Xi^{k-1}_j$ where $\Xi^{k-1}_j = \sum_{l=1}^{\bar{l}_c} \Xi^k_{jl}$. The $\bar{l}_c$ cells in which the mother cell Ξ^{k-1}_j is splitted are, for construction, conformal and on them a proper defined stencil can be defined. In the 2D and 3D cases, the linear reconstruction can be obtained providing a stencil with cardinality respectively of 3 and 4 cells. In figures 2.5 and 2.6, the stencils employed are represented, for all the possible positions of the cell Ξ^k_j, on which the polynomial $\mathcal{P}^k_j$ must be recovered.

The coefficients for the polynomial $\mathcal{P}^k_j$ are obtained by the same procedure of the 1D case. In particular, a system of three or four equations should be solved, respectively, for the 2D and 3D cases. The polynomials $\mathcal{P}^k_j$ for the 2D and 3D cases result

$$\mathcal{P}^k_j = \begin{cases} a(\xi_1 - \xi_{1,j}) + b(\xi_2 - \xi_{2,j}) + c & \text{if} \quad \boldsymbol{\xi} \in \mathbb{R}^2 \\ a(\xi_1 - \xi_{1,j}) + b(\xi_2 - \xi_{2,j}) + c(\xi_3 - \xi_{3,j}) + d & \text{if} \quad \boldsymbol{\xi} \in \mathbb{R}^3, \end{cases} \tag{2.54}$$

where the centroid of the cell Ξ^k_j is indicated as $(\xi_{1,j}, \xi_{2,j})$ and $(\xi_{1,j}, \xi_{2,j}, \xi_{3,j})$ in the 2D and 3D case, respectively. Of course, even in the multi-dimensional case, the system

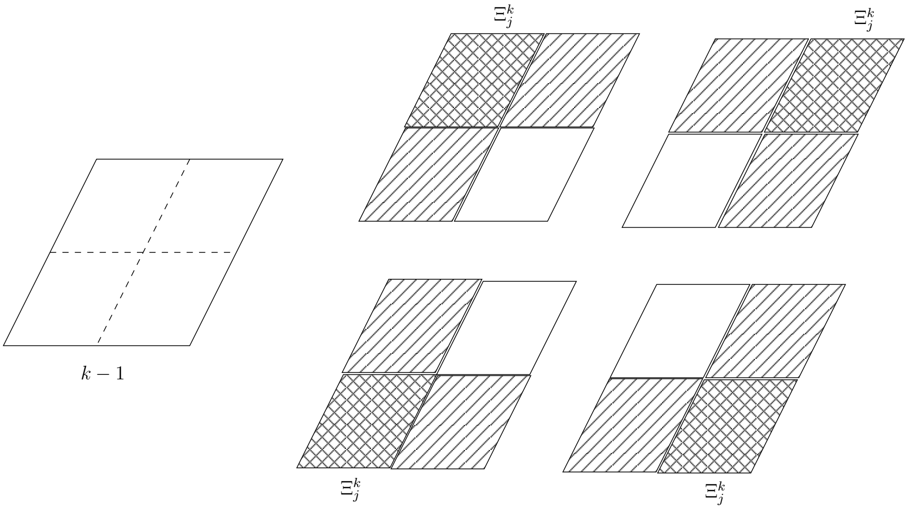


Figure 2.5: Stencil identification for the 2D stochastic linear reconstruction. On the left the *mother* cell at level $k-1$ is represented, while on the right the four possible positions for the cell Ξ_j^k and their stencil $\mathcal{S}_j^k$ are reported.

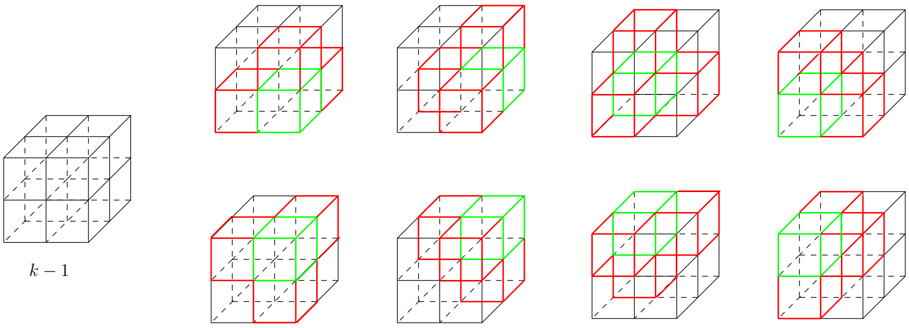


Figure 2.6: Stencil identification for the 3D stochastic linear reconstruction. On the left the *mother* cell at level $k-1$ is represented, while on the right the eight possible positions for the cell Ξ_j^k (in green) and their stencil $\mathcal{S}_j^k$ (red) are reported.

can be written as $\mathtt{A}\mathtt{x} = \mathtt{b}$ and the matrix $\mathtt{A}$ can be obtained analytically as function of the centroid coordinates (and the boundaries of Ξ_j^k) and the same holds for its inverse. Once identified the centroid coordinates of the cell Ξ_j^k, the vector of coefficients can be obtained analytically without matrix inversion. In this section, the procedure to select the stencil, for all the cases taken in consideration in this work, is presented. The polynomial reconstruction plays a fundamental role for both the MR framework, allowing to detect the regions in which a strong effort (more cells) is needed, and also in the SI scheme, where the reconstruction of the variables, explicitly known only in terms of conditional expectancies, over the stochastic space, is a mandatory step to compute the fluxes expectancies. In particular, the link between the MR and the SI, is presented in Chapter 3.

2.4 Closing remarks: choosing between point-value and the cell-average

In this chapter, the framework for a MR representation in the stochastic space, by means of both point-value or cell-average setting, has been presented. In papers ***P1***, ***P2*** and ***P3***, the point-value setting is used, while in papers ***P4*** and ***P5***, the cell-average setting constitutes the fundamental brick to design the aSI scheme. Even if only a rigorous comparison between the two approaches could provide some elements in order to evaluate the performances of the sTE and the aSI scheme, it is important to notice that the two approaches appear to be complementary.

Despite the virtually universal applicability of both the reconstruction techniques to any kind of numerical scheme, some further comments are necessary. Dealing with discrete data in many context is easier with point-value quantities. This is the case, for instance, of finite difference (FD) schemes or finite element (FE) methods (or even ODE). At the same time, even the discrete representation of a function is normally obtained by its point values. For instance, if a numerical model is available, the point values could be relatively easy to obtain, while the cell-average value could be known only after a numerical integration relying on (again) point-values. In this sense the sTE scheme appears to be the most general and natural one and the less intrusive between the two approaches. However, the introduction of a finite volume (FV) reconstruction in the seminal work of Abgrall [3], opened a new way of interpreting the representation of functions over the random space. Mimicking what normally is done in the deterministic context, the Abgrall's method, unifying the representation in the overall physical/stochastic space, should be viewed as a way to build well-balanced stochastic scheme. It is clear that the high-order representation of the data in the stochastic space cannot improve the overall accuracy of a numerical scheme in which the spatial accuracy is of a low order. This aspect has been pointed out, for instance, in the work [86] to comment the limit of their multiwavelet approach. Moreover, when dealing with adaptive strategies in the overall physical/stochastic space it is important to be able to preserve the coherence of the data. It is well known, in the deterministic context, the importance of the conservative reconstruction of the function for the flux evaluation, while an approach like the sTE algorithm *a priori* could generate non-conservative interpolations due to a Lagrange interpolation procedure. For this reason, the SI method appears to be more robust and potentially more efficient. The sTE scheme, despite its good results as a stand-alone scheme, is used in this work as a test ground to develop the new ideas and the algorithms. For instance, the TE algorithm for the cell-average setting is only an adaptation (improved) of the TE for the point-value setting. As a concluding remark, to resume the previous comments, the point-value setting remains very attractive in obtaining less intrusive schemes for applications where a function should be reproduced with the lowest number of evaluations (code running) or in problems where there is not a spatial dependence, as in the ordinary differential equations context (see for instance **P1** and also **P2**). On the contrary, in the stochastic partial differential context, the difference between the two approaches are not so evident. Further investigation to compare the two scheme are necessary. However, the possibility to represent, in a unified way and with a prescribed total accuracy, the solution of stochastic PDE makes the SI very attractive. For this reason, the MR cell-average framework is introduced in the SI scheme to obtain the overall aSI formulation presented in the next chapter. The

MR approach is useful to drive the refinement/coarsening of the random space, as a function of the evolution of the variables, permitting to represent the discrete data and to obtain conservative reconstructions.

CHAPTER 3

The adaptive semi-intrusive scheme

In this chapter, the SI scheme, briefly re-called in Chapter 1 with the MUSCL-Hancock method (MHM), is extended for including the MR representation of (discrete) data over the stochastic space. In this way, the aSI scheme is obtained. Two main elements should be clarified: the role of the discretization operator $\mathcal{D}_k$ and the link between the conservative reconstructions, performed in both the MR and SI. In particular the discretization operator $\mathcal{D}_k$ can be identified with the time-update step of the SI (see equation (1.36)), while the reconstruction operator can be chosen to be the same for MR and SI. The aim of the aSI scheme is to compute the conditional expectancies of the solution of partial differential equations. Let us suppose a tessellation of N_x cells $\mathcal{C}_i$ of the entire physical space $\Omega = \bigcup_{i=1}^{N_x} \mathcal{C}_i$ and a constant subdivision of the time line in N_t intervals of length $\Delta t = t_F/N_t$, $t_n = n\Delta t$. The overall aSI scheme is constituted by two external loops on time and physical space coordinates. For each couple of physical and time coordinates (x_i, t_n), the TE algorithm (in the cell average framework) is performed. At this level, the aSI scheme appears to be equivalent to the sTE algorithm (see the papers ***P2*** and ***P3***). However, the effect of the presence of the MR framework in the cell average setting is more evident analyzing the operation needed to perform the TE algorithm. In particular, the DAD algorithm presented in the Algorithm 3 is constituted, in the discretization step, by the time update step of the SI scheme (1.36). Moreover, the nested mesh sequence $\{\mathcal{G}^k\}_{k=0}^{L}$ will be dependent on the time t_n and also on the spatial coordinate x_i, instead of only depending from the resolution level k. In this sense, the aim is to obtain an adaptive refinement/coarsening of the stochastic space ables to capture the function regularity. For instance, the scheme applied to non-linear hyperbolic problems must be able to catch the development of a shock wave and to track it during the time evolution. Moreover, the scheme should be able to identify high-gradient or shock regions in the stochastic space even if they have no counterparts in the physical space. This property constitutes a strong difference with respect to the subcell resolution approach proposed by Witteven and Iaccarino in [91] where a shock sensor in the physical space is employed to reconstruct the position of discontinuities in the stochastic space. In the following Algorithm 5, the aSI scheme is presented, and the operations related directly to the SI scheme are reported in red.

Algorithm 5: Schematic presentation of the aSI algorithm

```
for $n = 1,\dots,N_t$ do
    for $i = 1,\dots,N_x$ do
        while $1 \leq k \leq L$ do
            for $\Xi_j^k(\mathbf{x}_i,t_n) \in \mathcal{G}^k(\mathbf{x}_i,t_n)$ do
                    DAD
                        Discretization
                        Agglomeration
                        Decimation
            end
                $\Longrightarrow v^{k-1}(\mathbf{x}_i,t_n), v^k(\mathbf{x}_i,t_n)$
            if $k \neq L$ then
                Encoding $(v^{k-1}(\mathbf{x}_i,t_n), v^k(\mathbf{x}_i,t_n)) \rightarrow d^k(\mathbf{x}_i,t_n)$
                Truncation $(d^k(\mathbf{x}_i,t_n), \varepsilon_k) \rightarrow \hat{d}^k(\mathbf{x}_i,t_n)$
                Level advancement $(\mathcal{G}^k(\mathbf{x}_i,t_n), \hat{d}^k(\mathbf{x}_i,t_n)) \rightarrow \mathcal{G}^{k+1}(\mathbf{x}_i,t_n)$
            end
        end
            $\Longrightarrow v_j^L(\mathbf{x}_i,t_n) \quad \forall \Xi_j^L(\mathbf{x}_i,t_n) \in \mathcal{G}^L(\mathbf{x}_i,t_n)$
        Statistics computation for $v_j^L(\mathbf{x}_i,t_n)$
    end
end
```

In the next, sequences of operations needed to obtain the Discretization §3.1 and the Level advancement §3.2 steps are reported. In particular, section §3.1 contains the details concerning the evaluation of the discrete data (the cell-averages) at a fixed physical/time coordinate $(\mathbf{x}_i, t_n)$ knowing the set of MR representations $v^L(\mathbf{x}_i, t_{n-1})$ of the solution at the previous time step t_{n-1} and for each physical coordinate $\mathbf{x}_i$ with $i = 1, \dots, N_x$. This problem concerns some operations between resolution levels at the same physical/time location and some operations between MR at different time and resolution levels. In this step, the conservative reconstruction procedure is also contained. The Level advancement step, instead, does not concern only the generation of a new resolution level knowing the actual resolution and the wavelet defined on it, as in the TE algorithm. On the contrary, it concerns also the distribution of the degrees-of-freedom of each cell between a level $k-1$ and the successive one k. A brief discussion of this issue is made in §3.2, while in §3.3 some comments on the statistics computations are reported. In the following section the discretization step is further analyzed.

3.1 From SI to the Discretization step of the aSI

In this section, the discretization step of the aSI scheme, Algorithm 5, is described in more details. In the general MR framework, the discretization is performed by means of the operator $\mathcal{D}_k$ described in the previous chapter. It is possible to note that the SI scheme naturally embeds a discretization operator: the conditional expected value operator $\mathbb{E}\left(\bullet \,|\, \Xi_j\right)$ (see equation (1.35)). However, recalling the equation (1.36) of the SI scheme, the conditional expected value of the (spatial) cell average at time $n+1$ is expressed as a function of quantities at time n. Making evident the dependence of the mesh from a specific spatial position, the ith cell $\mathcal{C}_i$, and the time step t_n, on a generic resolution level kth, the stochastic cell can be indicated as $\Xi_j^{k,n}$. The equation (1.36) can be recasted as follows

$$\begin{aligned}\mathbb{E}\left(\bar{u}_i^{n+1} \,|\, \Xi_j^{k,n+1}\right) = \mathbb{E}\left(\bar{u}_i^{n} \,|\, \Xi_j^{k,n+1}\right) - \frac{\Delta t}{|\mathcal{C}_i|}\Big(&\mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i-1_R}^{\Uparrow}, u_{i_L}^{\Uparrow}\right) \,|\, \Xi_j^{k,n+1}\right) \\ &- \mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i_R}^{\Uparrow}, u_{i+1_L}^{\Uparrow}\right) \,|\, \Xi_j^{k,n+1}\right)\Big).\end{aligned} \tag{3.1}$$

It is important to remark that each stochastic cell depends explicitly from the spatial coordinate x_i. For simplifying the notation, the index i is not explicitly reported as sub-index to the cell, because it is naturally evident considering the operator $\mathbb{E}\left(\bullet \,|\, \Xi_j\right)$. For instance, the term $\mathbb{E}\left(\bar{u}_i^{n+1} \,|\, \Xi_j^{k,n+1}\right)$ refers to the conditional expected value of the cell average $\bar{u}_i^{n+1}$ on the spatial cell $\mathcal{C}_i$ at time t_{n+1} over the stochastic cell $\Xi_j^k(\mathrm{x}_i, t_{n+1})$. The explicit dependence from the time cannot be eliminated from the notation of the stochastic cell because, as it will be more evident later, the time-advancement of the solution at the time step $n+1$ requires to update the conditional expected value, of the spatial cell-average, at time n, but computed on the stochastic cell at $n+1$. In principle, this stochastic cell $\Xi_j^k(\mathrm{x}_i, t_{n+1})$ could not exist at the final resolution level L at the same spatial location at the previous time step n. This case, for instance, commonly occurs when the low resolution levels should be computed at time $n+1$, while the solution at the previous time step has been obtained over a more refined stochastic space. The procedure to obtain, in an accurate way, the term $\mathbb{E}\left(\bar{u}_i^{n} \,|\, \Xi_j^{k,n+1}\right)$ is of a crucial importance and the problem will be addressed explicitly in the following. The time update (3.1) also requires the computation of the flux expectancies of the numerical fluxes over the interfaces. In this sense, the notation $\mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i-1_R}^{\Uparrow}, u_{i_L}^{\Uparrow}\right) \,|\, \Xi_j^{k,n+1}\right)$ indicates the expected value of the numerical (Riemann) fluxes over the jth stochastic cell at resolution level k and time $n+1$ between the spatial cells $\mathcal{C}_{i-1}$ and $\mathcal{C}_i$. It is important to note that, even if the SI scheme formulation is based on cell-average quantities over the combined physical/stochastic cells, the flux contribution is present only in the spatial direction, *i.e.* through the interfaces dividing the spatial cells. The flux contribution through the interfaces dividing two stochastic cells is always equal to zero because it does not exist. In figure 3.1, the typical pattern of the combined physical/stochastic cells in 1D spatial problem and 1D, 2D and 3D stochastic problems are reported.

The computation of both contributions, the cell-average value $\mathbb{E}\left(\bar{u}_i^{n} \,|\, \Xi_j^{k,n+1}\right)$ and the expectancies of the fluxes $\mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i-1_R}^{\Uparrow}, u_{i_L}^{\Uparrow}\right) \,|\, \Xi_j^{k,n+1}\right)$ and $\mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u_{i_R}^{\Uparrow}, u_{i+1_L}^{\Uparrow}\right) \,|\, \Xi_j^{k,n+1}\right)$, are closely connected to the quantities at the previous (coarser) resolution level. This aspect will be discussed in the next section and is determined by the quadrature rule used to compute the integral quantities. In this sense, the following exposition is carried out considering each resolution level as independent from the previous (coarser) ones at the same time step, while the efficient redistribution of the degrees-of-freedom is presented in §3.2.

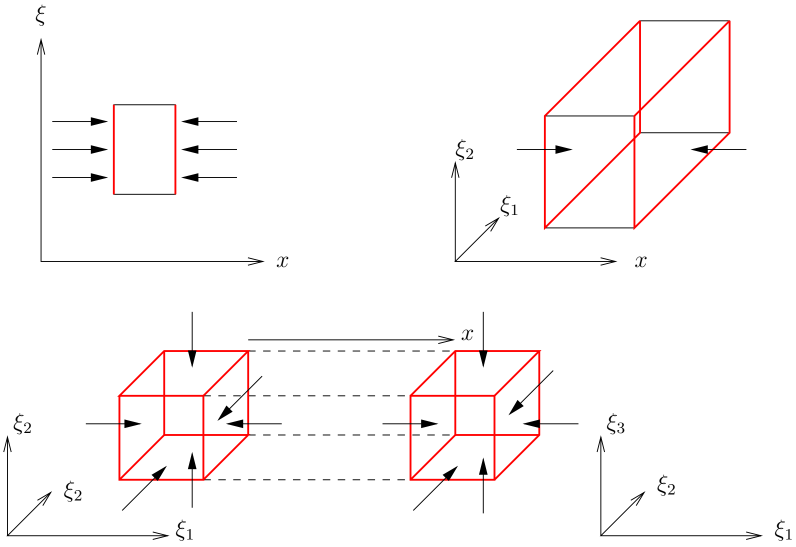


Figure 3.1: Schematic representation of the combined physical/stochastic cells for 1D-1D, 1D-2D and 1D-3D problems. The face through which a flux exists are colored in red.

The first contribution to compute is the conditional cell-average value $\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_j^{k,n+1}\right)$. The conditional cell-average value should be computed knowing a sequence of discrete data v^L on the tessellation of the stochastic space

$$\Xi = \bigcup_{j=1}^{N_\xi(\mathbf{x}_i,t_n)} \Xi_j^L(\mathbf{x}_i,t_n), \tag{3.2}$$

where the smaller (in a measure sense) cell is equal to the resolution level L and the dependence of the tessellation on the number of the elements N_ξ of the tessellation, from the physical and time coordinate, is explicitly reported. Two possibilities are of interest[1]:

$$\begin{aligned} &\Xi_{j_{n+1}}^{k,n+1}(\mathbf{x}_i) \subseteq \Xi_{j_n}^{L,n}(\mathbf{x}_i) \quad \text{or} \\ &\Xi_{j_{n+1}}^{k,n+1}(\mathbf{x}_i) = \bigcup_{l_n=1}^{\bar{l}_n} \Xi_{l_n}^{L,n}(\mathbf{x}_i). \end{aligned} \tag{3.3}$$

The reason is evident considering that each cell at the coarsest level can be divided at most until reaching the finest level L, but where, locally, the function is more regular even at a lower level ($k < L$). In any case, each divided cell is always formed by an entire number of cells. For this reason, if at the time step $n+1$ the function is more regular than at the previous time step, then a set of $\bar{l}_n$ cells, forming the cell $\Xi_{j_{n+1}}^{k,n+1}(\mathbf{x}_i)$ at time $n+1$, will exist. Otherwise, if a high-gradient or a shock region appears, the cell $\Xi_{j_{n+1}}^{k,n+1}(\mathbf{x}_i)$ will be entirely contained in a cell (to identify) $\Xi_{j_n}^{L,n}(\mathbf{x}_i)$. In this last case, the value can be obtained employing directly the predictor operator P_{k-1}^k; if a prediction operator is applied on the resolution level L, the cell $\Xi_{j_n}^{L,n}(\mathbf{x}_i)$ has

[1] It is important to note the introduction of the subscript j_n or j_{n+1} to make evident the change of numeration of the tessellation at the same spatial location $\mathbf{x}_i$ for different time steps. Hereafter this complete subscript will be introduced to avoid possible misunderstandings.

been generated at a lower resolution level and has not been divided any more arriving at the resolution level L carrying its original measure. The reconstruction operator $\mathcal{R}_k$ on the tessellation at time n, obtained at resolution level L, can be discretized on the cell $\Xi_{j_{n+1}}^{k,n+1}(\mathrm{x}_i)$ to obtain the prediction

$$\begin{aligned}\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{j_{n+1}}^{k,n+1}\right) &= \frac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\int_{\Xi_{j_{n+1}}^{k,n+1}} \bar{u}_i^n(\boldsymbol{\xi})\mathrm{d}\mu(\Xi_{j_{n+1}}^{k,n+1}) \\ &\simeq \frac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\int_{\Xi_{j_{n+1}}^{k,n+1}} \left(\mathcal{R}_{L,n}v^{L,n}\right)_{j_{n+1}} \mathrm{d}\mu(\Xi_{j_{n+1}}^{k,n+1}).\end{aligned} \tag{3.4}$$

The other situation that can occur is the existence of an entire number $\bar{l}_n$ of cells, at time n, forming $\Xi_{j_{n+1}}^{k,n+1}(\mathrm{x}_i)$ at time $n+1$. This is always the case when, for instance, the TE algorithm starts at the new time step $n+1$ from the lowest resolution level. In this case, the conditional expected value can be obtained by assembling the conditional expectancies of each cell at time n thanks to the additivity of integrals as follows

$$\begin{aligned}\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{j_{n+1}}^{k,n+1}\right) &= \frac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\sum_{l_n=1}^{\bar{l}_n}\int_{\Xi_{l_n}^{L,n}(\mathrm{x}_i)} \bar{u}_i^n(\boldsymbol{\xi})\mathrm{d}\mu(\Xi_{l_n}^{L,n}(\mathrm{x}_i)) \\ &= \frac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\sum_{l_n=1}^{\bar{l}_n} \mu(\Xi_{l_n}^{L,n}(\mathrm{x}_i))\,\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{l_n}^{L,n}(\mathrm{x}_i)\right).\end{aligned} \tag{3.5}$$

The set of values $\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{l_n}^{L,n}(\mathrm{x}_i)\right)$ is, of course, already available belonging to $v^{L,n}$. Resuming, the value of $\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{j_{n+1}}^{k,n+1}\right)$ can be computed as

$$\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{j_{n+1}}^{k,n+1}\right) = \begin{cases} \dfrac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\displaystyle\int_{\Xi_{j_{n+1}}^{k,n+1}} \left(\mathcal{R}_{L,n}v^{L,n}\right)_{j_{n+1}} \mathrm{d}\mu(\Xi_{j_{n+1}}^{k,n+1}) & \text{if} \quad \Xi_{j_{n+1}}^{k,n+1}(\mathrm{x}_i) \subseteq \Xi_{j_n}^{L,n}(\mathrm{x}_i) \\ \dfrac{1}{\mu(\Xi_{j_{n+1}}^{k,n+1})}\displaystyle\sum_{l_n=1}^{\bar{l}_n} \mu(\Xi_{l_n}^{L,n}(\mathrm{x}_i))\,\mathbb{E}\left(\bar{u}_i^n \,|\, \Xi_{l_n}^{L,n}(\mathrm{x}_i)\right) & \text{if} \quad \Xi_{j_{n+1}}^{k,n+1}(\mathrm{x}_i) = \displaystyle\bigcup_{l_n=1}^{\bar{l}_n} \Xi_{l_n}^{L,n}(\mathrm{x}_i). \end{cases} \tag{3.6}$$

The time update (3.1) requires also the computation of the expectancies of the fluxes. The evaluation of these contributions appears to be much more complex and intimately related to the deterministic scheme with respect to the previous contribution (3.6). The evaluation of the integral quantities cannot be performed without considering a proper quadrature formula. In this section, a generic set of quadrature points (on each cell) $\{\boldsymbol{\xi}_{\mathrm{ng}}\}_{q=1}^{\mathrm{Ng}}$ and a proper set of weight $\{w_{\mathrm{ng}}\}_{q=1}^{\mathrm{Ng}}$ are considered. In the next section, the choice of the quadrature rule will be discussed for obtaining, in an efficient way, the nodal value from the previous resolution level. Fixing a spatial cell x_i, the aim is to compute the expected value of the flux contributions at time n, at a generic local resolution level k. For simplicity of exposure, as already made in §1.3.2, the following exposition is made for a 1D conservation law. The sequence of operations, to compute the flux expected values, is reported in the Algorithm 6.

Algorithm 6: Computation of the flux expected values in the aSI scheme.

```
for $\mathrm{ng} = 1, \dots, \mathbb{N}\mathrm{g}$ do
    • Physical Vector assembling $\mathbb{PV}(\boldsymbol{\xi}_{\mathrm{ng}}) = \left\{\bar{u}^n_{i-2}(\boldsymbol{\xi}_{\mathrm{ng}}), \bar{u}^n_{i-1}(\boldsymbol{\xi}_{\mathrm{ng}}), \dots, \bar{u}^n_{i+2}(\boldsymbol{\xi}_{\mathrm{ng}})\right\}$:
        $\bar{u}^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) = \left(\mathcal{R}^n_L(x_\ell) v^{L,n}(x_\ell)\right)(\boldsymbol{\xi}_{\mathrm{ng}}) \quad \text{with} \quad \ell \in \{i-2, i-1, \dots, i+2\}$
    • Imposition of the boundary conditions (if $x_i = \{x_1, x_2, x_{N_x-1}, x_{N_x}\}$)
    • Slope computations (and limiting) $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}\}$:
        $\sigma^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) = \sigma\left(\bar{u}^n_{\ell-1}(\boldsymbol{\xi}_{\mathrm{ng}}), \bar{u}^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}), \bar{u}^n_{\ell+1}(\boldsymbol{\xi}_{\mathrm{ng}})\right)$
    • Extrapolation $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}\}$ (STEP 1 - MHM):
        $\begin{cases} u^n_{\ell_L}(\boldsymbol{\xi}_{\mathrm{ng}}) = \bar{u}^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) - \sigma^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) \dfrac{|\mathcal{C}_\ell|}{2} \\ u^n_{\ell_R}(\boldsymbol{\xi}_{\mathrm{ng}}) = \bar{u}^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) + \sigma^n_\ell(\boldsymbol{\xi}_{\mathrm{ng}}) \dfrac{|\mathcal{C}_\ell|}{2} \end{cases}$
    • Semi-time step evolution $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}\}$ (STEP 2 - MHM)[a]:
        $\begin{cases} u^{\Uparrow}_{\ell_R}(\boldsymbol{\xi}_{\mathrm{ng}}) = u^n_{\ell_R}(\boldsymbol{\xi}_{\mathrm{ng}}) + \dfrac{1}{2}\dfrac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u^n_{\ell_L}(\boldsymbol{\xi}_{\mathrm{ng}}), \boldsymbol{\xi}_{\mathrm{ng}}) - f(u^n_{\ell_R}(\boldsymbol{\xi}_{\mathrm{ng}}), \boldsymbol{\xi}_{\mathrm{ng}})\right) \\ u^{\Uparrow}_{\ell_L}(\boldsymbol{\xi}_{\mathrm{ng}}) = u^n_{\ell_L}(\boldsymbol{\xi}_{\mathrm{ng}}) + \dfrac{1}{2}\dfrac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u^n_{\ell_L}(\boldsymbol{\xi}_{\mathrm{ng}}), \boldsymbol{\xi}_{\mathrm{ng}}) - f(u^n_{\ell_R}(\boldsymbol{\xi}_{\mathrm{ng}}), \boldsymbol{\xi}_{\mathrm{ng}})\right) \end{cases}$
end
Flux Quadrature:
    $\begin{cases} \mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right) \mid \Xi^{k,n+1}_{j_{n+1}}\right) \simeq \sum_{\mathrm{ng}=1}^{\mathbb{N}\mathrm{g}} w_{ng}\, \mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}(\boldsymbol{\xi}_{\mathrm{ng}}), u^{\Uparrow}_{i_L}(\boldsymbol{\xi}_{\mathrm{ng}})\right) \\ \mathbb{E}\left(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right) \mid \Xi^{k,n+1}_{j_{n+1}}\right) \simeq \sum_{\mathrm{ng}=1}^{\mathbb{N}\mathrm{g}} w_{ng}\, \mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}(\boldsymbol{\xi}_{\mathrm{ng}}), u^{\Uparrow}_{i+1_L}(\boldsymbol{\xi}_{\mathrm{ng}})\right) \end{cases}$
```

When the expected values of the fluxes are computed and the conditional expected value $\mathbb{E}\left(\bar{u}^n_i \mid \Xi^{k,n+1}_{j_{n+1}}\right)$ is evaluated via the equation (3.6), the time update can be obtained by the equation (3.1). In the previous Algorithm 6, the numerical flux has been indicated as $\mathcal{F}^{\mathrm{RM}}$ referring the flux obtained by a Riemann solver, exact or approximated, as made in paper ***P4***. However, all the other techniques developed in the deterministic context can be applied. In this sense, the aSI scheme shows a flexibility equivalent to a *non-intrusive* technique. Moreover, in paper ***P5***, the aSI scheme is formulated for a classical predictor-corrector MUSCL approach and a more complex flux function is obtained by the Discrete Equation Method for multiphase flows. The procedure just described constitutes the discretization step of the Algorithm 5, while the level advancement step will be discussed in the next section.

[a] The flux function can depends separately from the vector of the random parameters and from the unknown, as in the linear advection equation solved in the Chapter 4, where a random advection velocity is present.

3.2 Level advancement: the role of the quadrature formula

In this section, the level advancement step of the Algorithm 5 is analyzed more in details. The TE algorithm, in both point-value and cell-averages frameworks, is based on the comparison between the local coordinates of the error vector e^k at a resolution level k with the threshold ε_k. In the cell-average framework, if the local error is greater than the local threshold, then the cell is divided into a number of $\bar{l}_c$ cells. In this work, the number of $\bar{l}_c$ has been fixed to 2^d, where d is the number of the stochastic dimensions. In the SI scheme, as already shown in the previous section, the discretization operator on a generic cell Ξ_j^k is exactly equal to the conditional expected value operator applied on the same stochastic cell. The set of all the conditional expected values, over all the cells belonging to the entire stochastic space Ξ, is constituted by the discrete values indicated with v^k in Chapter 2. However, the computation of the discrete data at a time step $n+1$ depends also on some integral quantities that are not available: the flux expected values (described in previous section). These quantities are closely related to the quadrature rule. In the seminal work of Abgrall [3], a Gauss quadrature rule with two nodes has been proposed. The 1D Gauss quadrature with two points for each cell can integrate exactly a polynomial function of third order. However, this kind of choice does not appear convenient to build efficient aSI scheme. Let us suppose to have a generic cell defined on the interval $[a,b]$. The quadrature points will be $\xi_{\mathrm{ng}} = \frac{a+b}{2} \pm \frac{\sqrt{3}}{2}\frac{b-a}{2}$. It is easy to verify that if the cell is divided in two cells, defined on $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$, the Gauss quadrature points of the two new cells will be, for both the two couples, different from the quadrature points of the mother cell. From a computational point-of-view, it is evident that a nested sequence of quadrature points would be preferable. For instance, if a trapezoidal rule is applied, *i.e.* the nodes are the two points a and b, they are also quadrature points for the children cells. In such a case, the splitting of a cell, from a quadrature points, is equivalent to add another cell. In general, from a computational point-of-view related to the degrees-of-freedoms belonging to each cell, a sequence of subdivisions of the cells has a cost that is equivalent to the number of cells obtained at the finest resolution level. Then, when the TE algorithm is applied to obtain the (optimal) non-uniform distribution of cells in the stochastic space, it has required a number of degrees-of-freedom exactly equal to an uniform distribution of the same number of cells.

The quadrature rules in the aSI scheme cover an important role to compute the flux expected values as shown in the previous section. The quadrature rule are applied in a space of 2^d dimensions, for each interface of each physical cell (see also the figure 3.1). For this reason, the quadrature rules have to be designed to be nested and easily extensible in many dimensions. In this context, as a first attempt, the family of Newton-Cotes rules could be a proper choice [71]. These *formulæ* are based on Lagrange interpolation with equally spaced nodes in $[a,b]$. Examples of Newton-Cotes *formulæ* are the midpoint, trapezoidal and the Cavalieri-Simpson rules with, respectively, one, two and three points. A significant properties of the Newton-Cotes rules is that the weights can be computed one time for all fixing the number of points. Moreover, considering a rule with $\mathrm{N\!g} \geq 2$ nodes, the weights w_{ng} and the nodes $\xi_{\mathrm{ng}} = a + \frac{b-a}{(\mathrm{N\!g}-1)}(\mathrm{ng}-1)$ are symmetric with respect to the interval $[a,b]$. In this work, the case of closed rules is of interest, in which both a and b belongs to the the set of quadrature nodes ξ_{ng}. In the table 3.1, the first four quadrature *formulæ* of the

Newton-Cotes family are reported.

$w_{\text{ng}} \setminus$ Ng	2	3	4	5
w_1	1/2	1/6	1/8	7/90
w_2	1/2	2/3	3/8	32/90
w_3	-	1/6	3/8	12/90
w_4	-	-	1/8	32/90
w_5	-	-	-	7/90

Table 3.1: First four closed *formulæ* of the Newton-Cotes family with Ng nodes $\xi_{\text{ng}} = a + \frac{b-a}{(\text{Ng}-1)}(\text{ng}-1)$.

Finally, to integrate a function $f(\xi)$ on the interval $[a, b]$, the Newton-Cotes rules are of the form

$$\int_a^b f(\xi)\mathrm{d}\xi \simeq (b-a) \sum_{\text{ng}=1}^{\text{Ng}} w_{\text{ng}} f(\xi_{\text{ng}}), \tag{3.7}$$

where weights are reported in the table 3.1. The degree of exactness m of a Newton-Cotes rules rule with Ng nodes is

$$m = \begin{cases} \text{Ng} - 1 & \text{if} \quad \text{Ng is even} \\ \text{Ng} & \text{if} \quad \text{Ng is odd.} \end{cases} \tag{3.8}$$

The extension to multidimensional space can be obtained by tensorization of the previous relation (3.7). If each rule applied over the ith generic direction and the approximated integral is indicated as I^i, over the hypercube $[a, b]^d$ of dimension d, it follows that

$$\int_{[a,b]^d} f(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \simeq I^1 \otimes \cdots \otimes I^d = (b-a)^d \sum_{\text{ng}_1=1}^{\text{Ng}} \cdots \sum_{\text{ng}_d=1}^{\text{Ng}} w_{\text{ng}_1} \cdots w_{\text{ng}_d} f(\xi_1, \ldots, \xi_d). \tag{3.9}$$

This tensorization approach in general is not efficient because the number of the nodes exhibits an exponential growth with the dimensional increase, *i.e.* the so-called *curse of dimensionality*. However, in the context of the present work, the overall number of cells is strongly limited by the aSI scheme itself. In this work, the aSI scheme is designed employing a number of nodes (for each dimension) equal to $\text{Ng} = 3$. Finally, in figure 3.2, the redistribution of the nodes of each cell at the resolution level $k-1$ is represented. In particular, the full dots indicates the nodes in which the functional evaluations have to be performed, while the circles indicate the nodes in which the functional evaluations have been already performed for the cell at level $k-1$ and the corresponding values are already available.

The so-called reallocation of the nodes, from a cell at level $k-1$ to the others at level k, has the direct consequence to reduce the loop over the quadrature nodes Ng (in the Algorithm 6 for the computation of the flux expected values) to only the unknown, *i.e.* not yet evaluated, quadrature points. The number of the unknown points varies from the position of the child cell in the original cell, at resolution level k; the policy of redistribution is presented in figure 3.2.

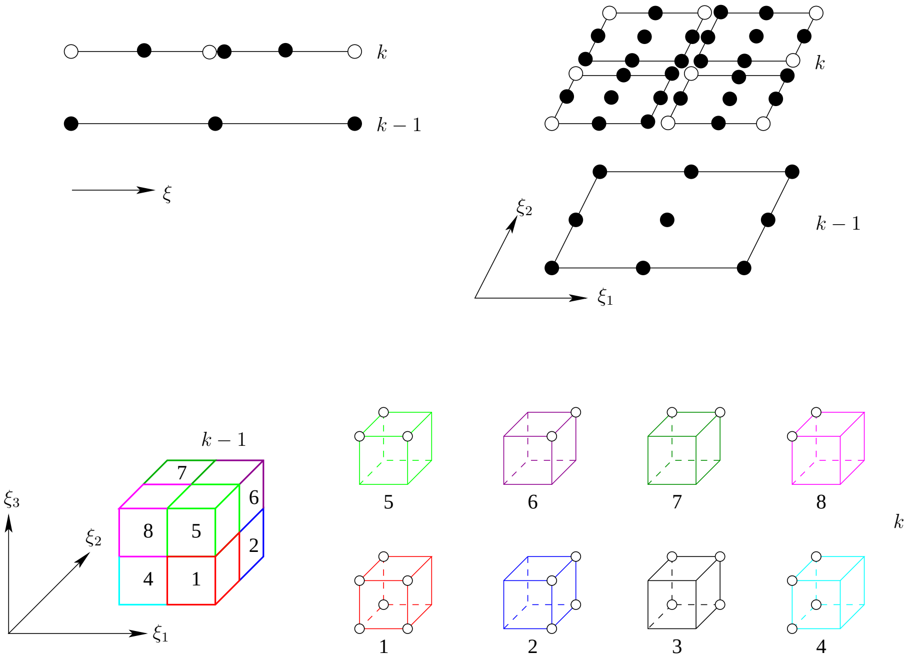


Figure 3.2: Repartition of the degrees-of-freedoms for the stochastic cells in the aSI scheme in 1D, 2D and 3D stochastic spaces. The full dots indicate the quadrature points to compute, while the white circles represent the quadrature points obtained by the mother cell at level $k-1$.

3.3 Statistics computations in the aSI scheme

In this section, some further comments on the computation of statistics are reported. In the last step of the internal loop of the aSI Algorithm 5, the statistics are computed for the discrete solution over the entire stochastic space. At the end of the internal loop, *i.e.* the application of the TE algorithm at each spatial location at a fixed time step, the result is the discrete solution $v^L(\mathrm{x}_i, t_n)$ over the tessellation $\Xi_j^L(\mathrm{x}_i, t_n)$. The final tessellation contains cells with different measures, because some among them derive directly from coarse level, being the local error under the threshold, while others are obtained at the resolution level L. One of the aim of the UQ analysis is to compute the quadrature of the solution over the stochastic space. Obviously, the quadrature rules described in the previous section can be applied knowing the continuous representation of the function $u(\xi) \in \mathcal{F}$. The continuous representation can be obtained applying the reconstruction operator $\mathcal{R}_L$ to the discrete data v^L. As already described in Chapter 2, in the Harten MR framework and hence in the approach developed in this work, the reconstruction operator is obtained locally by an eventually non-linear reconstruction, *i.e.* a local polynomial interpolant $\mathcal{P}_j$. This local reconstruction, as already shown, has a key role in the MR itself to obtain the predictor operator P_{k-1}^k. Moreover, in the aSI scheme, it has a role in order to obtain the conservative reconstruction useful to the computation of the fluxes and their expected values. The local conservative polynomial reconstruction can be employed also to compute statistics. In particular, at the last level of the TE algorithm, the reconstruction operator $\mathcal{R}_L$ is explicitly computed and the local polynomial is stored in each cell: obviously many of the local polynomial interpolants have already been computed to obtain the predictor operator P_{k-1}^{k} and hence stored in the corresponding cell. The statistics, for each spatial coordinate x_i and time step t_n, are then computed injecting the continuous reconstructed solution $\bar{u}_i^n(\boldsymbol{\xi}) \simeq \mathcal{R}_L^n(\mathrm{x}_i)v^{L,n}(\mathrm{x}_i)$, on the tessellation obtained at level L, into the integrals as follows in the case of the expected value and the variance

$$\begin{aligned}\mathbb{E}(\bar{u}_i^n) &= \int_{\Xi} \bar{u}_i^n(\boldsymbol{\xi})\mathrm{d}\mu(\boldsymbol{\xi}) \simeq \int_{\Xi} \left(\mathcal{R}_L^n(\mathrm{x}_i)v^{L,n}(\mathrm{x}_i)\right)\mathrm{d}\mu(\boldsymbol{\xi}) \\ \mathrm{Var}(\bar{u}_i^n) &= \int_{\Xi} \left(\bar{u}_i^n(\boldsymbol{\xi}) - \mathbb{E}(\bar{u}_i^n)\right)^2 \mathrm{d}\mu(\boldsymbol{\xi}) \simeq \int_{\Xi} \left(\mathcal{R}_L^n(\mathrm{x}_i)v^{L,n}(\mathrm{x}_i)\right)^2 \mathrm{d}\mu(\boldsymbol{\xi}) - \mathbb{E}^2(\bar{u}_i^n).\end{aligned} \tag{3.10}$$

Moreover the reconstruction operator will be employed at the successive time step in the Algorithm 6, in the Physical Assembling $\mathbb{P}\mathbb{V}$ step, to obtain the values of the physical cell average. Finally the integral quantities can be computed employing a quadrature rule or even analytically knowing the expression for the polynomial. For instance in papers ***P4*** and ***P5***, the Newton-Cotes rules with $\mathbb{N}\mathrm{g} = 5$, the so-called Boole's rule, has been applied to compute the variance, while in the multidimensional cases (reported in Chapter 4) the integration is performed analytically.

CHAPTER 4

Numerical results

In this chapter, some numerical results are presented. It is important to remark here that the complete set of numerical results can be found in the attached journal papers. In this chapter, only the unpublished results are presented. In particular, in Section §4.1, numerical results concerning the point-value framework with WENO interpolation are presented: the scheme is applied to steady functions in §4.1.1 and to a system of non-linear elastic wave propagation in a heterogeneous media in §4.1.2. The aSI scheme, obtained by the cell-average framework, is presented in §4.2 for 2D/3D results, while the 1D is discussed in paper ***P4***. Some steady functions are presented in section §4.2.1 for both 2D and 3D cases to demonstrate the convergence properties of the TE algorithm for cell-average discrete data. Three classical CFD test cases are also presented in §4.2.2 for a 2D stochastic space: the linear advection problems, with both continuous and discontinuous initial conditions, the Burgers equation and the Euler system of the gasdynamics equations. Finally, the 3D case for the linear advection equation is presented in §4.2.3.

4.1 Point-value TE/sTE scheme

In this section, the results obtained in the point-value setting are presented. First, the analysis of some steady problems, *i.e.* function depending only from the random variable ξ, is presented in terms of convergence properties. Different reconstruction operators are also described highlighting the importance of non-linear operators.

4.1.1 Steady-functions

In this section, a steady problem, in which a function $f = f(\xi)$ with $\xi \in \Xi \subset \mathbb{R}$ describes the stochastic output, is presented. The system is affected by an uncertain parameter ξ with distribution $p(\xi)$ here assumed uniform, *i.e.* $\xi \sim \mathcal{U}[0,1]$. The model function is discontinuous with the following equation

$$f(\xi) = \begin{cases} sin(2\xi^2\pi) & \text{if} \quad \xi \leq 11/20 \\ sin(2\xi^2\pi) + 1 & \text{otherwise.} \end{cases} \tag{4.1}$$

In Figure 4.1, the function defined in (4.1) is represented. Hereafter the resolution levels are designed indicating the index $m \in \mathbb{N}$ where, the coarse level contains 2^{m_0} intervals, while the finest one contains 2^{m_L} intervals. The TE algorithm has been applied with the following parameters: coarser level $m_0 = 3$, finest level m_L ranging between 6 and 20 and a threshold equal to $\varepsilon = 10^{-1}$.

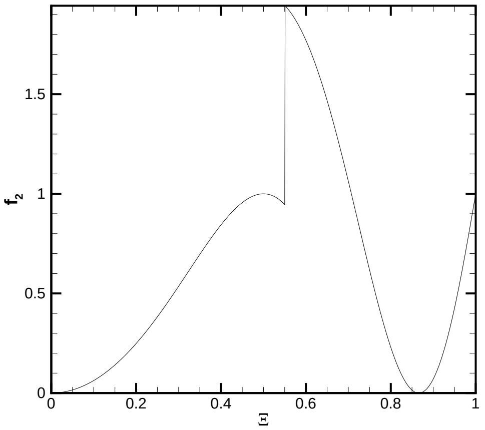


Figure 4.1: Representation of the steady functions $f = f(\xi)$ defined in the equation (4.1).

In the table 4.1, the results related to the application of the point-value TE Algorithm, 1, are reported. Different reconstruction operators are employed: the linear reconstruction ($r = 1$), the centered cubic polynomial reconstruction ($r = 3$), the ENO selection of the less oscillatory cubic polynomial ($r = 3$) and the WENO interpolation based on the cubic polynomials. The advantages of employing an high-order reconstruction operator is evident, considering the high compression ratios obtained, both μ and τ, for the cubic-based interpolation with respect to the linear one. Moreover, the non-linear, *i.e.* data-dependent interpolations, are clearly superior both in terms of compression and errors in L^1 and L^∞ norms. The two ratios of compression μ and of evaluations τ, defined in (2.20) and (2.21), are also reported. Knowing the analytical description of the function, the norms in L^1 and L^∞ can be measured as follows

$$\begin{cases} \mathrm{err}_{L^1} = ||v^L - \hat{v}^L||_{L^1} & = & \dfrac{1}{N_L + 1} \displaystyle\sum_{j=0}^{N_L} |v_j^L - \hat{v}_j^L| \\ \mathrm{err}_{L^\infty} = ||v^L - \hat{v}^L||_{L^\infty} & = & \max\limits_j |v_j^L - \hat{v}_j^L|, \end{cases} \tag{4.2}$$

in which v^L represents the function discretized on the finest level and $\hat{v}^L$ its counterpart obtained by the application of the TE algorithm.

To make evident the reason of the superiority of the data-dependent interpolation techniques the patterns, of the non-null wavelets, *i.e.* the points in which the discretization operator $\mathcal{D}_k$ is applied, are reported in the figure 4.2. The parameters are $m_0 = 3$, $m_L = 13$ with a threshold equal to $\varepsilon = 10^{-1}$. In the first row the comparison between the linear interpolation 4.2(a) and the high-order one 4.2(b) is evident: the linear scheme need the entire $k = 3$ resolution level and the most part of the level $k = 4$, while in the smooth regions the interpolation with $r = 3$ achieve the required accuracy with a lower resolution level. However, in correspondence of the discontinuity, the number of activated wavelets is greater in the high-order scheme than in the linear one. This is due to the degradation of the interpolation in all the intervals, belonging to the stencil, when the discontinuity is present in one of them. In the case of linear interpolation, the stencil is formed by only one cell and the number of activated wavelets is only two. In the case of $r = 3$ the stencil is formed by three intervals (with four point[1]); when the discontinuity is present, in at least one of them, two new points are generated for each interval. This produces a total of six new wavelets reducing the compression with respect the linear interpolation. This behavior is well known in the MR context and it is called *pollutted* region (see for instance [16]). The non-linear multiresolution schemes aims to avoid the generation of the pollutted regions: if the discontinuity is located, the stencil can be biased and the interpolation would be degraded only in the interval containing the discontinuity, but not in the others. The result is quite evident comparing the centered interpolation 4.2(b) with the ENO interpolation 4.2(c). The ENO selection of the stencil is able to locate, and avoiding, the discontinuity with the direct result to eliminate the *pollutted* region. However, the ENO selection of the stencil is well-known in literature to be very dependent to the round-off errors in the computation of the divided differences useful to identify the smoothest stencil (see the Algorithm 2). The drawback is evident in the smooth regions where, due to the round-off error, a biased stencil is selected instead of the centered one. As demonstrated in the paper ***P4***, the interpolation error is the smaller possible for centered stencils (or the one biased by one interval for r even), for smooth function. For instance, at the resolution level $k = 3$, at the left side of the discontinuity, the ENO interpolation 4.2(c) performs worst than the centered one 4.2(b). To recover the best possible interpolation, the WENO interpolation can be introduced. The pattern of the activated wavelets is presented in 4.2(d) and the direct comparison, with both the centers interpolation 4.2(b) and the ENO one 4.2(c), demonstrates that the WENO interpolation combine the advantage of both the strategies (high-order interpolation and non-linear stencil selection).

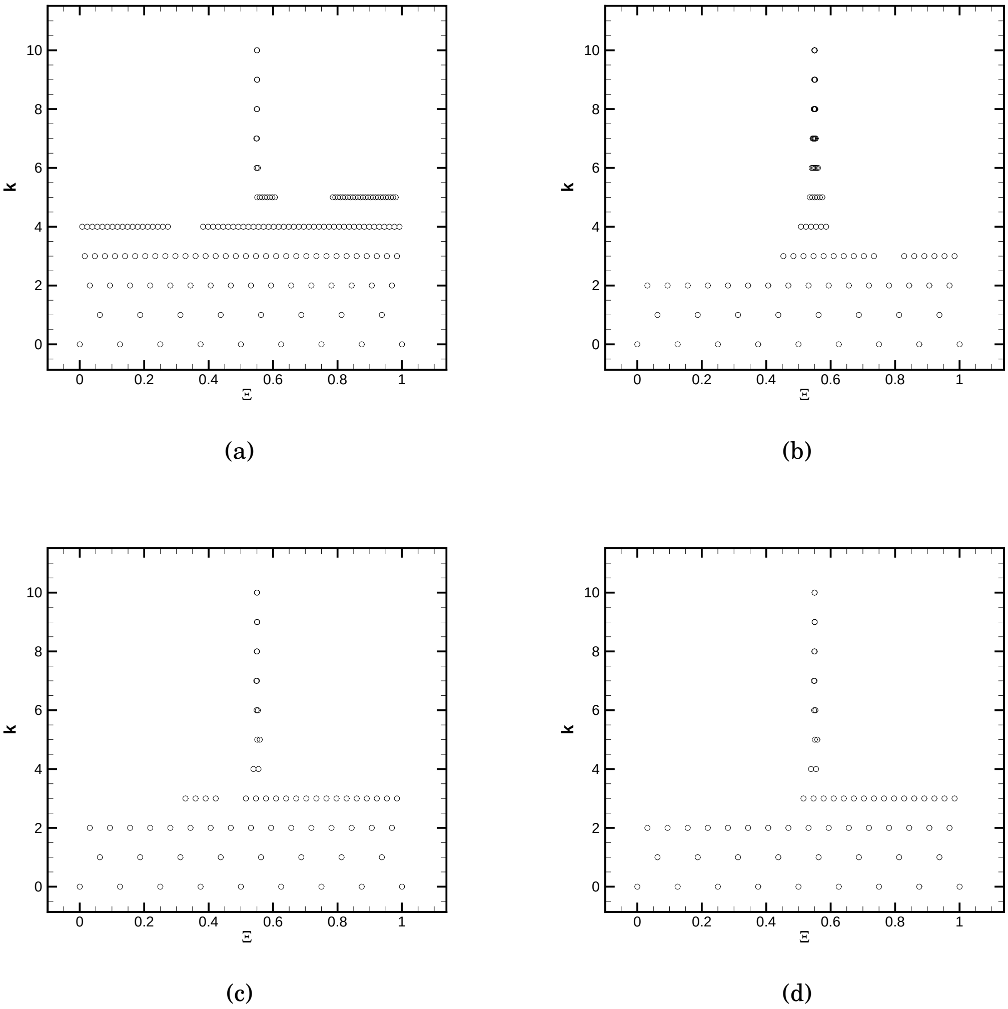
In a UQ perspective, the interest is not only to obtain a compressed representation of the function at a fixed time, but to have the possibility to compute statistics in a more efficient way, *i.e.* with the lower possible number of simulations for a prescribed accuracy. Results in terms of the error for computing the expected value and variance, with respect to the analytical solution, are reported in figure 4.3 (as a function of the number of applications of the discretization operator $\mathcal{D}_k$)

$$\begin{cases} \mathrm{err}_{\mathbb{E}} = \dfrac{\mathbb{E} - \mathbb{E}_{\mathrm{ex}}}{\mathbb{E}_{\mathrm{ex}}} \\ \mathrm{err}_{\mathrm{Var}} = \dfrac{\mathrm{Var} - \mathrm{Var}_{\mathrm{ex}}}{\mathrm{Var}_{\mathrm{ex}}}. \end{cases} \tag{4.3}$$

The errors are reported for the TE algorithm with $r = 1$, $r = 3$ (with and without the ENO selection of the stencil and for the WENO reconstruction). Moreover, as comparison, also a non-intrusive Polynomial Chaos (PC), with a number of simulations between 6 and 1041 (with steps of 15 simulations), and a quasi-Monte Carlo method, with Sobol sequences and simulation ranging between 10 and 1050 (with steps of 20 simulations), have been employed.

The TE algorithm achieves the best efficiency in term of reached accuracy, with a prescribed number of simulations, *i.e.* with a fixed number of exact evaluations of the model via the discretization operator $\mathcal{D}_k$, with respect to both the MC and the PC. The introduction of the high-order interpolation ($r = 3$) for the reconstruction operator increases the performances, for both the computation of expected value $\mathbb{E}$ and variance Var, with respect to the linear ($r = 1$) scheme. Moreover, the non-linear re-

[1] See the paper ***P3*** for further details on the cubic interpolation.



(a)

(b)

(c)

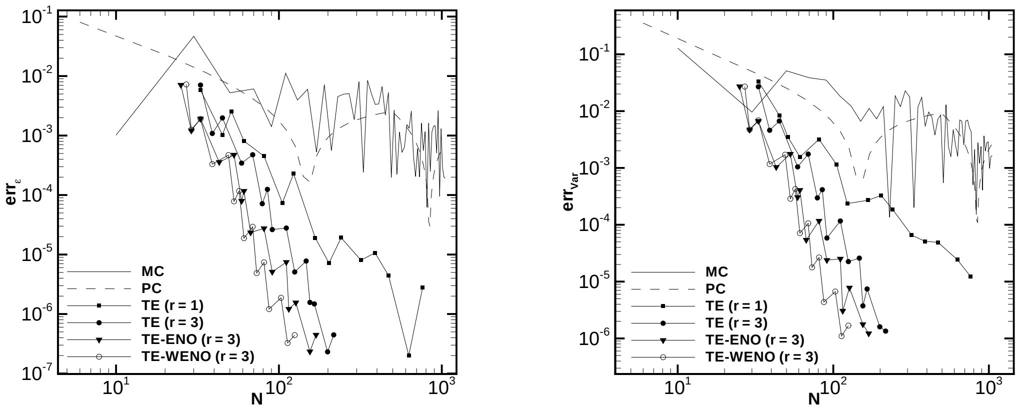
(d)

Figure 4.2: Distribution of the evaluated points for the steady function (4.1) with $m_L = 13$, $m_0 = 3$ and $\varepsilon = 10^{-1}$ for the linear reconstruction (a), the cubic centered reconstruction (b), with the ENO selection (c) and with WENO interpolation (d).

construction techniques allows to reach a stronger compression then a greater overall efficiency.

The TE algorithm can be employed as the basis to build an efficient algorithm solving also stochastic ordinary differential equations. In the paper ***P1***, some examples of stochastic ODEs, even with discontinuous and time dependent pdfs, have been reported. Other examples, of stochastic ODEs, are reported also in the paper ***P2*** and ***P3*** with, respectively, linear and high-order (non-linear) reconstructions.

In the next section, instead, the solution of the non-linear system of partial differential equations governing the propagation of elastic waves in heterogeneous media, in presence of random input, will be presented.



(a) (b)

Figure 4.3: Statistical errors for the expected value (a) and variance (b) following the definitions (4.3).

4.1.2 Non-linear elastic wave propagation in heterogeneous media

In this section the non-linear equations, governing the elastic wave propagation in a heterogeneous media in the Lagrangian frame, are solved in presence of uncertainty in the constitutive laws of the material. The propagation of compressional waves, in a one dimensional rod with density $\rho(x) > 0$ and a stress-strain relation $\sigma(\epsilon, x)$ satisfying $\dfrac{\partial \sigma(\epsilon, x)}{\partial \epsilon} > 0$ everywhere, is addressed. Dealing with heterogeneous media implies the presence of *non-autonomous* fluxes, *i.e.* fluxes depending directly from the physical space. From a FV point-of-view the problem can be modeled with both cell-centered flux functions or cell-edge fluxes. If the equations are written in a Lagrangian frame the cell-centered flux functions, in which a flux function holds throughout each cell and can jump only between cells, can be chosen.

The system can be written as

$$\frac{\partial \mathbf{u}(x,t)}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u},x)}{\partial x} = 0, \tag{4.4}$$

where $\mathbf{u}(x,t) \in \mathbb{R}^2$ and $\mathbf{f}(\mathbf{u},x) \in \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$. In particular, if the conservative variables are the strain $\epsilon(x,t)$ and the momentum $m = \rho(x)u(x,t)$ (where $u(x,t)$ is the longitudinal velocity), it follows

$$\mathbf{u}(x,t) = \begin{bmatrix} \epsilon \\ \rho u \end{bmatrix} = \begin{bmatrix} \epsilon \\ m \end{bmatrix} \qquad \mathbf{f}(\mathbf{u},x) = \begin{bmatrix} -\, m/\rho(x) \\ -\, \sigma(\epsilon, x). \end{bmatrix} \tag{4.5}$$

The two equations (4.4) express the kinematic relation $\dfrac{\partial \epsilon(x,t)}{\partial t} = \dfrac{\partial u(x,t)}{\partial x}$ and the Newton's second law. The Jacobian matrix for the system is

$$\mathbf{f}_{\mathbf{u}}(\mathbf{u},x) = \begin{pmatrix} 0 & -1/\rho(x) \\ -\sigma_\epsilon(\epsilon,x) & 0 \end{pmatrix} \tag{4.6}$$

and the speed of sound can be expressed as $c(\mathbf{u}, x) = \sqrt{\frac{\sigma_\epsilon(\epsilon, x)}{\rho(x)}}$.

To solve the system (4.4) employing a MUSCL-Hancock approach, the fluxes at the interfaces must be provided. The problem reduces to the solution of a Riemann problem between two cells (with eventually discontinuous material properties) in which the conservative variables $\mathbf{u}(x, t)$ are constant in each cell. Formally the problem is to find the solution $\mathbf{u}(x, t)$, of the system (4.4), for $t > 0$ with the following initial condition

$$\mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L & \text{if} \quad x < 0 \\ \mathbf{u}_R & \text{if} \quad x > 0, \end{cases} \tag{4.7}$$

where the interface between the cells is located at $x = 0$ and the material properties are constant in each cell:

$$\rho(x) = \begin{cases} \rho_L & \text{if} \quad x < 0 \\ \rho_R & \text{if} \quad x > 0, \end{cases} \qquad \text{and} \qquad \sigma(\epsilon, x) = \begin{cases} \sigma_L = \sigma(\epsilon_L) & \text{if} \quad x < 0 \\ \sigma_R = \sigma(\epsilon_R) & \text{if} \quad x > 0, \end{cases} \tag{4.8}$$

Let consider the situation sketched in the figure 4.4, where the structure of the Riemann solution between a left and a right cell is illustrated. The Riemann solution consists in general of two waves (shocks or rarefactions) one moving in the left cell and the other moving in the right cell. Being each waves confined entirely in each cell (there is always a negative and a positive eigenvalues $\lambda_{1,2} = \mp c(\mathbf{u}, x)$) this will be a standard shock or rarefaction relative to each own material; the characteristic speeds never cross zero (the eigenvalues never changes sign in the Lagrangian frame) and hence no entropy fix needs to be designed.

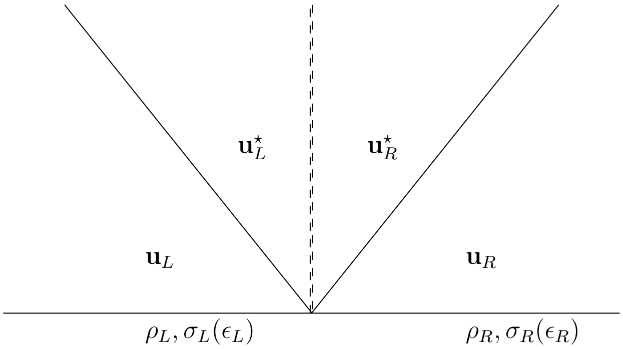


Figure 4.4: Structure of the Riemann solution for the elastic waves propagation problem (4.4).

The Riemann problem (4.7) can be solved exactly solving a system of two nonlinear equations, for the strain in both sides of the star region, arising from the continuity of the fluxes at the interface [54]. It is important to note that, due to the heterogeneous properties of the material at two neighboring cells, two stationary waves are present at the interface carrying one a jumps in the strain ϵ and one a jump in the momentum m. However, the flux must be continuous for $t > 0$ and this is physically reasonable in which the flux components are the momentum and the stress that need to be continuous at the interface. In [20, 54] an approximate Riemann problem has been proposed for this kind of problems. The same is employed in this work and it is described in the following. Defining the impedance as $Z(\mathbf{u}, x) = \rho(x)c(\mathbf{u}, x)$ the eigenvectors of the Jacobian matrix (4.6) are (for the two eigenvalues $\lambda_{1,2} = \mp c(\mathbf{u}, x)$)

$$\mathbf{r}_1(\mathbf{u}, x) = \begin{bmatrix} 1 \\ Z(\mathbf{u}, x) \end{bmatrix} \qquad \mathbf{r}_2(\mathbf{u}, x) = \begin{bmatrix} 1 \\ -Z(\mathbf{u}, x) \end{bmatrix}. \tag{4.9}$$

If in each cell the wave speed $s_{1,2}$ is chosen to be the sound speed in the appropriate cell (this is a good approximation it the non-linearity is not too strong)

$$s_1 = -\sqrt{\frac{\sigma_L(\epsilon_L)}{\rho_L}} \quad s_2 = \sqrt{\frac{\sigma_R(\epsilon_R)}{\rho_R}}, \tag{4.10}$$

the corresponding impedances can be computed as

$$\begin{aligned} Z_L(\mathbf{u}_L) &= -\rho_L s_1 \\ Z_R(\mathbf{u}_R) &= \rho_R s_2. \end{aligned} \tag{4.11}$$

Finally it is possible to define the Jacobian matrix, at the interface, in terms of its eigenvalues ($s_{1,2}$) and eigenvectors

$$R = \begin{pmatrix} 1 & 1 \\ Z_L & -Z_R \end{pmatrix}. \tag{4.12}$$

The standard approach consists in decomposing the jump of the conservative variables $\mathbf{u}$ in term of eigenvectors

$$\mathbf{u}_R - \mathbf{u}_L = R\alpha, \tag{4.13}$$

where $\alpha \in \mathbb{R}^2$. However, the scheme will result conservative only if the Jacobian matrix at the interface $\mathbf{f}^\star_\mathbf{u}$ satisfies

$$\mathbf{f}^\star_\mathbf{u} (\mathbf{u}_R - \mathbf{u}_L) = \mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L). \tag{4.14}$$

A such approach will fail for conservation laws with *non-autonomous* flux functions due to the presence of stationary waves at the interface that are not explicitly taken into account in the decomposition (4.13). A more convenient decomposition in term of flux differences can be searched as

$$\mathbf{f}_R - \mathbf{f}_L = R\beta, \tag{4.15}$$

where $\beta \in \mathbb{R}^2$. This decomposition, in term of flux difference, is already adequate to compute the flux at the interface (as needed in the MHM):

$$f(x = 0) = f_L(\mathbf{u}_L) + \beta_1 \begin{bmatrix} 1 \\ Z_L \end{bmatrix} = f_R(\mathbf{u}_L) - \beta_2 \begin{bmatrix} 1 \\ -Z_R \end{bmatrix}. \tag{4.16}$$

The conservative properties (4.14) is then satisfied defining α as

$$\mathbf{f}^\star_\mathbf{u} (\mathbf{u}_R - \mathbf{u}_L) = \mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = \mathbf{f}^\star_\mathbf{u} R\alpha = R\Lambda\alpha = R\beta, \tag{4.17}$$

where $\mathbf{f}^\star_\mathbf{u} = R\Lambda R^{-1}$ with $\Lambda = \mathrm{Diag}(s_1, s_2)$.

In this work, the non-linear propagation of the compressional waves is considered for a rod (defined on $\Omega = [0,1]$), with heterogeneous material (there is a discontinuity in the density of the material, see figure 4.6)

$$\rho(x) = \begin{cases} 1.4 & \text{if } \dfrac{1}{2} \leq x \leq \dfrac{4}{5} \\ 1 & \text{otherwise,} \end{cases} \tag{4.18}$$

within a quadratic constitutive relation (equal over the whole physical domain Ω) of the form

$$\sigma(\epsilon) = K_1\epsilon + K_2\epsilon^2, \tag{4.19}$$

where $K_1 = 2$ and $K_2 = K_2(\xi) = \xi$ with $\xi \sim \mathcal{U}(0,1)$. The following initial condition (independent from the random parameter) is considered (see also the figure 4.5)

$$\mathbf{u}(x,0) = \begin{bmatrix} 2e^{\frac{-(x-7/20)^2}{1/250}} \\ -2e^{\frac{-(x-7/20)^2}{1/250}} \end{bmatrix} \tag{4.20}$$

Boundary conditions of solid wall type are imposed to model a clamped rod situation.

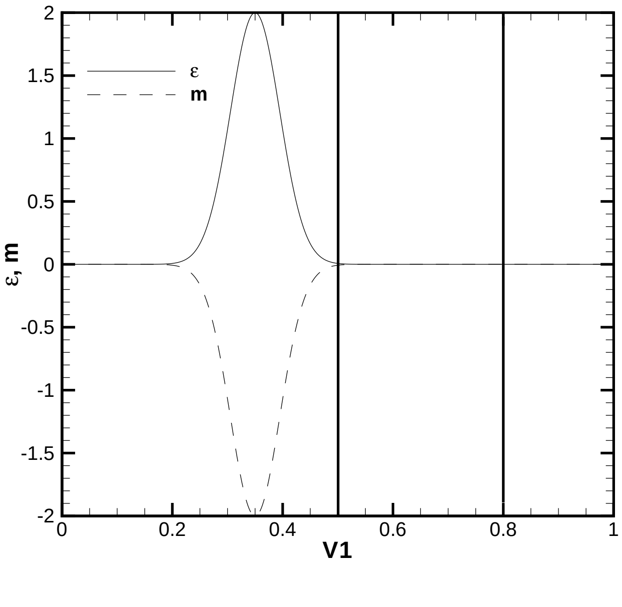


Figure 4.5: Initial condition for the elastic waves propagation in a heterogeneous rod (4.18).

The problem, in presence of the random parameter ξ, is represented in the figure 4.6.

The physical space has been discretized on a mesh of $N_x = 201$ equally spaced points with a resolution of $\Delta x = 5 \times 10^{-3}$, while the time space $T = [0, 0.5]$ has been divided in $N_t = 500$ time steps of length $\Delta t = 10^{-3}$. A reference solution has been obtained considering a resolution level equal to $\Delta\xi = 1/(5 \times 2^{18})$ corresponding to $N_\xi = 1\,310\,721$ equally spaced points on Ξ. The reference solution, in term of expected

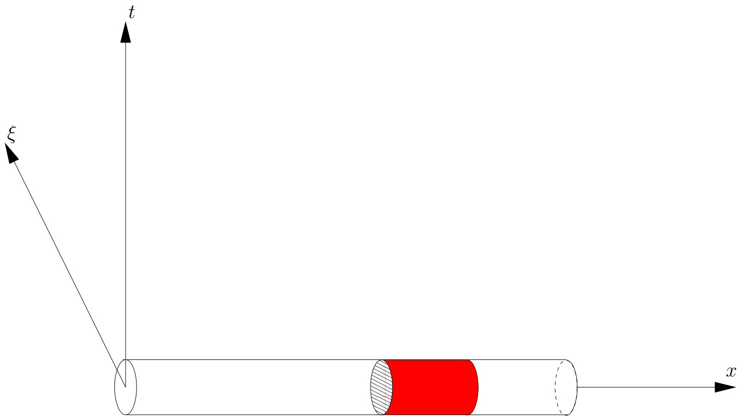


Figure 4.6: Sketch of the rod with heterogeneous material. The red part indicates the inclusion with the abrupt change in the density (4.18).

value and variance, for the two components of the vector of the conservative variables $\mathbf{u}$, is reported in figure 4.7 in the plane $x - t$.

From the figure 4.7, the effect of the change in the density of the rod is evident. The impedance, itself, results to be a non-linear random function

$$Z(\mathbf{u}, x, \xi) = \rho(x)c(\mathbf{u}, x, \xi) = \sqrt{\rho(x)\frac{\partial \sigma}{\partial \epsilon}(\mathbf{u}, x, \xi)} = \sqrt{\rho(x)\left(K_1 + 2K_2(\xi)\epsilon(x,t,\xi)\right)}. \quad (4.21)$$

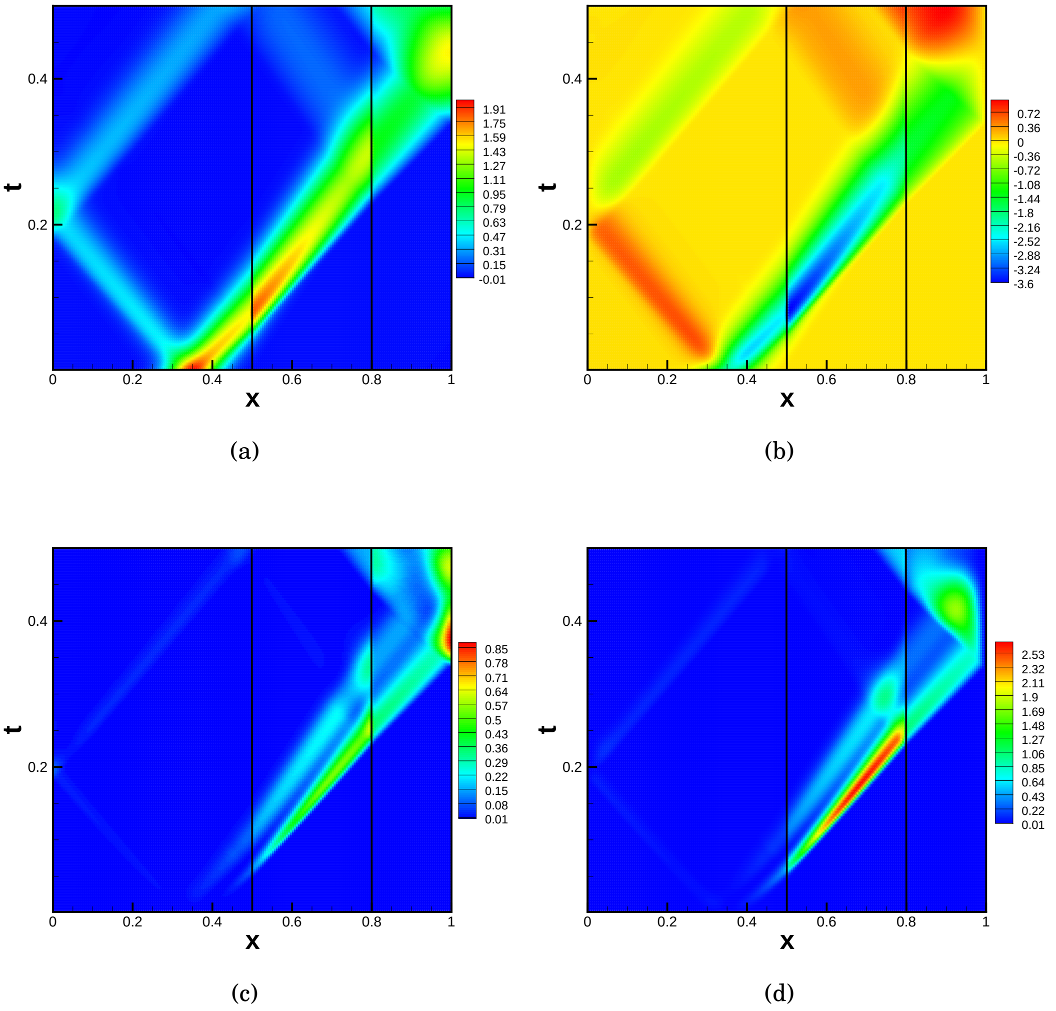
A traveling elastic wave moves locally with a speed dependent from the local speed of sound $c(\mathbf{u}, x, \xi)$, hence, in the region of the rod with different density, the waves decelerate (where $\rho = 1.4$) or accelerate. However, a secondary effect appears: there is a partial reflection of the waves when passing through an interface dividing two materials with different impedance.

Knowing the reference solution on the uniform tessellation of $N_\xi = 1\,310\,720$ intervals, the error norms for the statistics are computed on the whole physical-time space $x - t$ using the following relations

$$\begin{aligned} \mathrm{err}_{\mu^m}|_{L^p} &= ||\mu^m - \mu^m_{\mathrm{ref}}||_{L^p} = \left(\frac{1}{N_t \times N_x} \sum_{i=1}^{N_t} \sum_{j=1}^{N_x} \left|\mu^m_{ij} - \mu^m_{\mathrm{ref},ij}\right|^p \right)^{1/p} \\ \mathrm{err}_{\mu^m}|_{L^\infty} &= ||\mu^m - \mu^m_{\mathrm{ref}}||_{L^\infty} = \max_{ij} \left|\mu^m_{ij} - \mu^m_{\mathrm{ref},ij}\right|, \end{aligned} \quad (4.22)$$

with $\mu_{ij} = \mu(x_j, t_i)$ the generic statistical moment. The sTE algorithm is applied with the following parameters: $m_0 = 3$, $m_L = [6, 14]$ (with step of 2) and $\varepsilon = 10^{-1}$. All the reconstruction operators are employed: linear $r = 1$, centered cubic $r = 3$, ENO and WENO based on cubic polynomials. For comparison, both the quasi-MC ($N_\xi = [20, 140]$) and the PC (with degree between 20 and 60 with step of 10) results are also reported in the figures 4.8 and 4.9.

To demonstrate the efficiency of the sTE algorithm the results, obtained without compression (W/O compression), are also reported in the figures 4.8 and 4.9. It is evident that, even without compression, the scheme performs better then the MC, but worse with respect to the PC. The results of the non-compressed scheme are improved by the sTE scheme with linear reconstruction, but the overall performance



(a)

(b)

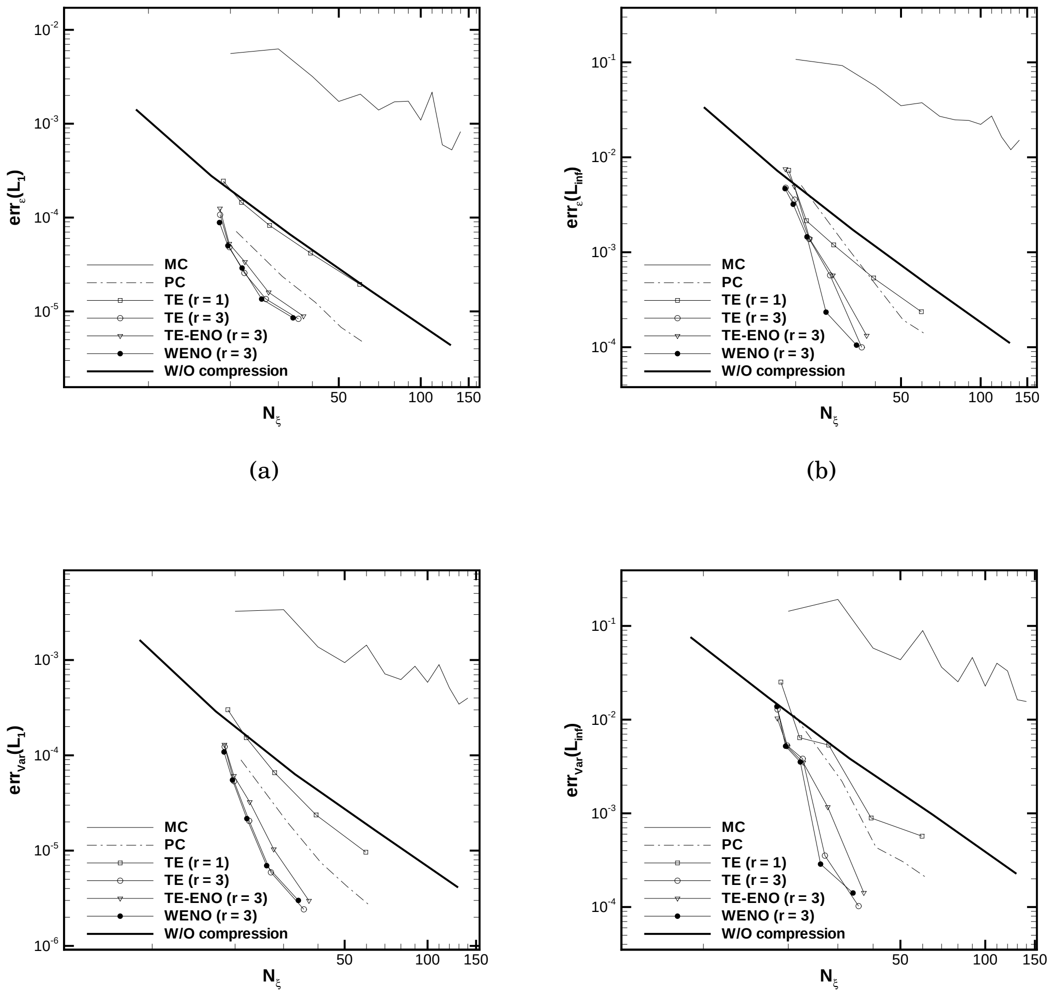
(c)

(d)

Figure 4.7: Statistics for the compressional wave propagation problem in a heterogeneous rod. In the first row the expectancies for the strain ϵ (a) and the momentum m (b) are reported, while in the second row their variance are shown. The two bold line indicates the region in which the density of the rod changes (see also the equation (4.18)).

of the scheme remain worse with respect to the PC. The introduction of the high-reconstruction operator improves the results and in particular the introduction of the WENO reconstruction allows to recover the better convergence without loosing accuracy in the smooth regions. The introduction of the ENO causes a slight deterioration of the compression capabilities, due to non-optimal selection of the stencils in the smooth regions, producing a scheme that does not perform well as the centered scheme. However, the optimal compression is recovered by the WENO reconstruction. Obviously, the application of the sTE algorithm produces a time-space varying mesh in the stochastic space, hence in all the numerical results the number of points indicates the average number of points employed during the entire simulation.

In this section some results on the sTE scheme have been presented for the non-



(a)

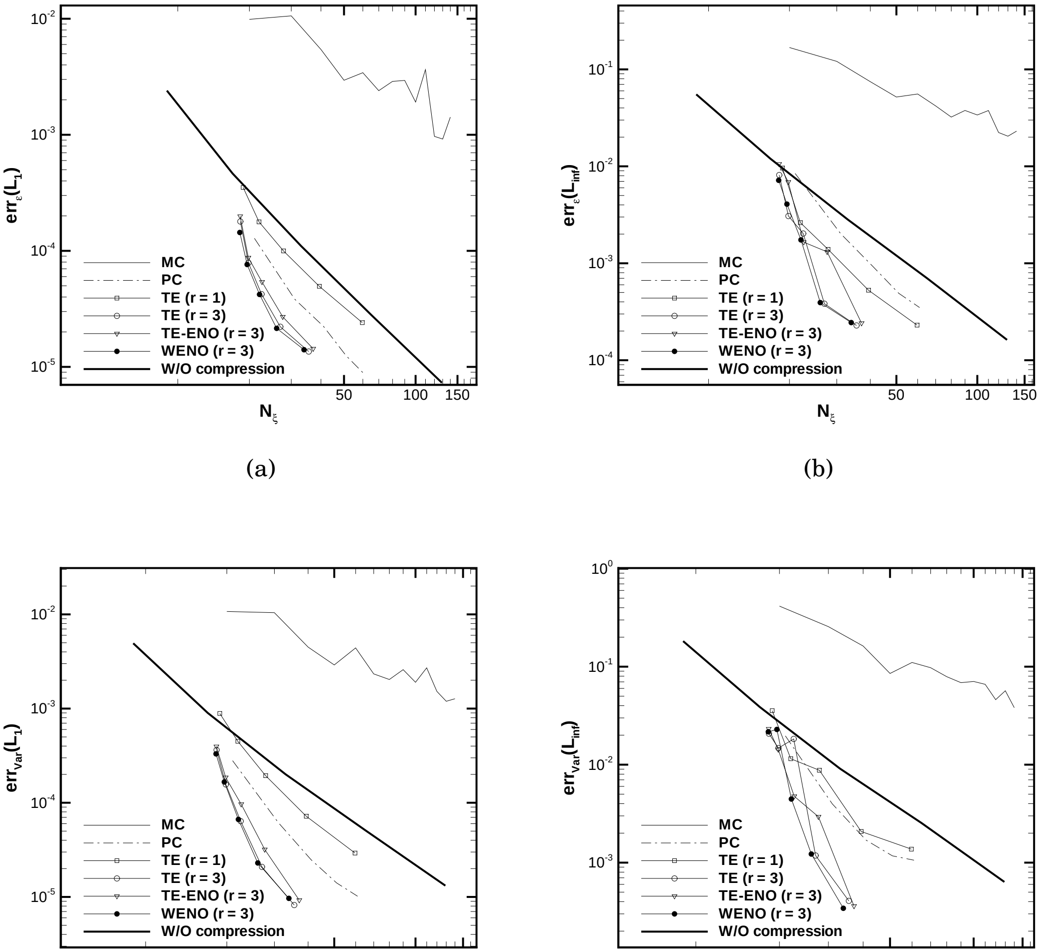
(b)

(c)

(d)

Figure 4.8: Error for the statistics of the strain ϵ in norms L^1 and L^∞ following (4.22).

linear elastic wave propagation problem. However the sTE is demonstrated to performs better than the MC and PC for a wide range of problems (also for stochastic ODEs). Several examples are reported in the papers ***P1***, ***P2*** and ***P3***. Numerical results for the Euler system of equations have been obtained and presented during the 25th Summer Program 2012 of the *Center for Turbulence Research* at the Stanford University and they are also reported in ***C8***. In the next section some results for the TE/aSI schemes, in the context of cell-average MR framework, in the multidimensional 2D and 3D stochastic cases, will be presented. For the results relative to the 1D stochastic aSI scheme, it is necessary to refer to the paper ***P4***.



(a)

(b)

(c)

(d)

Figure 4.9: Error for the statistics of the momentum m in norms L^1 and L^∞ following (4.22).

m_L	N_w	N_{eval}	μ	τ	err_{L_1}	err_{L_∞}
			$r = 1$			
6	18	33	0.3611111E+01	0.1969697E+01	0.1076455E-01	0.2036121E-02
7	24	45	0.5375000E+01	0.2866667E+01	0.5504224E-02	0.1247305E-02
8	27	51	0.9518518E+01	0.5039216E+01	0.3773556E-02	0.1120426E-02
9	32	61	0.1603125E+02	0.8409836E+01	0.3072174E-02	0.9275682E-03
10	42	81	0.2440476E+02	0.1265432E+02	0.1877892E-02	0.4878891E-03
11	54	105	0.3794444E+02	0.1951429E+02	0.9460814E-03	0.2630980E-03
12	63	123	0.6503175E+02	0.3330894E+02	0.7898986E-03	0.2049418E-03
13	85	167	0.9638824E+02	0.4905988E+02	0.3863797E-03	0.1053899E-03
14	103	203	0.1590777E+03	0.8071429E+02	0.2366288E-03	0.6703958E-04
15	122	241	0.2685984E+03	0.1359709E+03	0.1965388E-03	0.5040153E-04
16	161	319	0.4070621E+03	0.2054451E+03	0.9923035E-04	0.2763382E-04
17	196	389	0.6687398E+03	0.3369486E+03	0.5915678E-04	0.1750254E-04
18	237	471	0.1106097E+04	0.5565711E+03	0.4904217E-04	0.1264509E-04
19	315	627	0.1664410E+04	0.8361866E+03	0.2434267E-04	0.6904394E-05
20	383	763	0.2737799E+04	0.1374282E+04	0.1478987E-04	0.4403922E-05
			$r = 3$			
6	18	33	0.3611111E+01	0.1969697E+01	0.7117425E-03	0.8735547E-04
7	21	39	0.6142857E+01	0.3307692E+01	0.7117425E-03	0.9449274E-04
8	24	45	0.1070833E+02	0.5711111E+01	0.7554024E-03	0.9610891E-04
9	31	59	0.1654839E+02	0.8694915E+01	0.2774332E-03	0.3455729E-04
10	36	69	0.2847222E+02	0.1485507E+02	0.1440048E-03	0.1581990E-04
11	41	79	0.4997561E+02	0.2593671E+02	0.5072283E-04	0.7745037E-05
12	44	85	0.9311364E+02	0.4820000E+02	0.5072283E-04	0.7747611E-05
13	47	91	0.1743192E+03	0.9003297E+02	0.5072283E-04	0.7748743E-05
14	57	111	0.2874561E+03	0.1476126E+03	0.5072513E-04	0.3817509E-05
15	64	125	0.5120156E+03	0.2621520E+03	0.1184295E-04	0.1398270E-05
16	75	147	0.8738267E+03	0.4458299E+03	0.4712994E-05	0.4914680E-06
17	79	155	0.1659152E+04	0.8456323E+03	0.3622172E-05	0.4409493E-06
18	84	165	0.3120774E+04	0.1588758E+04	0.2246242E-05	0.3779013E-06
19	101	199	0.5190980E+04	0.2634618E+04	0.1560115E-05	0.1507988E-06
20	110	217	0.9532519E+04	0.4832152E+04	0.7982875E-06	0.9490253E-07
			$r = 3$ (ENO)			
6	14	25	0.4642857E+01	0.2600000E+01	0.3700040E-02	0.3713698E-03
7	16	29	0.8062500E+01	0.4448276E+01	0.7526094E-03	0.1680260E-03
8	18	33	0.1427778E+02	0.7787879E+01	0.8017963E-03	0.1342630E-03
9	23	43	0.2230435E+02	0.1193023E+02	0.2852286E-03	0.5378220E-04
10	28	53	0.3660714E+02	0.1933962E+02	0.1601275E-03	0.1929446E-04
11	31	59	0.6609677E+02	0.3472881E+02	0.4692711E-04	0.1011992E-04
12	32	61	0.1280312E+03	0.6716393E+02	0.4692711E-04	0.1012630E-04
13	35	67	0.2340857E+03	0.1222836E+03	0.4408086E-04	0.7037415E-05
14	42	81	0.3901190E+03	0.2022840E+03	0.2570142E-04	0.3276826E-05
15	47	91	0.6972128E+03	0.3600989E+03	0.1224823E-04	0.1940414E-05
16	57	111	0.1149772E+04	0.5904234E+03	0.4841718E-05	0.6950600E-06
17	59	115	0.2221576E+04	0.1139765E+04	0.2960107E-05	0.6360469E-06
18	65	127	0.4033000E+04	0.2064134E+04	0.2399054E-05	0.4181538E-06
19	79	155	0.6636570E+04	0.3382510E+04	0.1513161E-05	0.1847706E-06
20	86	169	0.1219276E+05	0.6204598E+04	0.8283872E-06	0.1256176E-06
			$r = 3$ (WENO)			
6	15	27	0.4333333E+01	0.2407408E+01	0.1637357E-02	0.8552741E-04
7	16	29	0.8062500E+01	0.4448276E+01	0.1637357E-02	0.9293030E-04
8	18	33	0.1427778E+02	0.7787879E+01	0.7678849E-03	0.3087795E-04
9	21	39	0.2442857E+02	0.1315385E+02	0.2378888E-03	0.1082544E-04
10	26	49	0.3942308E+02	0.2091837E+02	0.4168824E-04	0.2128466E-05
11	28	53	0.7317857E+02	0.3866038E+02	0.4168824E-04	0.2008684E-05
12	30	57	0.1365667E+03	0.7187719E+02	0.4168824E-04	0.1306688E-05
13	32	61	0.2560312E+03	0.1343115E+03	0.4168824E-04	0.1080161E-05
14	36	69	0.4551389E+03	0.2374638E+03	0.1079712E-04	0.2653878E-06
15	38	73	0.8623421E+03	0.4488904E+03	0.2284037E-05	0.1415316E-06
16	42	81	0.1560405E+04	0.8090988E+03	0.2284037E-05	0.8807150E-07
17	47	87	0.2788787E+04	0.1506586E+04	0.2284037E-05	0.6575382E-07
18	55	103	0.4766273E+04	0.2545097E+04	0.2284039E-05	0.3413663E-07
19	60	113	0.8738150E+04	0.4639726E+04	0.3728225E-06	0.1972595E-07
20	66	125	0.1588753E+05	0.8388616E+04	0.3728225E-06	0.1949176E-07

Table 4.1: Results of the application of the TE algorithm on the steady functions $f = f(\xi)$ (see equation 4.1).

4.2 Preliminary results for the multidimensional TE/aSI schemes

In the previous section, some numerical results on the TE and sTE schemes are presented. The MR framework can be formulated relying on discretization operators of cell-average kind, as described in the previous chapters. The aim of this section is to present some preliminary results on the TE scheme obtained in the context of the cell-average framework. In particular, in Section §4.2.1, the TE algorithm is applied on steady functions in 2D and 3D stochastic spaces to show the convergence properties of the approach. In Sections §4.2.2 and §4.2.3, the aSI scheme is applied to stochastic PDE problems. The aim is to demonstrate the applicability and the efficiency of the scheme to the classical CFD test case: the linear advection equation, the Burgers equation and the Euler system of equations. These are only preliminary results and the work on multidimensional aSI scheme is still in progress. However, a complete set of results, demonstrating the efficiency of the aSI scheme, with respect to the original SI approach, is presented in paper ***P4***.

4.2.1 2D/3D steady functions

In this section the TE algorithm based on the cell-average discretization is applied to some steady functions, *i.e.* functions depending only on the stochastic space $f = f(\boldsymbol{\xi})$ where $\boldsymbol{\xi} \in \Xi \subset \mathbb{R}^d$ with $d = 2, 3$. The numerical results concern the statistics (expected value and variance) and the error norms between the exact function and the reconstructed function $\mathcal{R}_L \hat{v}^L$.

The functions are the following

$$\begin{aligned}
\textit{circles:}\quad & f(\boldsymbol{\xi}) = \begin{cases} 1 & \text{if} \quad \xi_1^2+\xi_2^2 \leq \dfrac{1}{16} \text{ or } (\xi_1-1)^2+(\xi_2-1)^2 \leq \dfrac{1}{16} \text{ or} \\ & \quad (\xi_1-1)^2+\xi_2^2 \leq \dfrac{1}{16} \text{ or } \xi_1^2+(\xi_2-1)^2 \leq \dfrac{1}{16} \\ 0 & \text{otherwise} \end{cases} \\
\textit{spheres:}\quad & f(\boldsymbol{\xi}) = \begin{cases} 1 & \text{if} \quad \xi_1^2+\xi_2^2+\xi_3^2 \leq \dfrac{1}{16} \text{ or } (\xi_1-1)^2+(\xi_2-1)^2+(\xi_3-1)^2 \leq \dfrac{1}{16} \text{ or} \\ & \quad (\xi_1-1)^2+\xi_2^2+(\xi_3-1)^2 \leq \dfrac{1}{16} \text{ or } \xi_1^2+(\xi_2-1)^2+(\xi_3-1)^2 \leq \dfrac{1}{16} \\ & \quad (\xi_1-1)^2+(\xi_2-1)^2+\xi_3^2 \leq \dfrac{1}{16} \text{ or } (\xi_1-1)^2+\xi_2^2+\xi_3^2 \leq \dfrac{1}{16} \\ & \quad \xi_1^2+(\xi_2-1)^2+\xi_3^2 \leq \dfrac{1}{16} \text{ or } \xi_1^2+\xi_2^2+(\xi_3-1)^2 \leq \dfrac{1}{16} \\ 0 & \text{otherwise} \end{cases} \\
\textit{corner peak:}\quad & f(\boldsymbol{\xi}) = \frac{1}{\left(1+\sum_{i=1}^{d} c_i \xi_i\right)^q} \\
\textit{discontinuous:}\quad & f(\boldsymbol{\xi}) = \begin{cases} \dfrac{1}{50} e^{\sum_{i=1}^{d} c_i \xi_i} & \text{if} \quad \boldsymbol{\xi} \leq \boldsymbol{\xi}^0 \\ \qquad 0 & \text{otherwise,} \end{cases}
\end{aligned} \tag{4.23}$$

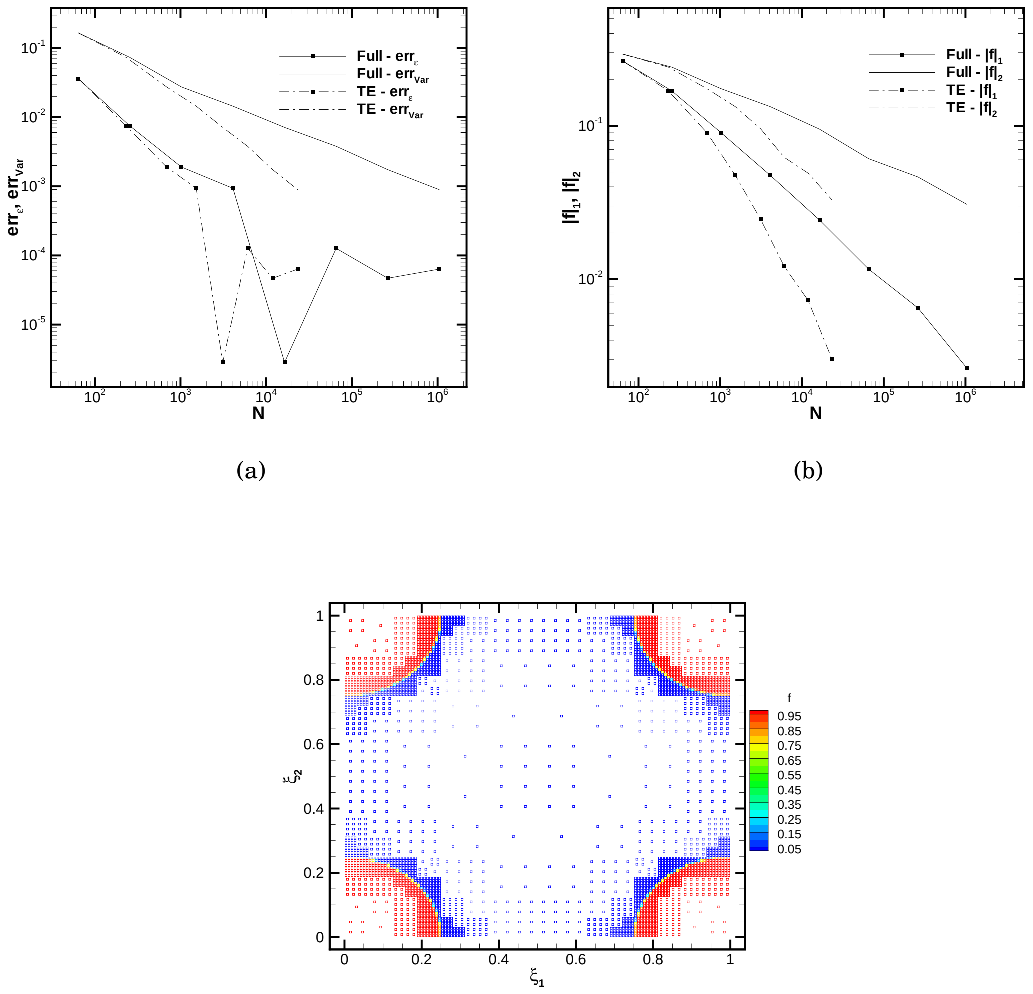
where both the *corner peak* and *discontinuous* functions are employed in 2D and 3D stochastic spaces with the following parameters: $q = 4$ with $c = \{1, 6\}$ (2D) or $c = \{1, 6, 12\}$ (3D) for the *corner peak* while $c_i = 5$ (2D/3D) and $\boldsymbol{\xi}^0 = (0.53, 0.33)$ (2D) or $\boldsymbol{\xi}^0 = (0.53, 0.33, 0.63)$ (3D) for the *discontinuous* function.

All the functions chosen are analytical and the statistics can be computed exactly, hence the error can be evaluated, as already done in the point-value case (see equations (4.3)). In this case the convergence properties of the scheme are presented computing the norms of the error as

$$
\begin{aligned}
|f|_1 &= \frac{\int_\Xi |f(\boldsymbol{\xi}) - \hat{f}(\boldsymbol{\xi})| \mathrm{d}\boldsymbol{\xi}}{\int_\Xi |f(\boldsymbol{\xi})| \mathrm{d}\boldsymbol{\xi}} \simeq \frac{\sum_{j=1}^N |f(\boldsymbol{\xi}_j) - \hat{f}(\boldsymbol{\xi}_j)|}{\sum_{j=1}^N |f(\boldsymbol{\xi}_j)|} \\
|f|_2 &= \frac{\int_\Xi |f(\boldsymbol{\xi}) - \hat{f}(\boldsymbol{\xi})|^2 \mathrm{d}\boldsymbol{\xi}}{\int_\Xi |f(\boldsymbol{\xi})|^2 \mathrm{d}\boldsymbol{\xi}} \simeq \frac{\sum_{j=1}^N |f(\boldsymbol{\xi}_j) - \hat{f}(\boldsymbol{\xi}_j)|^2}{\sum_{j=1}^N |f(\boldsymbol{\xi}_j)|^2},
\end{aligned}
\tag{4.24}
$$

where the reconstructed function is obtained applying the reconstruction operator $\mathcal{R}_k$ at the last resolution level L: $\hat{f}(\boldsymbol{\xi}) = \mathcal{R}_L \hat{v}^L$. The norms (4.24) are computed on a fixed set of points ($N = 10000$) generated by a quasi-MC Sobol sequence. In the figures 4.10, 4.11 and 4.12 the results relative to the 2D cases are reported. In each figure the error for the statistics (mean and variance) and the errors in norms L^1 and L^2 following (4.24) are reported. The parameters for the TE algorithm are $m_0 = 1$, $m_L = 10$ and $\varepsilon = 10^{-4}$. To make evident the advantage of the application of the TE algorithm, the results relative to the full approach are also reported. The full approach consists in generating a uniform tessellation on which the reconstruction operator is applied directly, obtaining the linear approximation of the function on each cell. This reconstruction operator is employed to perform the quadrature and also the computation of the point values for the norm computations. In both cases, and in all the multidimensional stochastic cases presented in this manuscript, the reconstruction operator is based on local linear polynomial (conservative) interpolation as described in §2.3.1. Of course, this is actually a limitation of the approach, but the extension, as already made in the 1D case, is underway. However, from a global accuracy point-of-view of the aSI scheme, the conservative linear interpolation is the same employed in the physical space and hence it appears less restrictive for the solution of stochastic partial differential equation; at the same time the reconstruction operator plays a fundamental role in the compression, *i.e.* efficiency, of the scheme. For that reason the extension of the algorithm to high-order reconstruction techniques appears to be a necessary further step. Moreover, in the figures 4.10, 4.11 and 4.12 the patterns of the cells, obtained applying the TE scheme with the parameters $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$, are also reported, for all the cases, making evident the capability of the algorithm locating the discontinuous/high-gradient regions and concentrating the computational effort. In the figures 4.13, 4.14 and 4.15 the results for the 3D cases are reported.

For all the cases presented, the results are quite similar. In particular, it is evident the capability compression of the scheme with respect the full approach and the consequent improvement in term of efficiency. The effect is a translation of the curves towards left (less functional evaluations) for all the quantities. The advantage increases augmenting the number of evaluations. This results is valid for both the statistics and the error norms. In this section, the TE algorithm has been presented to show its capability to reduce the computational cost with respect the full approach. The SI scheme (see Chapter 1), with the introduction of the TE algorithm, as described in the previous chapters, results in the aSI scheme allowing a reduction of the global computational cost. In the following sections some preliminary multidimensional results are presented.



(a)

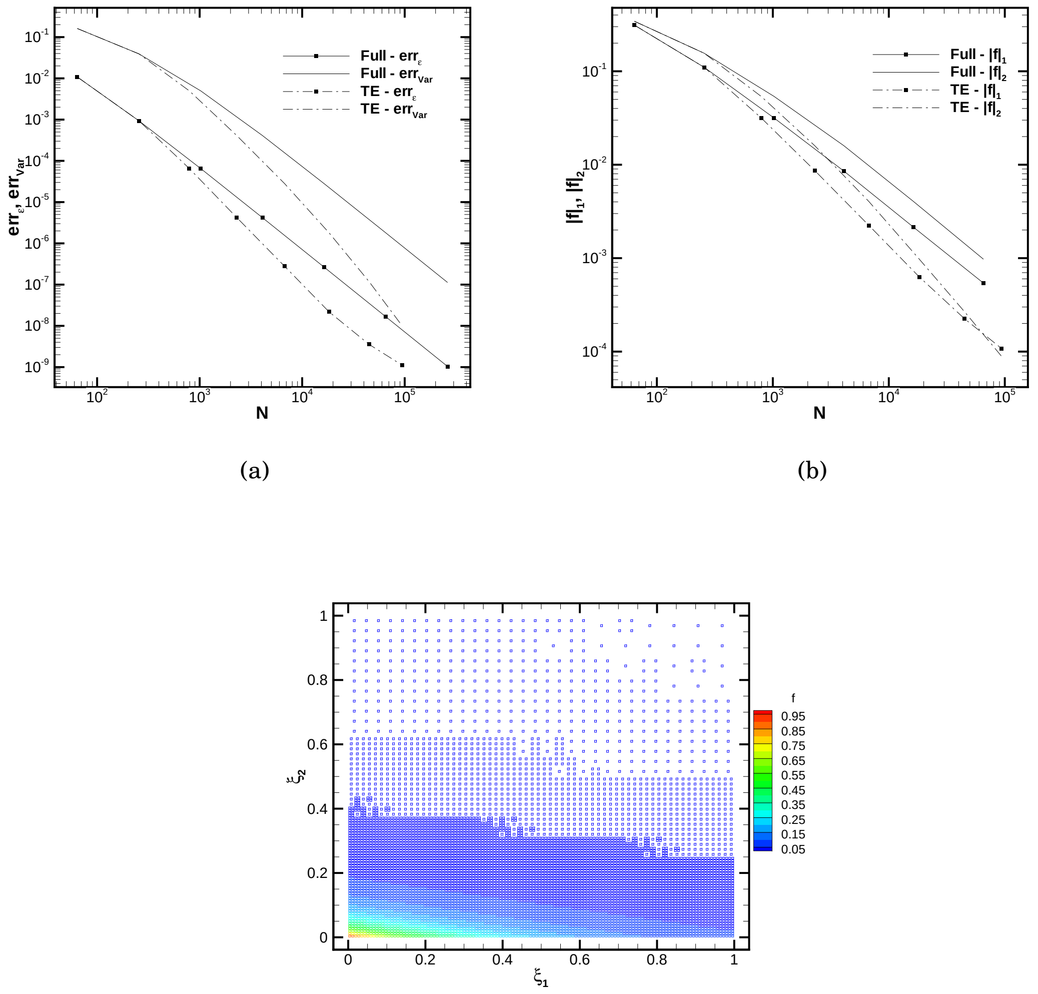
(b)

(c)

Figure 4.10: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 2D *circles* function (4.23).

4.2.2 2D test problems for stochastic PDE

In this section the numerical results for the aSI scheme in two stochastic dimensions are presented. The test cases are an extension of the classical (deterministic) CFD problems (see the paper ***P4*** for the 1D stochastic space analysis). The test cases are of increasing complexity and ranging from the linear advection problem, with both random advection velocity and initial conditions, the Burgers equation, with uncertain smooth initial condition, and the Euler equations of the gasdynamics. For all the test cases the physical space is one-dimensional. This is only a simplification in term of computational cost, because in principle the aSI scheme does not suffer from any problem regarding the extension of the physical space. However, the application of intrusive adaptive schemes in one-dimensional physical space, represents actually the state-of-the art, see for instance [66, 86].



(a)

(b)

(c)

Figure 4.11: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 2D *corner peak* function (4.23).

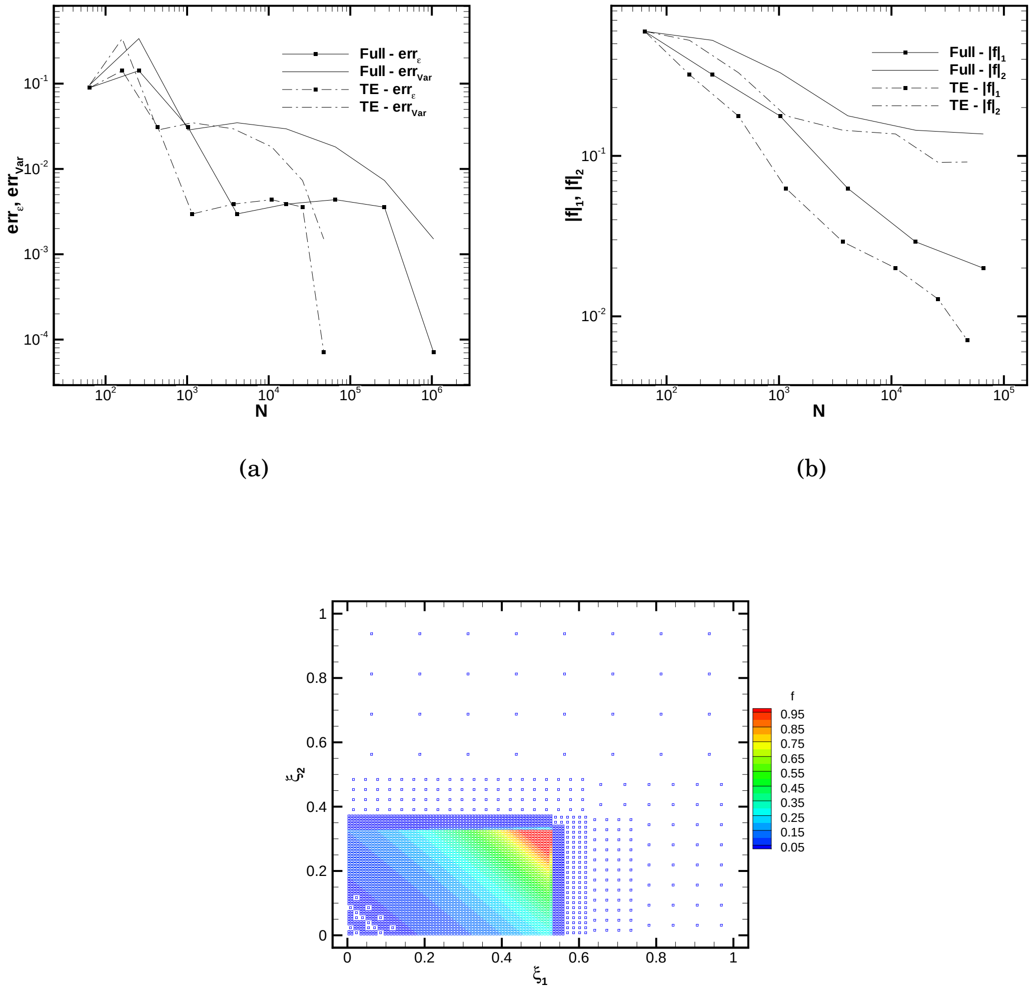
Linear advection equation

The first test case is the classical linear advection problem

$$\begin{cases} \dfrac{\partial u(x,\boldsymbol{\xi},t)}{\partial t} + a(\boldsymbol{\xi},t)\dfrac{\partial u(x,\boldsymbol{\xi},t)}{\partial x} = 0 \quad t > 0 \\ u(x,\boldsymbol{\xi},0) = u_0(x,\boldsymbol{\xi}) \quad x \in \Omega \subset \mathbb{R} \text{ and } \boldsymbol{\xi} \in \Xi \subset \mathbb{R}^2, \end{cases} \tag{4.25}$$

where both the advection velocity a and the initial condition u_0 can depend on a random vector $\boldsymbol{\xi}$. Each of the random parameter is uniformly distributed $\xi_i \sim \mathcal{U}(0,1)$ with $i = 1,2$. Two case are considered: the first one with smooth random initial condition $u_0 = u_0(x,\boldsymbol{\xi})$

$$u_0(x,\boldsymbol{\xi}) = \sin(2\pi x + \xi_1^2 + \xi_2^2), \tag{4.26}$$



(a)

(b)

(c)

Figure 4.12: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 2D *discontinuous* function (4.23).

with constant advection velocity $a = 0.1$; the second one has a discontinuous initial condition $u_0 = u_0(x)$

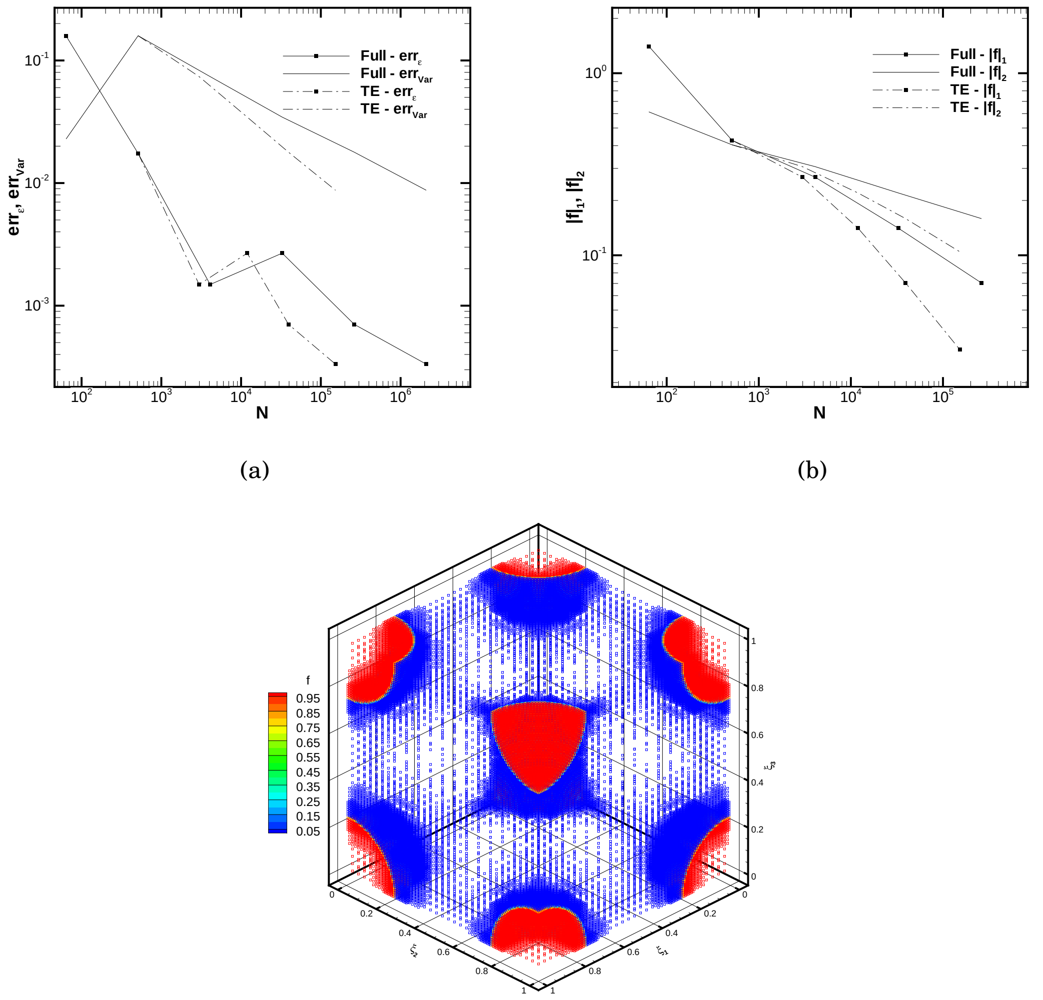
$$u_0(x) = \begin{cases} 1 & \frac{2}{5} \leq x \leq \frac{3}{5} \\ 0 & \text{otherwise} \end{cases} \tag{4.27}$$

and uncertain advection velocity

$$a(\boldsymbol{\xi}) = \xi_1^2\xi_2^2 - \xi_1^2\xi_2 - \xi_1\xi_2^2 - c\,\xi_1^2 - c\,\xi_2^2 + \xi_1\xi_2 + c\,\xi_1 + c\,\xi_2 + c^2, \tag{4.28}$$

where $c = 0.75$ in the numerical examples. Both the cases are solved in this section with periodic boundary conditions.

In the following the norms L^1 of the statistics are computed as following (for the



(a)

(b)

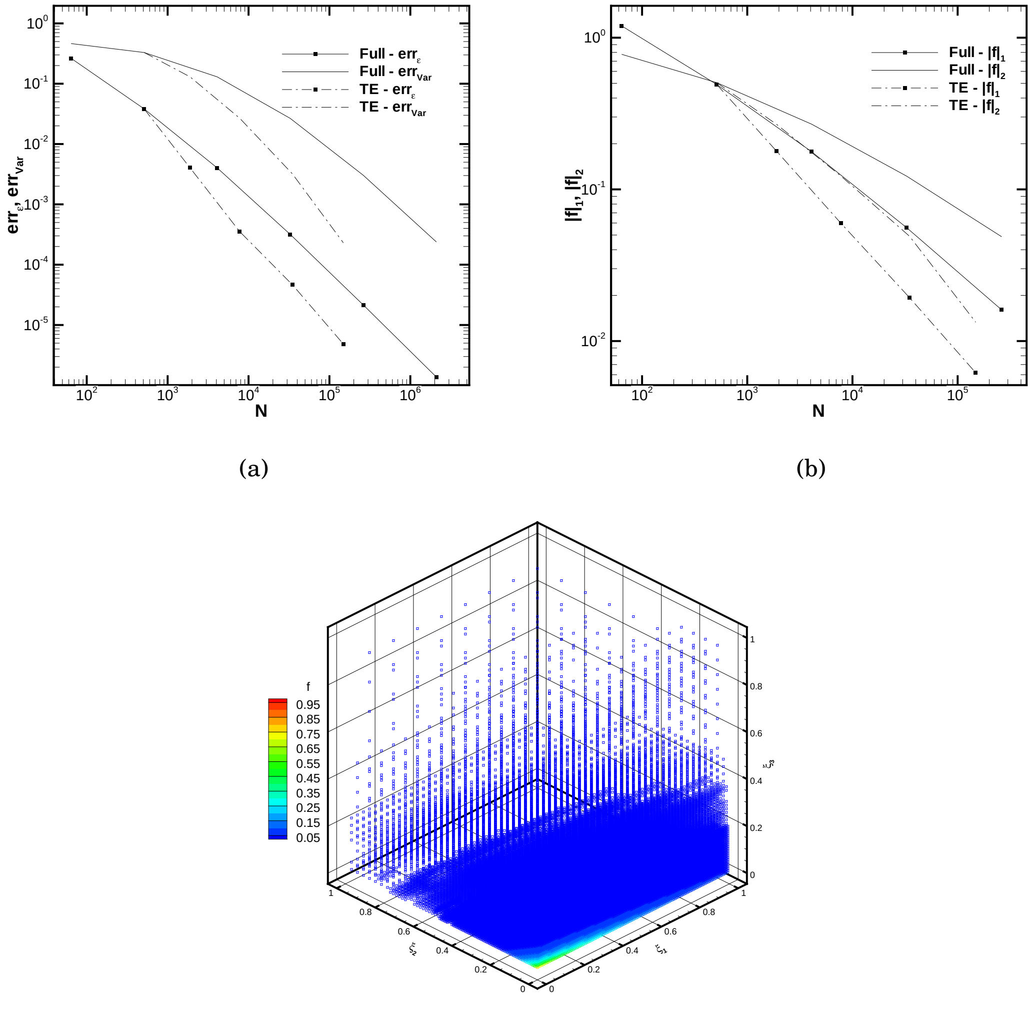
(c)

Figure 4.13: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 3D *spheres* function (4.23).

final time step $t = t_F$)

$$
\begin{aligned}
|\mathrm{err}_{\mathbb{E}}|_1 &= \int_\Omega |\mathbb{E}(u(x,t_F)) - \mathbb{E}_{\mathrm{ex}}(u(x,t_F))|\,\mathrm{d}x \\
&\simeq \Delta x \sum_{i=1}^{N_x} |\mathbb{E}(\bar{u}(x_i,t_F)) - \mathbb{E}_{\mathrm{ex}}(\bar{u}(x_i,t_F))| \\
|\mathrm{err}_{\mathrm{Var}}|_1 &= \int_\Omega |\mathrm{Var}(u(x,t_F)) - \mathrm{Var}_{\mathrm{ex}}(u(x,t_F))|\,\mathrm{d}x \\
&\simeq \Delta x \sum_{i=1}^{N_x} |\mathrm{Var}(\bar{u}(x_i,t_F)) - \mathrm{Var}_{\mathrm{ex}}(\bar{u}(x_i,t_F))|
\end{aligned}
\tag{4.29}
$$

In the figure 4.16, the results for the application of the aSI scheme are reported in term of statistical error L^1 norms, and statistics over the whole physical space. The



(a)

(b)

(c)

Figure 4.14: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 3D *corner peak* function (4.23).

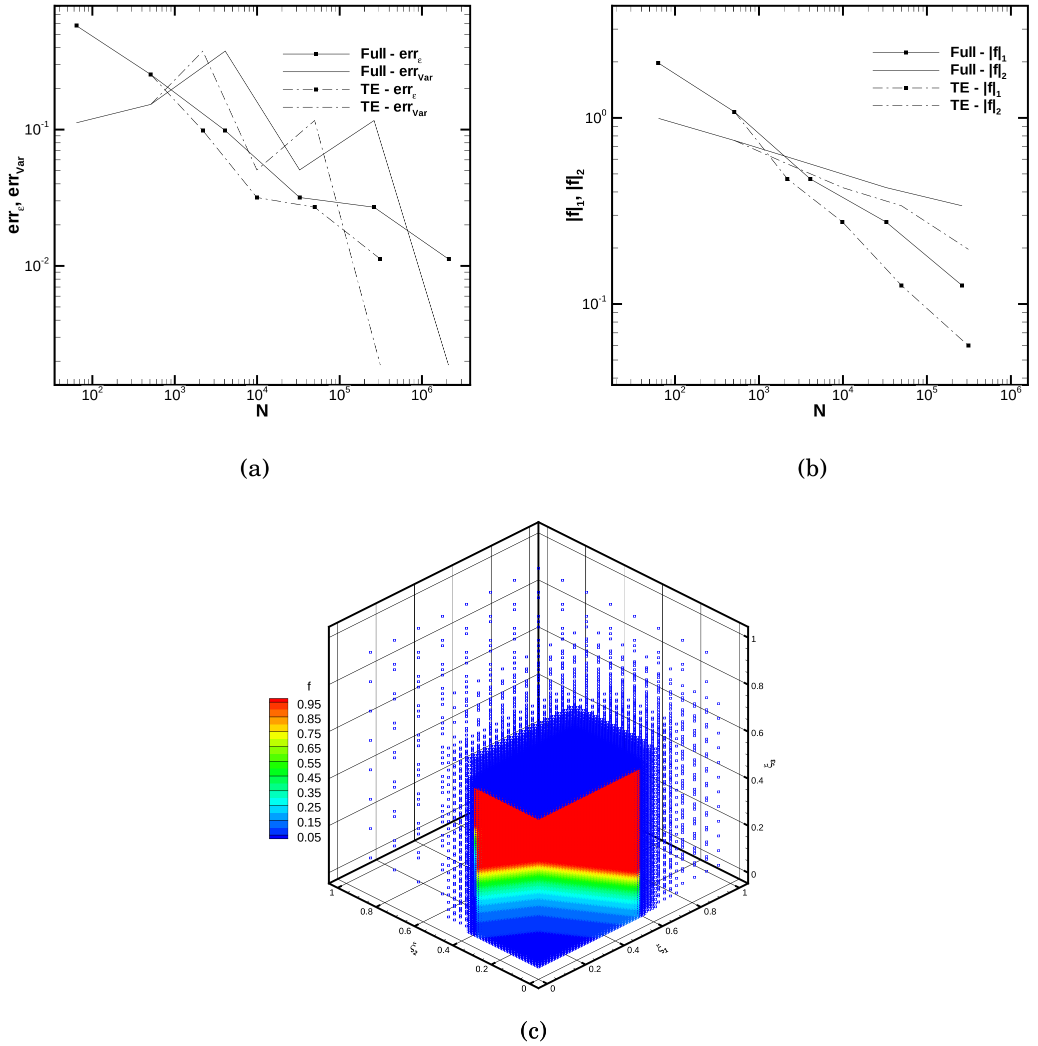
smooth advection problem is solved for increasing spatial resolutions on meshes of N_x equal to 21, 51 and 101 points, dividing the time line $T = [0, 1/2]$ in $N_t = 100$ intervals with equal length $\Delta t = 5 \times 10^{-3}$. The centered limiter can be employed due to the smooth properties of the solution. The parameters for the aSI scheme are $m_0 = 3$, $m_L = 7$ and $\varepsilon = 10^{-4}$. The exact solution can be computed by means of integration of the exact solution

$$u(x, \xi, t) = \sin\left(2\pi(x - at) + \xi_1^2 + \xi_2^2\right). \tag{4.30}$$

In this analytical case the integral are computed with the software MAPLE.

In the figure 4.16 it is evident that the effect of the adaptation, *i.e.* the time-dependent stochastic refinement/coarsening (see the figure 4.16(d)), does not modify the expected second order convergence rate of the solution in space.

The advection linear equation can be also solved with discontinuous initial condition and the random advection equation (4.28). The problem is solved for $N_t = 350$



(a)

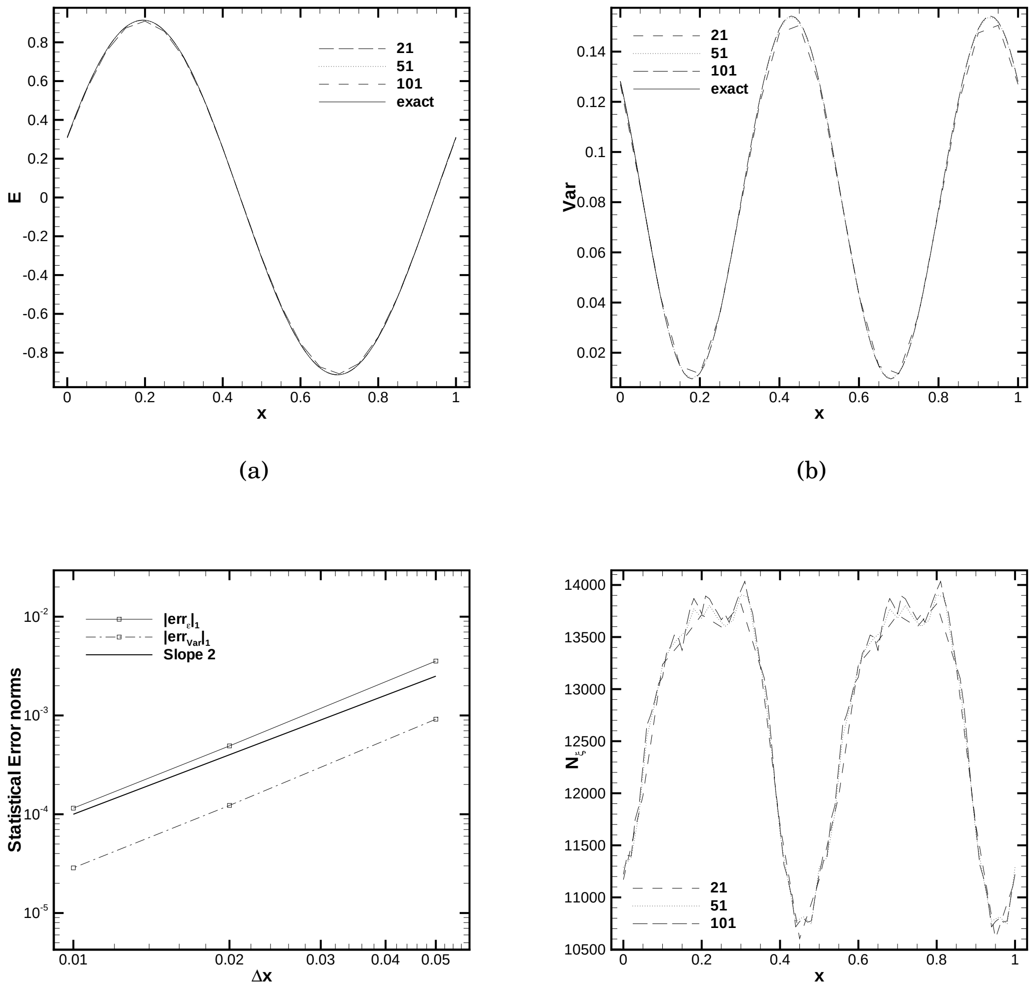
(b)

(c)

Figure 4.15: Error on the statistics (a), the norms L^1 and L^2 (b) and the adapted cell with $m_0 = 1$, $m_L = 7$ and $\varepsilon = 10^{-4}$ of the 3D *discontinuous* function (4.23).

and $\Delta t = 1 \times 10^{-3}$ for the equally spaced meshes with N_x, equal to 51, 101, 201 and 501, points. In this case the limiter is the Roe's superbee limiter (1.28). The reference (numerical) solution is obtained employing the scheme with the following parameters: $N_x = 1001$ physical points and the aSI scheme on uniform stochastic mesh with $N_\xi = 64 \times 64$ cells (this is equivalent to the level $m_L = 6$). The aSI scheme is applied with $m_0 = 1$, $m_L = 6$ and $\varepsilon = 10^{-4}$. The error L^1 norm for the statistics, the expected value and the variance, over the whole physical domain, and the evolution of the number of stochastic cells employed, for each physical location, are reported in 4.17.

From the figure 4.17 it is evident that the scheme converges to the reference solution with the expected second order rate of convergence and the solutions with 201 and 501 physical points appear nearly indistinguishable from the reference solution for the expected value 4.17(a) while the difference is more evident for the variance

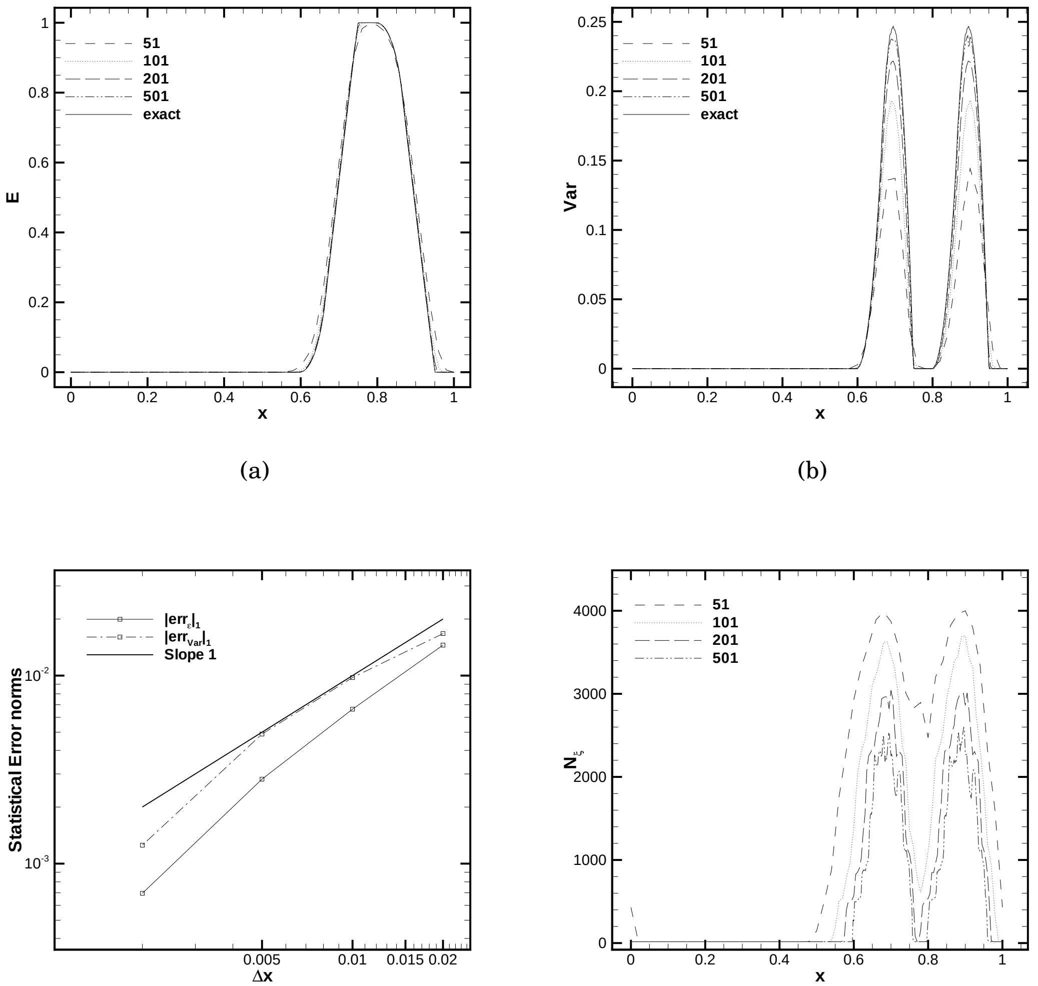


(a) (b)

(c) (d)

Figure 4.16: The expected value (a) and the variance (b) for the linear advection problem with smooth initial condition (4.26) are reported for different physical space resolutions over Ω. The statistic errors in L^1 norms, computed with respect the exact analytical solution following (4.29), are reported in (c). The number of stochastic cells N_ξ at each physical location is reported, for all the spatial resolutions, in (d).

4.17(b). The number of stochastic cells over the physical space 4.17(d) reveals the effect of the representation in the combined physical/stochastic space: finer grids in the physical space allow sharper representations of the discontinuities even in the stochastic space. The computational effort can be concentrated in these high-gradients regions, in the stochastic space, obtaining a narrow region, over the physical space, in which a large number of stochastic cells needs to be placed. For instance, comparing the solution with 51 and 501 physical points the maximum number of stochastic cells required is about $4\,000$ and $2\,600$ respectively. Moreover, the high-gradient region in the stochastic space, for the case of 51 points, smear in a very



(a) (b)

(c) (d)

Figure 4.17: The expected value (a) and the variance (b) for the linear advection problem with random advection velocity (4.28) are reported for different physical space resolutions over Ω. The statistic errors in L^1 norms, computed with respect the numerical reference solution ($N_x = 1001$) following (4.29), are reported in (c). The number of stochastic cells N_ξ at each physical location is reported, for all the spatial resolutions, in (d).

broad region leaving the domain from the right side and re-entering through the left boundary due to the periodic boundary conditions. The aSI scheme, as already noted in the paper ***P4*** for the 1D stochastic case, results to be more and more efficient as the physical resolution increases. In this section the classical linear advection problem has been presented, even in the presence of a non-linear random velocity in the stochastic space. A classical scalar non linear case, for hyperbolic problems, is the inviscid Burgers equation. This case is presented in the following section.

The inviscid Burgers equation

In this section the aSI scheme is applied to the solution of the inviscid Burgers equation

$$\frac{\partial u(x,\boldsymbol{\xi},t)}{\partial t}+\frac{\partial f(u(x,\boldsymbol{\xi},t))}{\partial x}=0 \quad x\in\Omega \quad \text{and} \quad t\in[0,t_F], \tag{4.31}$$

where the random flux function is defined as $f=f(u(x,\xi,t))=\frac{1}{2}u^2(x,\xi,t)$ and $\boldsymbol{\xi}\in\Xi\subset\mathbb{R}^2$. Each of the random parameter is uniformly distributed $\xi_i\sim\mathcal{U}(0,1)$ with $i=1,2$. The physical domain considered here is $\Omega=[0,1]$ with wall boundary conditions. The time line is discretized by means of $N_t=600$ intervals of equal length $\Delta t=7.5\times10^{-4}$. The following initial condition is employed

$$u(x,\boldsymbol{\xi},0)=\begin{cases} 0 & \text{if} \quad x<\dfrac{1}{10} \\ \dfrac{2}{5}\xi_1^2+\dfrac{1}{10000}\xi_1+\dfrac{9}{10} & \text{if} \quad \dfrac{1}{10}\leq x\leq\dfrac{1}{2} \\ \dfrac{1}{5}\xi_2 & \text{if} \quad x>\dfrac{1}{2} \end{cases} \tag{4.32}$$

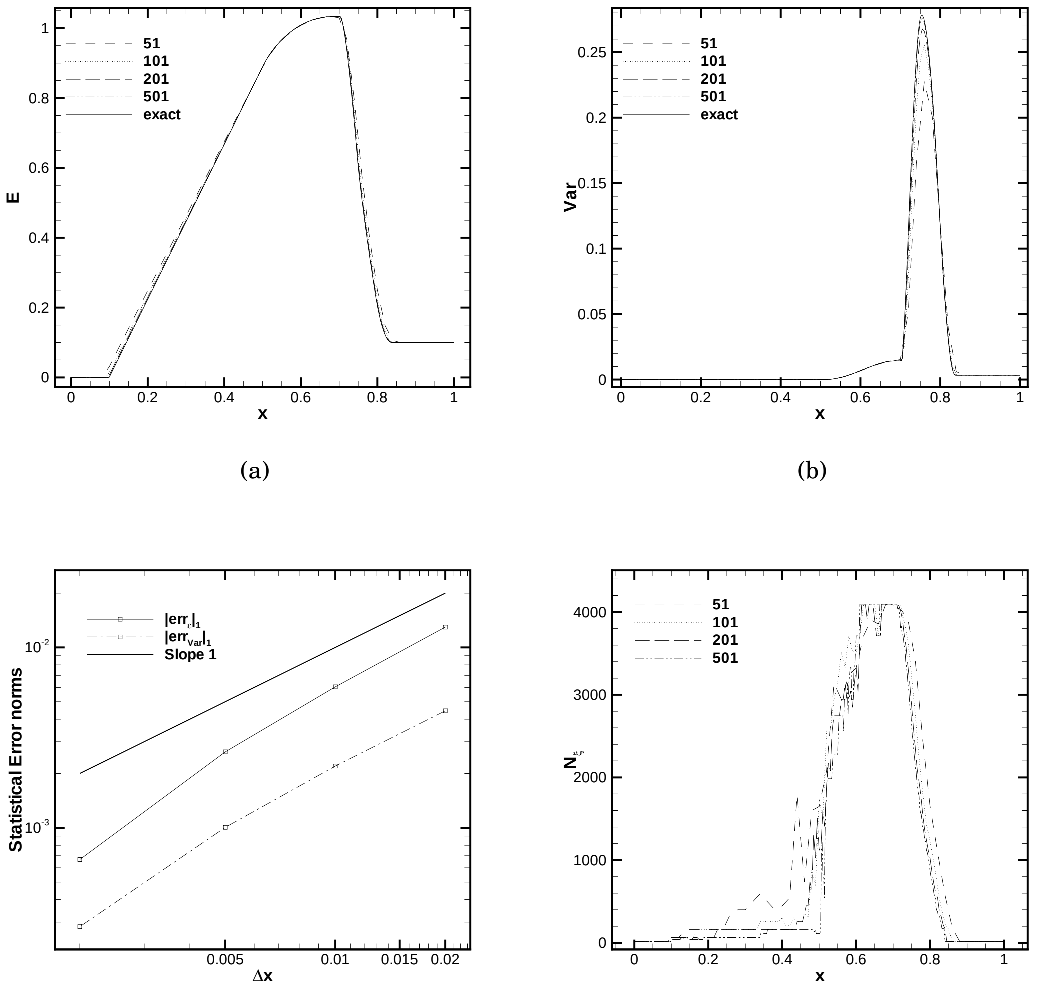
This initial condition is the extension of the test case presented in the paper ***P4*** where, its 1D stochastic counterparts, is employed to obtain a stochastic exact solution even for shocked problems. Here the problem is solved for increasing spatial resolutions corresponding to equally spaced meshes with N_x, 51, 101, 201 and 501, points. For this test case, the limiter is the Roe's superbee (1.28). The aSI scheme is applied with the following parameters $m_0=1$, $m_L=6$ and $\varepsilon=10^{-4}$. A numerical reference solution is computed, without compression, on a resolution level corresponding to $m_L=6$ with $N_\xi=64\times64$ stochastic cells on a physical mesh of 1 001 equally spaced points. The initial condition (4.32) is chosen to produce a random shock front for $t>0$ on the right side of the hat, while a random expansion fan is produced at the left side (see the paper ***P4*** for more details). The error norms in L^1, for both the expected value and the variance, are presented in the figure . The scheme exhibits the expected rate of convergence even if the adaptive procedure refines and coarsens the mesh in the stochastic space according to the local regularity of the function. The expected value 4.18(a) and the variance 4.18(b) are also reported for the different spatial resolutions. The expected value for the solutions on the mesh of 201 and 501 points is almost indistinguishable from the reference solution while, for the variance, the effect of the physical discretization is more evident. Even in this case, higher spatial resolutions, correspond to narrow regions of high number stochastic cells. This is evident from the figure 4.18(d) where the evolution of the stochastic cells is reported over the physical domain.

All the test cases, presented until this points, are scalar test cases. In the next section the Euler equations of the gasdynamics are solved to obtain the statistics of the vector of conservative variables.

Euler equations of the gasdynamics

In this section a vectorial case is presented. The Euler equations of the gas dynamics are solved here to analyze a classical (random) shock tube problem. The problem can be modeled by the well-known 1D Euler equations

$$\frac{\partial\mathbf{u}}{\partial t}+\nabla\cdot\mathbf{f}(\mathbf{u})=\mathbf{0} \tag{4.33}$$



(a)

(b)

(c)

(d)

Figure 4.18: The expected value (a) and the variance (b) for the inviscid Burgers equation with random initial condition (4.32) are reported for different physical space resolutions over Ω. The statistic errors in L^1 norms, computed with respect the numerical reference solution ($N_x = 1001$) following (4.29), are reported in (c). The number of stochastic cells N_ξ at each physical location is reported, for all the spatial resolutions, in (d).

where the vector of conservative variables, the density ρ, the momentum $m = \rho u$ and the total Energy E^t, $\mathbf{u} \in \mathbb{R}^3$ and the flux vector $\mathbf{f}(\mathbf{u}) \in \mathbb{R}^3 \to \mathbb{R}^3$ are

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ E^t \end{pmatrix} \quad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} m \\ \dfrac{m^2}{\rho} + \Pi(\mathbf{u}) \\ \dfrac{m}{\rho}\left(E^t + \Pi(\mathbf{u})\right). \end{pmatrix} \tag{4.34}$$

The pressure $\Pi(\mathbf{u})$ (as function of the conservative variables) can be derived for a polytropic ideal gas [70] as follows

$$\Pi(\mathbf{u}) = (\gamma - 1)\left(E^t - \frac{1}{2}\frac{|m^2|}{\rho}\right). \tag{4.35}$$

The initial condition, for the uncertain shock tube problem, is derived from the classical Sod test case [76], where both the density and the pressure at the left side of the diaphragm are uncertain. The left state ($x < x_d$ for $t = 0$) and the right side are defined as follows:

$$\mathbf{u}_L(x, \boldsymbol{\xi}, t) = \begin{pmatrix} \rho_L(\xi_1) \\ 0 \\ \dfrac{p_L(\xi_2)}{\gamma - 1} \end{pmatrix} \quad \mathbf{u}_R(x, t) = \begin{pmatrix} \rho_R \\ 0 \\ \dfrac{p_R}{\gamma - 1}. \end{pmatrix} \tag{4.36}$$

In particular, the density on the left state is dependent from an uniformly distributed random parameter $\xi_1 \sim \mathcal{U}[0,1]$: $\rho_L(\xi_1) = 0.9 + 0.4\xi_1$. The values of the pressures is instead dependent from the random parameter $\xi_2 \sim \mathcal{U}[0,1]$ as $\rho_L(\xi_2) = 0.8 + 0.4\xi_2$. The other data are the right pressure $p_R = 0.1$ and the right value of the density is $\rho_R = 0.125$. The total energy E^t is obtained (considering the gas at the rest in the whole domain) as a function of the local pressure and the ratio between specific heats, that for a diatomic gas can be assumed equal to $\gamma = 1.4$. The initial position of the diaphragm is fixed to $x_d = 0.42$.

Simulations are performed, over a physical domain $\Omega = [0,1]$, for $N_t = 3\,900$ time steps of length $\Delta t = 6.25 \times 10^{-5}$. The simulations are carried out over equally spaced meshes of N_x, 51, 101, 201 and 501, points employing the aSI scheme based on the MHM with a van Leer limiter (see equation (1.29)). The reference solution is obtained over a physical mesh of $1\,001$ equally spaced nodes and a stochastic mesh of $N_\xi = 32 \times 32$ cells. The aSI scheme is applied with the following parameters $m_0 = 1$, $m_L = 6$ and $\varepsilon = 10^{-3}$. The statistics, expected values and variances, for all the conservative variables are reported in the figure 4.19.

For the Euler test case the influence of the spatial resolution is more evident: the error on the maximum values for the variance of all the conservative variables is quite high for the 51 points mesh. The situation improves rapidly obtaining very close values with respect to the reference solution for the 501 case. The error norms computed in L^1 following (4.29), for all the conservative variables, for both expected value and variance, with respect to the reference solution, are reported in the figure 4.20(a). In the figure 4.20(b) the evolution of the number of points over the physical space is reported for all the physical resolutions. Even here, as in the previous cases, the higher physical resolution allows to recover a sharp representation of the high-gradients and shocks regions in the combined space and a local computational effort more concentrated in their neighboring.

Another test case is performed to make evident the ability of the aSI scheme. The Euler problem is solved for a different initial condition: the initial position of the diaphragm is placed at $x_d = 0.65$. The simulation is carried out for $N_t = 6000$ time steps of length $\Delta t = 6.25 \times 10^{-5}$ to obtain the reflection of the random shock waves at the right boundary (wall boundary conditions are employed). The physical mesh is constituted by 401 equally spaced points in Ω and the parameter for the aSI scheme are $m_0 = 1$, $m_L = 6$ and $\varepsilon = 10^{-3}$. Some results, for three different time steps, are reported in the figure 4.21. In the left column, the adapted mesh in the overall $x - \xi_1 - \xi_2$ space is shown, while in the right one the number of stochastic cells, over the whole physical space is reported. In particular the first time step corresponds to a time of $t = 0.1875$ ($N_t = 3000$) 4.21(a) where the front of the shock waves (well captured) start to interact with the right boundary and reflects. This is evident in the upper corner at the right boundary in the figure 4.21(a) where each point represents the center of a cell, in the overall $\Omega - \Xi$ space, and its color is the local value of the conditional expected value for the density $\mathbb{E}(\rho(x_i)|\Xi_j^L)$. In the figure 4.21(b) the aSI scheme captures the three main structures that are, starting from left, the expansion fans, the contact discontinuities and the shock fronts. At the second time step $t = 0.2875$ ($N_t = 4600$) 4.21(c) the shock front is totally reflected and, in the upper domain ($\xi_2 > 0.5$ roughly), already interacted with the incoming contact discontinuities front. In the figure 4.21(d) the corresponding evolution of the number of stochastic cells over Ω is reported and the aSI scheme is demonstrated to be able coarsening, the local stochastic mesh, just after the reflection of the shock front from the boundary. At the final time step the shock front entirely passed through the contact discontinuity front and appears well captured 4.21(f). The contact discontinuity front starts to interact with the right boundary and it is partially reflected. In this region, the number of stochastic cells is quite high due to the presence of not yet well separated structures. However, it is evident the ability of the aSI scheme in tracking the different structures capturing also new high-gradient regions when forming 4.21(f).

4.2.3 Preliminary 3D results: the smooth advection case

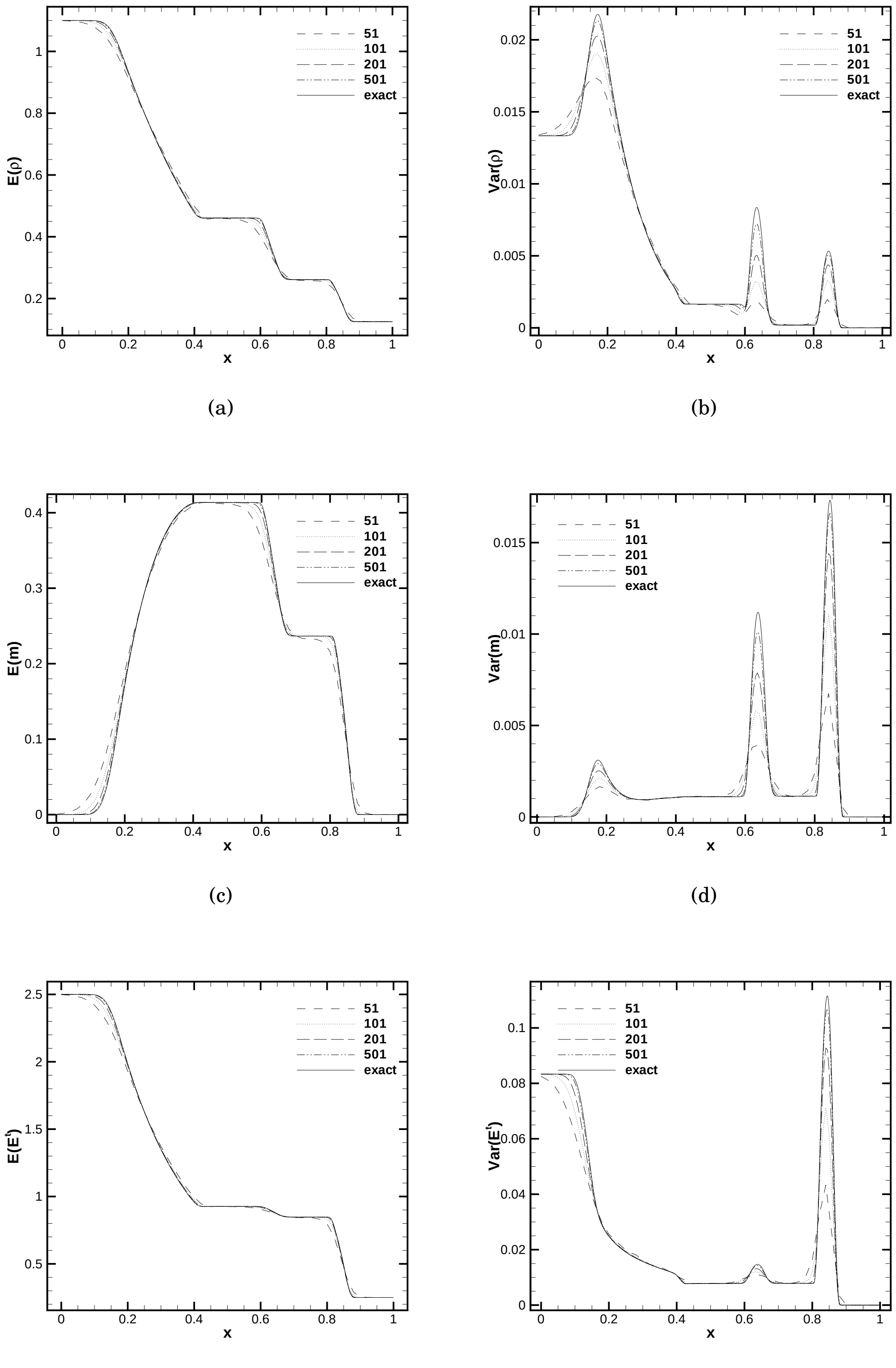
In this section some preliminary results for the linear advection equation are presented. The problem is formulated as already discussed in the previous section (see equation (4.25)). The stochastic problem is obtained considering the following random initial condition

$$u_0(x, \boldsymbol{\xi}) = \sin(2\pi x + \xi_1^2 + \xi_2^2 + \xi_3^2), \tag{4.37}$$

where each random parameter is $\xi_i \sim \mathcal{U}(0,1)$ for $i = 1, 2, 3$. As the 2D problem the advection velocity is fixed to $a = 0.1$. The simulation is carried out for $N_t = 500$ time steps of constant length $\Delta t = 5 \times 10^{-3}$. Even in this case, the exact reference solution can be obtained by integration of the solution over the physical space at each physical cell location

$$u(x, \boldsymbol{\xi}, t_F) = \sin\left(2\pi(x - at_f) + \xi_1^2 + \xi_2^2 + \xi_3^2\right)$$
$$\longrightarrow \begin{cases} \mathbb{E}_{\text{ex}}(\bar{u}_i(t_F)) = \displaystyle\int_\Xi u(x_i, \boldsymbol{\xi}, t_F)\mathrm{d}\mu(\boldsymbol{\xi}) \\ \mathrm{Var}_{\text{ex}}(\bar{u}_i(t_F)) = \displaystyle\int_\Xi u^2(x_i, \boldsymbol{\xi}, t_F)\mathrm{d}\mu(\boldsymbol{\xi}) - (\mathbb{E}_{\text{ex}}(\bar{u}_i(t_F)))^2 \end{cases} \tag{4.38}$$

The aSI scheme is applied with increasing physical meshes of N_x equal to 11, 21, 51 and 101 equally spaced points and the following parameters $m_0 = 1$, $m_L = 5$ and $\varepsilon = 10^{-2}$. As evident, due to the computational cost of the code not heavily parallelized, the maximum resolution level it is not fine enough and the threshold is higher than the previous 2D test cases. The effect is evident in the figure 4.22(a) where the norms L^1 for both the statistics, expected value and variance, are reported following the definitions (4.29). The rate of convergence very soon decreases and, already with 51 and 101 physical points, the error related to the stochastic space discretization is so high to produce a stagnation of the physical space error. These preliminary results are added to the present manuscript to demonstrate the possibility to extend the analysis to 3D stochastic cases with time-dependent adaptivity in the combined space, as evident from the figure 4.22(d) where the evolution of the stochastic cells over Ω is reported. In the figures 4.22(a) and 4.22(b) the expected value and the variance, of the solution, are reported respectively. For the mean the solutions are almost indistinguishable from the exact solution for the 51 and 101 cases. However, the error on the variance is higher: in the peaks regions the differences with respect the analytical solutions are evident even for the mesh with 101 physical points.



(a)

(b)

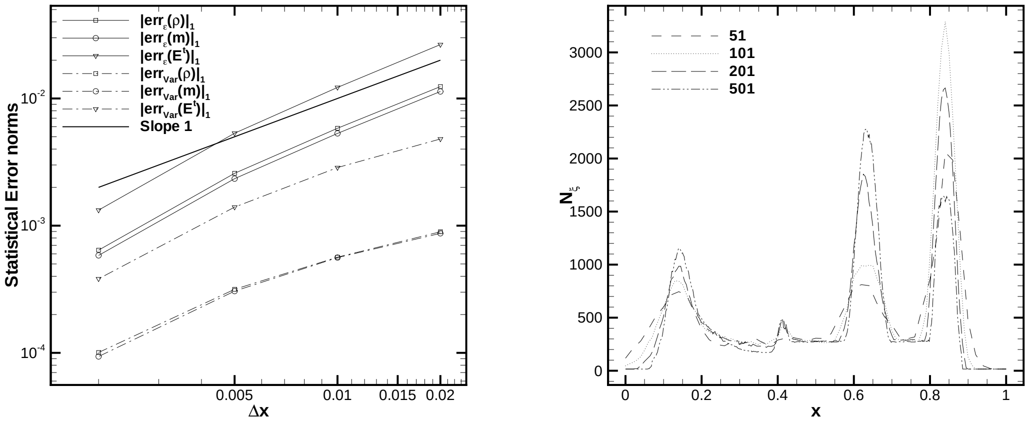
(c)

(d)

(e)

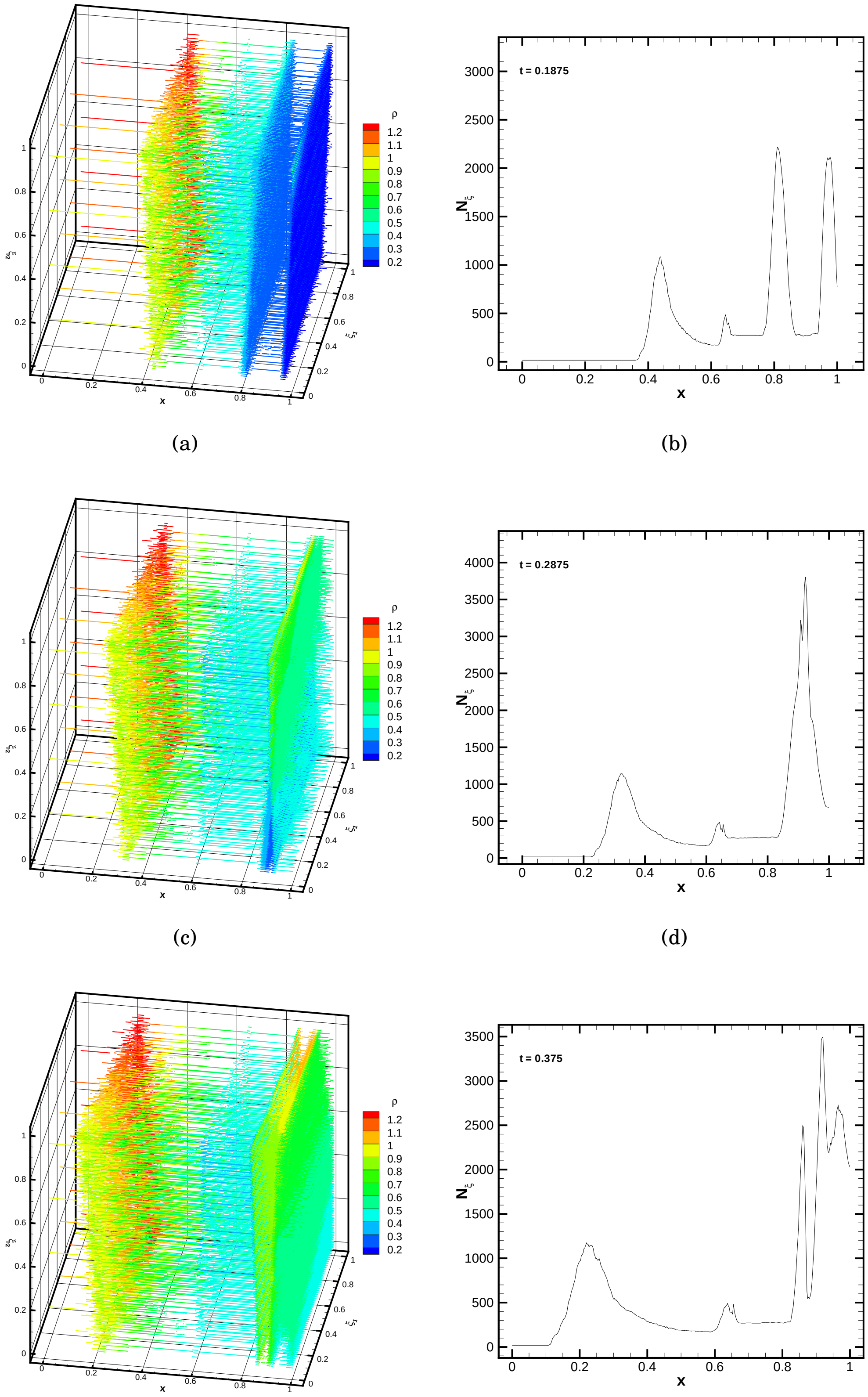
(f)

Figure 4.19: Expectancies for the shock tube (4.36) for the density (a), momentum (c) and total energy (e). Their variances are reported respectively in (b), (d) and (f).



(a) (b)

Figure 4.20: The statistical L_1 norms, computed with respect the numerical reference solution ($N_x = 1001$) following (4.29), are reported in (a). The numbers of stochastic cells N_ξ at each physical location are reported for all the spatial resolutions in (b).



(a)

(b)

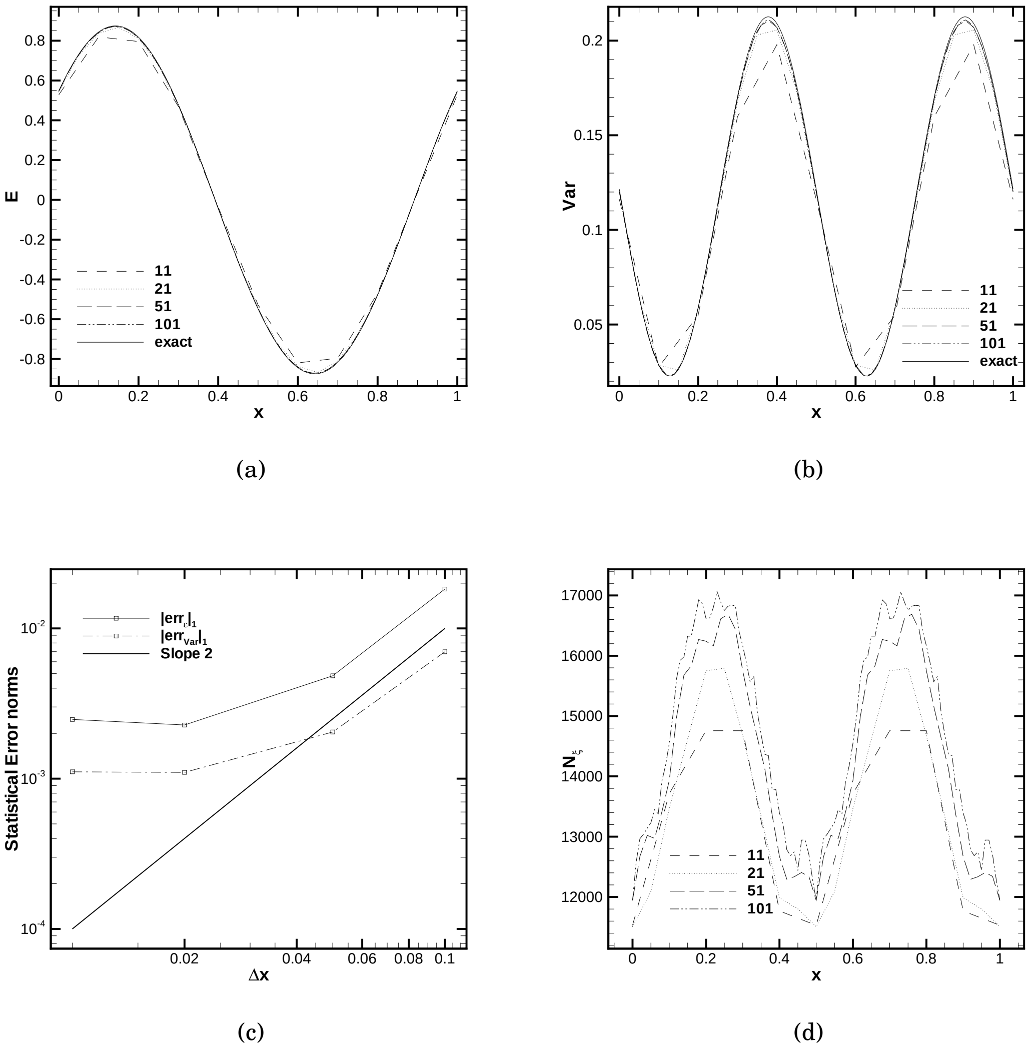
(c)

(d)

(e)

(f)

Figure 4.21: Center of the cells and number of stochastic cells over Ω for $N_t = 3000$ (a) and (b), $N_t = 4600$ (c) and (d) and $N_t = 6000$ (e) and (f). The color of each cell is based on the value of the conditional expected value for the density $\mathbb{E}(\rho(x_i)|\Xi^L_j)$.



(a)

(b)

(c)

(d)

Figure 4.22: The expected value (a) and the variance (b) for the linear advection equation with smooth random initial condition (4.37) are reported for different physical space resolutions over Ω. The statistic errors in L^1 norms, computed with respect the analytical reference solution following (4.29), are reported in (c). The number of stochastic cells N_ξ at each physical location is reported, for all the spatial resolutions, in (d).

CHAPTER 5

Concluding remarks and perspectives

This thesis deals with efficient weakly intrusive numerical schemes, based on Harten multiresolution framework, to propagate uncertainties in the context of stochastic differential equations. The use of multiresolution framework has a twofold aim. First, there is the reduction of the dimensionality of the discrete space of function representation, defined in a proper stochastic space. This reduction permits a gain in terms of computational efficiency, *i.e.* by reducing the number of explicit evaluations required to represent the function. Multiresolution constitutes also the basis for representing the function by exploiting the locality of this basis. Moreover, the multiresolution analysis offers a natural tool to investigate the local regularity of a function and can be employed to build an efficient refinement strategy. Second aim is a general procedure to refine/coarsen the stochastic space in unsteady problems. The strategy has been verified, by means of several test cases, on both ordinary and partial stochastic equations. Main feature, of the proposed approach, is the possibility to adapt the discretization of the stochastic space, in function of the physical and time coordinates. This approach suits very well fluid dynamics applications, where the problems are intrinsically multi-scale and also non-linear with traveling high-gradient/discontinuities. This strategy is able to capture and follow all types of flow structures.

Let us describe the main differences of the proposed approach, with respect to the other multiresolution/multiwavelet techniques proposed in literature. The approach presented here, is not limited to *intrusive* propagation of uncertainties and is not limited to independent random parameters. In principle, any stochastic space can be employed, even containing holes. However, in this thesis, the random parameters are always considered independent and no holes can be present. The possibility to consider general geometries, for the stochastic space, represents a great difference with respect to more classical multiresolution/multiwavelet approaches present in literature. These latter could require the solution of a dilating equation to obtain the basis of approximation; this solution, on general geometries, could be cumbersome. Moreover, the classical multiresolution approaches, present in literature, are limited to linear scheme, *i.e.* data-independent basis. However, it is well demonstrated in literature, and it is shown in some numerical test cases in this thesis, the superiority of the non-linear multiresolution approaches in terms of compression capabilities, specially for non-smooth problems.

Finally, the present approach opens a novel way to build efficient *intrusive* techniques. Its generality well suits multi-scales applications not limited to fluid flows problems, as demonstrated in the numerical section. Another interesting property, with respect to all the other *intrusive* techniques, remains the moderate intrusive behavior, unlike for example in spectral Galerkin projection, where a theoretical manipulation of the original system is needed. This could be of great interest for very complex applications, where the deterministic scheme is already available and cannot be easily modified. In some sense, an example of this kind of flexibility could be found in paper ***P5*** where the aSI scheme is coupled with the Discrete Equation Model.

5.1 Perspectives

The approach presented in this thesis demands several improvements in the future. Concerning the aSI scheme, the first improvement should be devoted to a strong parallelization of the numerical code. The aSI scheme is conceived to be highly parallel due to the external loop on the spatial coordinates. In principle, a number of parallel threads, equal to the number of spatial cells, could be employed. The numerical computation presented in this thesis are performed on 12 processors (in the multidimensional case) employing the automatic parallelization of the Intel® `ifort` compiler (`-parallel`). This approach could be heavily improved by a parallelization 'by hand' of the code.

Potentialities of the aSI approach could be exploited when treating unsteady and discontinuous probability density functions, as already made for the sTE scheme. The other key distinctive element, with respect to the other approaches, *i.e.* the possibility to employ dependent random variables should be further exploited.

Both the compression capabilities and the accuracy of the scheme (in the stochastic space) could be, in a straightforward manner, extended employing an high-order polynomial reconstruction. To obtain a multidimensional conservative reconstruction of high-order, and possibly non-linear (ENO/WENO), two different approaches seem to be feasible. The first one should be to recover the stencil on the resolution level to which the cell belong. Otherwise, the stencil could be built employing the non-conformal mesh obtained at a certain resolution level employing an algorithm of 'crystal grown' type, *i.e.* moving from each cluster of cells and adding the neighboring one allowing the least oscillatory reconstruction. These two possibilities have both advantages and disadvantages and should be verified before on the so-called steady functions.

Another interesting possibility, potentially capable to greatly increase the efficiency of the scheme, is the anisotropic refinement of the mesh. Limiting to the stochastic space, the importance of the different dimensions could be locally different. In literature, several *criteria* exist for the anisotropic refinement in different context (also in the UQ framework) and a preliminary study of the literature should be performed to choose the best option.

The final goal of this kind of *intrusive* approach is to obtain very accurate numerical schemes by performing a strong coupling between the physical and stochastic spaces, with an unsteady procedure for refinement (coarsening). Of course this possibility, even if very attractive, remains very complex from both the theoretical and algorithmic point-of-view.

Bibliography

[1] Remi Abgrall. Design of an essentially non-oscillatory reconstruction procedure on finite-element type meshes. Technical report, INRIA, 1992.

[2] Remi Abgrall. On Essentially Non-oscillatory Schemes on Unstructured Meshes: Analysis and Implementation, September 1994.

[3] Remi Abgrall. A simple, flexible and generic deterministic approach to uncertainty quantifications in non linear problems: application to fluid flow problems. Technical report, 2007.

[4] Rémi Abgrall and Pietro Marco Congedo. A semi-intrusive deterministic approach to uncertainty quantification in non-linear fluid flow problems. *Journal of Computational Physics*, 235:828–845, February 2013.

[5] Rémi Abgrall, Pietro Marco Congedo, Stephane Galéra, and Gianluca Geraci. Semi-intrusive and non-intrusive stochastic methods for aerospace applications. In *4TH EUROPEAN CONFERENCE FOR AEROSPACE SCIENCES, EUCASS 2011. July 2011 Saint-Petersburg, Russia*, number 1, pages 1–8, Saint petersbourg, 2011.

[6] Remi Abgrall and Ami Harten. Multiresolution Representation in Unstructured Meshes. *SIAM Journal on Numerical Analysis*, 35(6):2128–2146, 1998.

[7] Rémi Abgrall and Vincent Perrier. Asymptotic Expansion of a Multiscale Numerical Scheme for Compressible Multiphase Flow. *Multiscale Modeling & Simulation*, 5(1):84–115, January 2006.

[8] Rémi Abgrall and Richard Saurel. Discrete equations for physical and numerical compressible multiphase mixtures. *Journal of Computational Physics*, 186(2):361–396, April 2003.

[9] Rémi Abgrall and T Sonar. On the use of Mühlbach expansions in the recovery step of ENO methods. *Numerische Mathematik*, (1997):1–25, 1997.

[10] Nitin Agarwal and N.R. Aluru. A domain adaptive stochastic collocation approach for analysis of MEMS under uncertainties. *Journal of Computational Physics*, 228(20):7662–7688, November 2009.

[11] AIAA. *Guide for the verification and validation of computational fluid dynamics simulations*. 1998.

[12] Sergio Amat, F Aràndiga, Albert Cohen, and Rosa Donat. Tensor product multiresolution analysis with error control for compact image representation. *Signal Processing*, 82:587–608, 2002.

[13] Sergio Amat, S. Busquier, and J.C. Trillo. Nonlinear Harten's multiresolution on the quincunx pyramid. *Journal of Computational and Applied Mathematics*, 189(1-2):555–567, May 2006.

[14] JD Anderson. *Computational Fluid Dynamics. The Basics with Applications.* McGraw-Hill, 1995.

[15] F. Aràndiga, a. M. Belda, and P. Mulet. Point-Value WENO Multiresolution Applications toStable Image Compression. *Journal of Scientific Computing*, 43(2):158–182, February 2010.

[16] F Arandiga and Rosa Donat. Nonlinear multiscale decompositions: The approach of A. Harten. *Numerical Algorithms*, 23:175–216, 2000.

[17] Francesc Aràndiga, G. Chiavassa, and Rosa Donat. Harten framework for multiresolution with applications: From conservation laws to image compression. *Boletín SEMA*, 31(31):73–108, 2009.

[18] Francesc Aràndiga, Rosa Donat, and Ami Harten. Multiresolution based on weighted averages of the hat function I: Linear reconstruction techniques. *SIAM Journal on Numerical Analysis*, 36(1):160–203, 1998.

[19] Ivo Babuška, Fabio Nobile, and Raul Tempone. A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. *SIAM Journal on Numerical Analysis*, 45(3):1005, 2007.

[20] DS Bale, Randall J. LeVeque, S Mitran, and JA Rossmanith. A wave propagation method for conservation laws and balance laws with spatially varying flux functions. *SIAM Journal on Scientific Computing*, 24(3):955–978, 2003.

[21] Richard Ernest Bellman and Bellmann Richard. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[22] Barna L. Bihari and Ami Harten. Multiresolution Schemes for the Numerical Solution of 2-D Conservation Laws I. *SIAM Journal on Scientific Computing*, 18(2):315, 1997.

[23] Russel E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

[24] R.H. Cameron. Transformations of Weiner Integrals Under Translations. *The Annals of Mathematics*, 45(2):386–396, 1944.

[25] Tonkid Chantrasmi and Gianluca Iaccarino. Computing Shock Interactions under Uncertainty. *AIAA Paper 2009-2284*, (May):1–14, 2009.

[26] A Cohen, I Daubechies, and P Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1:54–81, 1993.

[27] Thierry Crestaux, Olivier Le Maître, and Jean-Marc Martinez. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7):1161–1172, July 2009.

[28] Ingrid Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics.

[29] Bert J. Debusschere, Habib N. Najm, Philippe P. Pebay, Omar M. Knio, Roger G. Ghanem, and Olivier Le Maître. Numerical Challenges in the Use of Polynomial Chaos Representations for Stochastic Processes. *SIAM Journal on Scientific Computing*, 26(2):698, 2004.

[30] G. Deodatis. Weighted Integral Method. I: Stochastic Stiffness Matrix. *Journal Eng. Mech.*, 117(8):1851–1864, 1991.

[31] Alireza Doostan and Houman Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. *Journal of Computational Physics*, 230(8):3015–3034, April 2011.

[32] J.H. Ferziger and M. Peric. *Computational Methods for Fluid Dynamics*. Springer Verlag, 2002.

[33] Marc Gerritsma, Jan-Bart van der Steen, Peter Vos, and George Em Karniadakis. Time-dependent generalized polynomial chaos. *Journal of Computational Physics*, 229(22):8333–8363, November 2010.

[34] Pascal Getreuer and Francois G. Meyer. ENO multiresolutions Schemes with General Discretizations. *SIAM Journal on Numerical Analysis*, 46(6):2953–2977, 2008.

[35] Roger G. Ghanem and Pol D. Spanos. *Stochastic Finite Elements. A spectral approach*. Springer Verlag, 1991.

[36] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics*, 230(10):3668–3694, February 2011.

[37] F.H. Harlow and J.E. Fromm. Computer experiments in fluid dynamics. *Scientific American*, 212(3):104–110, 1965.

[38] Ami Harten. Discrete multi-resolution analysis and generalized wavelets, May 1993.

[39] Ami Harten. Adaptive multiresolution schemes for shock computations. *Journal of Computational Physics*, 115(2):319–338, August 1994.

[40] Ami Harten. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Communications on Pure and Applied Mathematics*, 48(12):1305–1342, 1995.

[41] Ami Harten. Multiresolution Representation of Data : A General Framework. *SIAM Journal on Numerical Analysis*, 33(3):1205–1256, 1996.

[42] Ami Harten, Bjorn Engquist, and Stanley Osher. Uniformly high order accurate essentially non-oscillatory schemes, III. *Journal of Computational Physics*, 71(2):231–303, August 1987.

[43] Charles Hirsch. *Numerical computation of internal and external flows*. Elsevier, 2007.

[44] Gianluca Iaccarino. Quantification of Uncertainty in Flow Simulations Using Probabilistic Methods. Technical report, 2008.

[45] Gianluca Iaccarino. Introduction to Uncertainty Quantification. In *RTO-AVT-VKI Short COurse, Von Karman Institute, Belgium*, 2011.

[46] GS Jiang and CW Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 228(126):202–228, 1996.

[47] M. Kleiber and T.D. Hien. *The Stochastic Finite Element Method*. John Wiley & Sons Ltd, 1992.

[48] Andrey Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. 1933.

[49] Olivier Le Maître. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *Journal of Computational Physics*, 197(2):502–531, July 2004.

[50] Olivier Le Maître. Uncertainty propagation using WienerHaar expansions. *Journal of Computational Physics*, 197(1):28–57, June 2004.

[51] Olivier Le Maître and Omar M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Springer Verlag, 2010.

[52] Bram Van Leer. Upwind and high-resolution methods for compressible flow: From donor cell to residual-distribution schemes. *Communications in Computational Physics*, 1(2):192–206, 2006.

[53] Randall J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser Basel, Basel, 1992.

[54] Randall J. LeVeque. Finitevolume methods for nonlinear elasticity in heterogeneous media. *International Journal for Numerical Methods in Fluids*, (May 2001):93–104, 2002.

[55] RJ LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[56] P. Lipton. Testing hypotheses: prediction and prejudice. *Science*, 307:219–221, 2005.

[57] Xu-Dong Liu, Stanley Osher, and Tony Chan. Weighted Essentially Non-oscillatory Schemes. *Journal of Computational Physics*, 115(1):200–212, 1994.

[58] Yuan Liu and Yong-Tao Zhang. A Robust Reconstruction for Unstructured WENO Schemes. *Journal of Scientific Computing*, 54(2-3):603–621, May 2012.

[59] By D Lucor, J Witteveen, P Constantine, D Schiavazzi, and G Iaccarino. Comparison of adaptive uncertainty quantification approaches for shock wave-dominated flows. In *Center For Turbulence Research, Prooceedings of the Summer Program 2012*, pages 219–228, 2012.

[60] Didier Lucor. Spectral and High Order Methods for Partial Differential Equations. In Jan S. Hesthaven and Einar M. Rø nquist, editors, *Spectral and High Order Methods for Partial Differential Equations. Lecture Notes in Computational Science and Engineering*, volume 76 of *Lecture Notes in Computational Science and Engineering*, pages 293–307. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[61] Didier Lucor. *Stochastic Spectral Approach to Uncertainty Quantification of Computational Fluid Dynamics*. Habilitation à Diriger des Recherches, 2011.

[62] Didier Lucor and George Em Karniadakis. Predictability and uncertainty in flowstructure interactions. *European Journal of Mechanics - B/Fluids*, 23(1):41–49, February 2004.

[63] Xiang Ma and Nicholas Zabaras. An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. *Journal of Computational Physics*, 228(8):3084–3113, May 2009.

[64] Nicolas Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[65] WL Oberkampf and CJ Roy. *Verification and validation in scientific computing*. 2010.

[66] Per Pettersson, Gianluca Iaccarino, and Jan Nordström. An intrusive hybrid method for discontinuous two-phase flow under uncertainty. *Computers & Fluids*, 86:228–239, November 2013.

[67] Gaël Poëtte, Bruno Després, and Didier Lucor. Uncertainty quantification for systems of conservation laws. *Journal of Computational Physics*, 228(7):2443–2467, 2009.

[68] Gaël Poëtte and Didier Lucor. Non intrusive iterative stochastic spectral representation with application to compressible gas dynamics. *Journal of Computational Physics*, 231(9):3587–3609, May 2012.

[69] Stephen B. Pope. *Turbulent flows*. Cambridge University Press, 2000.

[70] Luigi Quartapelle and Franco Auteri. *Fluidodinamica comprimibile*. Casa Editrice Ambrosiana, 2013.

[71] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical Mathematics*. Springer Verlag, 2000.

[72] By D Schiavazzi, A Doostan, and G Iaccarino. Sparse multiresolution stochastic approximation for uncertainty quantification. In *Center For Turbulence Research, Annual Research Briefs*, pages 93–102, 2012.

[73] S. Schlesinger. Terminology for model credibility. *Simulation*, 32(3):103–104, 1979.

[74] S. A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Math. Dokl.*, (4):240–243, 1963.

[75] Ilya M. Sobol. Distribution of points in a cube and approximate evaluation of integrals. *U.S.S.R Comput. Maths. Math. Phys.*, 7:86–112, 1967.

[76] G.A. Sod. Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws. *Journal of Computational Physics*, 27:1–31, 1978.

[77] James R. Stewart. Simulation Based Predictive Science. VKI Lecture series, 2011.

[78] Gilbert Strang. Wavelets and dilation equations: A brief introduction. *SIAM review*, 31(4):614–627, 1989.

[79] W Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, 1998.

[80] J.C. Tannehill, D.A. Anderson, and R.H. Pletcher. *Computational Fluid Mechanics And Heat Transfer*. Taylor&Francis, 1997.

[81] Terry Bahill and Steven J. Henderson. Requirements development, verification, and validation exhibited in famous failures. *Systems Engineering*, 8(1):1–14, 2005.

[82] EF Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Verlag, 2009.

[83] J Tryoen, Olivier Le Maître, M Ndjinga, and A Ern. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems q. *Journal of Computational Physics*, 229:6485–6511, 2010.

[84] Julie Tryoen. *Methodes de Galerkin stochastiques adaptatives pour la propagation d'incertitudes parametriques dans les systemes hyperboliques*. PhD thesis, Univeriste Paris-Est.

[85] Julie Tryoen, Olivier Le Maître, M. Ndjinga, and a. Ern. Roe solver with entropy corrector for uncertain hyperbolic systems. *Journal of Computational and Applied Mathematics*, 235(2):491–506, November 2010.

[86] Julie Tryoen, OL Maitre, and A Ern. Adaptive anisotropic spectral stochastic methods for uncertain scalar conservation laws. *SIAM Journal on Scientific Computing*, 34(5), 2012.

[87] Bram van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, 32(1):101–136, 1979.

[88] Xiaoliang Wan and George Em Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 209(2):617–642, November 2005.

[89] Norbert Wiener. The Homogeneous Chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

[90] J.A.S. Witteveen and Gianluca Iaccarino. Simplex Elements Stochastic Collocation in Higher-Dimensional Probability Spaces. In *51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 12 - 15 April 2010, Orlando, Florida*, number April, 2010.

[91] J.A.S. Witteveen and Gianluca Iaccarino. Subcell resolution in simplex stochastic collocation for spatial discontinuities. *Journal of Computational Physics*, 251:17–52, October 2013.

[92] Dongbin Xiu. Efficient collocational approach for parametric uncertainty analysis. *Communications in computational physics*, 2(2):293–309, 2007.

[93] Dongbin Xiu. Fast numerical methods for stochastic computations: a review. *Communications in computational physics*, 5(2):242–272, 2009.

[94] Dongbin Xiu and JS Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 3:1118–1139, 2005.

[95] Dongbin Xiu and George Em Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191(43):4927–4948, September 2002.

[96] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

[97] F. Yamazaki, M. Shinozuka, and G. Dasgupta. Neumann expansion for stochastic finite element analysis. *Journal Eng. Mech.*, 114(8):1335–1354, 1988.

[98] D. Zhang. *Stochastic Methods for Flow in Porous Media*. Academic Press, 2002.

Part II: Papers

Paper *P1*

Journal of Computational Physics 257 (2014) 19–56

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

A one-time truncate and encode multiresolution stochastic framework

R. Abgrall, P.M. Congedo, G. Geraci *

INRIA Bordeaux–Sud-Ouest, Bacchus Team, 200 Avenue de la Vieille Tour, 33405 Talence Cedex, France

ARTICLE INFO

Article history:
Received 23 May 2012
Received in revised form 1 August 2013
Accepted 5 August 2013
Available online 9 October 2013

Keywords:
Multiresolution
Uncertainty quantification
Stochastic collocation
Adaptive grid

ABSTRACT

In this work a novel adaptive strategy for stochastic problems, inspired from the classical Harten's framework, is presented. The proposed algorithm allows building, in a very general manner, stochastic numerical schemes starting from a whatever type of deterministic schemes and handling a large class of problems, from unsteady to discontinuous solutions. Its formulations permits to recover the same results concerning the interpolation theory of the classical multiresolution approach, but with an extension to uncertainty quantification problems. The present strategy permits to build numerical scheme with a higher accuracy with respect to other classical uncertainty quantification techniques, but with a strong reduction of the numerical cost and memory requirements. Moreover, the flexibility of the proposed approach allows to employ any kind of probability density function, even discontinuous and time varying, without introducing further complications in the algorithm. The advantages of the present strategy are demonstrated by performing several numerical problems where different forms of uncertainty distributions are taken into account, such as discontinuous and unsteady custom-defined probability density functions. In addition to algebraic and ordinary differential equations, numerical results for the challenging 1D Kraichnan–Orszag are reported in terms of accuracy and convergence. Finally, a two degree-of-freedom aeroelastic model for a subsonic case is presented. Though quite simple, the model allows recovering some physical key aspect, on the fluid/structure interaction, thanks to the quasi-steady aerodynamic approximation employed. The injection of an uncertainty is chosen in order to obtain a complete parameterization of the mass matrix. All the numerical results are compared with respect to classical Monte Carlo solution and with a non-intrusive Polynomial Chaos method.

1. Introduction

Nowadays, the prediction of the numerical simulations is a fundamental task to attain for the optimization and the control of engineering devices. However, estimating the confidence of a numerical simulation remains very challenging. In recent years, a strong effort has been devoted to develop efficient numerical methods for taking into account the randomness in the numerical simulations.

The most popular and known method for the uncertainty quantification (UQ) is the Monte Carlo. Its development led back to the research on the nuclear devices in the context of the Manhattan project and is due to Fermi, von Neumann and Ulam. This method is based on a stochastic procedure to represent realizations of the solution for which the statistic moments can be computed. Despite its solid mathematical framework it represents a very expensive approach for most

* Corresponding author.
E-mail address: gianluca.geraci@inria.fr (G. Geraci).

0021-9991/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcp.2013.08.006

practical application because it requires a great number of realizations. Several improved versions of the classical Monte Carlo method have been proposed in literature for increasing the convergence rate, see for instance the recent work presented in [1], but they still remain unfeasible for complex problems when the evaluation of samples is expensive, as in most engineering problems.

One of the most important class of methods for UQ is based on the Polynomial Chaos (PC) representation. In the original work of Wiener [2], the solution is expanded in a polynomial Hermite basis, the so-called homogeneous chaos expansion. In recent years, Xiu and Karniadakis [3] demonstrated that the optimal convergence, with respect to non-Gaussian probability distributions, can be achieved if orthogonal basis are chosen following the so-called Wiener–Askey scheme. This leads to the well-known generalized Polynomial Chaos (gPC). Following the procedure introduced by Xiu and Karniadakis, i.e. employing the probability density function as a weight function for searching the orthogonal basis, an optimal expansion basis could be virtually obtained from every kind of pdf, see for details [4,5]. Actually the gPC is often used in combination with Galerkin projection [6] techniques following the idea of Ghanem and Spanos [7], who first extended the applications of the PC in combination with finite elements. The gPC is recognized as one of the most efficient techniques thanks to its exponential rate of convergence. However, problems with discontinuities in the random space can lead to slow convergence. Similarly long-time integration problems could be encountered [8], where this behavior is due to the modification in time of the statistic properties of the solution that induces an efficiency loss of the polynomial basis in time. Recently, Gerritsma et al. [9] proposed a time-dependent generalized Polynomial Chaos scheme based on the research of a time varying optimal polynomial basis.

Another class of method for the UQ is based on the stochastic collocation (SC) approach [10]. This strategy is based on building interpolants (polynomial), of the same dimensionality as the stochastic space, in order to approximate the solution. In order to reduce the computational cost for high-dimension problems, these methods are often coupled to sparse grids techniques. The sparse grid strategy has been proposed by Smolyak [11] allowing interpolation of the function in a reduced subset of points with respect to the full tensorization set. This strategy is a cure against the so-called curse of dimensionality [12] problem, i.e. the exponential growth of the number of points with respect to the stochastic dimensions [13,14].

Actually, handling a non-smooth behavior for high-dimension problems remains a very challenging issue. It is not completely solved even for low or moderate dimension problems. In the context of gPC schemes, Wan and Karniadakis introduced an adaptive class of methods for solving the discontinuity issues by using local basis functions, the multi-element generalized Polynomial Chaos (ME-gPC) [15]. This strategy deals with an adaptive decomposition of the domain on which local basis are employed. In order to treat discontinuous response surfaces, Le Maître et al. applied a multiresolution analysis to Galerkin projection schemes [16,17]. This class of schemes relies on the projection of the uncertain data on a multiwavelets basis consisting of piecewise polynomial (smooth) functions. This approach is shown to be very CPU demanding. Consequently, two cures are then explored in the context of adaptive methods: automatically refine the multi-wavelets basis or adaptively partitioning the domain.

More recently, unsteady stochastic problems have been solved by means of multi-elements techniques, employing the collocation simplex method [18]. Also for these stochastic collocation methods, adaptive strategies have been proposed in order to tackle the discontinuity issues. In the work of Agarwal and Aluru [19], an adaptive stochastic collocation method, based on the recursive splitting of the domain, has been proposed. In this case the splitting of the domain and the adaptivity is applied directly to the sparse grid basis. A sparse grid collocation strategy, based on piecewise multi-linear hierarchical basis functions, has been adopted by Ma and Zabaras [20] to recover the convergence loss by a global polynomial approximations in presence of discontinuities.

Recently, Abgrall et al. [21–23] introduced a new class of finite volume schemes capable to deal with discontinuous problems both in the physical and stochastic space for shock-dominated flows. The so-called semi-intrusive scheme (SI) exhibits promising results in term of accuracy and efficiency compared to more classical Monte Carlo and gPC methods. The idea is to extend to the stochastic space the finite volume representation used for the deterministic scheme. The established framework of the reconstruction techniques (ENO/WENO) in finite volume schemes can be, very easily, employed in the stochastic space with the SI scheme. This approach can lead to some advantages such as an extreme flexibility with respect the form of the pdf (that can be discontinuous and unsteady), an easy implementation, a slight modification of the deterministic solver preserving the number of equations.

The aim of the present work is to provide a framework, inspired from the classical multiresolution representation of Harten [24], capable to recover the same results of this theory but including new features for the extension to stochastic problems. The proposed algorithm, the Truncate and Encode (TE) strategy, displays very good properties in terms of convergence and efficiency. Moreover, it allows handling adaptively a stochastic mesh in a very general way. This could allow in the future a very easy coupling with different kinds of numerical methods as, for example SC and SI schemes. While in this work no dependence on the physical spaces is considered, the long-term objective is to build accurate numerical scheme, for low or moderate number of uncertainties, permitting to deal with unsteady discontinuous solutions and using unsteady refinement/derefinement capabilities both in the physical and stochastic space.

The approach proposed in the present work is based on a multiresolution concept, as already made in Le Maître's work. However, the approach differs completely since here no spectral projection is employed, as it will be explained in the next section. Moreover, the possibility to reject a wavelets (equal to an interpolation error as in the original Harten framework) is based only on local tests, then is different from Galerkin projection approach where 1D energy estimator along stochastic

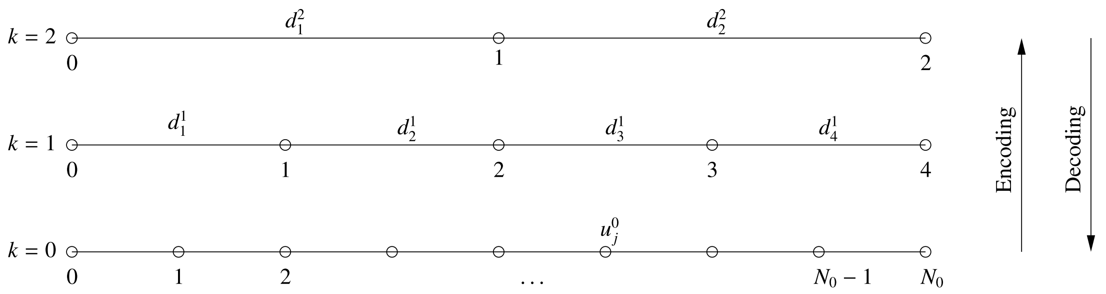


Fig. 1. Harten's multiresolution classical notation.

dimensions are used. For details on the multiresolution approach applied to Galerkin projection schemes, the reader can refer to the extremely exhaustive reference [6].

The paper is organized as follows. In Section 2, the classical multiresolution framework of Harten is illustrated. Some remarks are pointed out over the analytical results associated to the compression procedure. The algorithm proposed in this work, i.e. the Truncate and Encode (TE) strategy is presented in Section 3, while the accuracy preserving time-advancement procedure in Section 4. Differences between the classical MR framework and the present approach are discussed in Section 5. Accuracy and efficiency of the TE algorithm are verified on several numerical experiences in Section 6. In particular, algebraic and ordinary differential equation (both scalar and vectorial) and the 1D Kraichnan–Orszag problem with uniform and discontinuous unsteady probability distribution functions are considered. Finally, a two degree-of-freedom aeroelastic model is used to estimate the unsteady statistics for the motion. Remark that all the presented results are compared with respect to classical Monte Carlo methods and with gPC non-intrusive approach. Some concluding remarks and perspectives are reported in Section 7.

2. Harten's multiresolution framework

In this section, we briefly recall the classical Harten MR framework [25,24] extended here to the case of non-uniform measure. Consider a function $u(\xi)$ defined on a domain $\Xi=[0,1]$. Let us suppose to know the values $u^0=\{u_j^0\}_{j=0}^{N_0}$ on a uniform grid, defined as follows

$$\mathcal{G}^0=\{\xi_j^0\}_{j=0}^{N_0},\quad \xi_j^0=jh_0,\ h_0=1/N_0.$$

This grid is assumed as the finest one, i.e. the highest resolution level. Now, let us consider the set of nested dyadic grids $\mathcal{G}^k$ with $0\leqslant k\leqslant L$

$$\mathcal{G}^k=\{\xi_j^k\}_{j=0}^{N_k},\quad \xi_j^k=jh_k,\ h_k=2^kh_0,\ N_k=N_0/2^k,$$

where $k=0$ represents the finest level and L the coarsest. Remark that the use of a nested dyadic structure allows obtaining a grid $\mathcal{G}^k$ from the finest adjacent level $\mathcal{G}^{k-1}$ by removing only the odd points and preserving the condition $\mathcal{G}^k\cap\mathcal{G}^{k-1}=\mathcal{G}^k$. The following relation holds

$$u_j^k=u_{2j}^{k-1}\quad \text{for } 0\leqslant j\leqslant N_k.$$

In this paper, only structured uniform nested grids are employed, but the algorithm can be extended to an unstructured mesh as already shown in [26] provided that a nested structure is used between successive levels. The assumption of a regular mesh is then not exhaustive even in the 1D context. It is made here only for the sake of simplicity. The algorithm presented in this work, even described only in the 1D context, could be extended in a straightforward way following, for example, [27] or [28]. These works show how the classical Harten's framework is extended to 2D using a tensor product approach.

The convention adopted in this work is reported in Fig. 1.

In a general way, an approximation problem can be solved on the mesh to obtain an interpolation operator $\mathcal{I}(\xi;u)$ chosen to approximate the function u at the coordinate ξ. The problem can be formulated (for the generic level kth) as follows: find $\mathcal{I}_j(\xi;u^k)\in\mathbb{P}^r(\Xi)$ such that for any point ξ_j one has $L_j(\mathcal{I}(\xi;u^k))=L_j(u^k)$ for $0\leqslant j\leqslant N_k$, where $\{L_j\}_{1\leqslant j\leqslant r}$ is a family of linear forms defined on $\mathcal{C}^{\bar{r}}(\mathbb{R})$ $(\bar{r}>r)$. With this assumption, two different kinds of linear forms could be considered:

- Lagrange interpolation: the linear form is directly equal to the point value of the function

$$L_j\bigl(u^k\bigr)=u(\xi_j)\quad 0\leqslant j\leqslant N_k.$$

- (Conditional) average reconstruction, where at each point ξ_j, its surrounded cell Ξ_j is associated such that

$$L_j(u^k) = \bar{u}_j^k = \frac{\int_{\Xi_j} u^k \, d\xi}{\int_{\Xi_j} d\xi}. \tag{1}$$

In the following, no distinction is made by the two reconstruction procedures, because they are similar even if not exactly equivalent. An exact equivalence between the two reconstruction procedures is not possible, even in the case of constant reconstruction. In fact, a set of nested cells cannot be obtained on the same set of nested meshes, using the operator $\mathcal{I}(\xi; u^k)$. In particular, it is necessary to predict the value of the function in the *missing points*, i.e. the points $\xi \in \mathcal{G}^{k-1}$ but not to $\mathcal{G}^k \cap \mathcal{G}^{k-1}$ or its cell averaged value in its cell. The expression for both predicted values are the following (for $1 \leqslant j \leqslant N_k$):

$$\begin{aligned} \tilde{u}_{2j-1}^{k-1} &= \mathcal{I}(\xi_{2j-1}^{k-1}; u^k), \\ \tilde{u}_{2j-1}^{k-1} &= \frac{\int_{\Xi_{2j-1}^{k-1}} \mathcal{I}_{2j-1}^{k-1}(\xi; u^k) \, d\mu}{\int_{\Xi_{2j-1}^{k-1}} d\mu}. \end{aligned} \tag{2}$$

The interpolation errors d_j^k can be defined as follows

$$d_j^k = \bar{u}_{2j-1}^{k-1} - \tilde{u}_{2j-1}^{k-1}, \quad \text{for } 1 \leqslant j \leqslant N_k, \tag{3}$$

and generally they are called *wavelets coefficients* or *details*.

In this work, the point-value setting is employed. Anyway, the presented approach holds also when a cell average setting is employed.

The MR strategy is based on the observation that the knowledge of the couple (d^k, u^k), where $d^k = \{d_j^k\}$ and $u^k = \{u_j^k\}$, called multiresolution representation of u^{k-1}, permits to compute the solution on the grid $\mathcal{G}^{k-1}$. Of course, the *vice versa* also holds

$$u^{k-1} \leftrightarrow (d^k, u^k).$$

Proceeding recursively from u^0 to u^L

$$u^0 \leftrightarrow (d^1, u^1) \leftrightarrow (d^1, d^2, u^2) \leftrightarrow \cdots \leftrightarrow (d^1, d^2, \ldots, d^L, u^L) \stackrel{\text{def}}{=} (u_M)^{\mathrm{T}}$$

it is possible to obtain the multi resolution representation of the solution $(u_M)^{\mathrm{T}}$.

Remark that it is possible to obtain the multiresolution representation $(u_M)^{\mathrm{T}}$ with any interpolation technique, i.e. any degree of interpolation is allowed provided the appropriate stencil. In this framework two different operations can be defined: the *encoding* procedure that allows obtaining the multiresolution representation from the knowledge of u^0 (the solution at the finest level) and the *decoding* procedure that allows obtaining the function at the finest resolution level from the multiresolution representation. If a matrix notation is employed the *encoding* procedure could be formulated as

$$u_M = M u^0, \tag{4}$$

where M is an $(N_0 + 1) \times (N_0 + 1)$ matrix. The *decoding* procedure is the inverse procedure, then obviously

$$u^0 = M^{-1} u_M. \tag{5}$$

This matrix form is possible if the set of stencil is fixed and there is no adaptation, as for instance it happens in ENO reconstructions. In the case of automatic procedure to adapt the stencil the matrix M, that would not be linear, could be computed term by term in a closed form. The same is valid for the ‘inverse’ operator, see [29].

The MR framework, as presented in this section, could be seen only as hierarchical evaluation of a solution or, if applied to a numerical scheme, a hierarchical recasting of the same scheme. However one of the main advantages of the MR is the possibility to obtain more efficient schemes introducing a *truncation* procedure. The data compression capabilities of MR framework are then treated in the next section.

2.1. Data compression

As depicted in the previous section, the finest level can be reconstructed by the *decoding* procedure exploiting the multiresolution representation of the function, that is constituted by all the *details* d_j^k and all the point values of the function u_j^k. If we consider Eq. (3) rearranged in the following form

$$u_{2j-1}^{k-1} = d_j^k + \tilde{u}_{2j-1}^{k-1}, \tag{6}$$

it is evident that the value of the function u_{2j-1}^{k-1} can be obtained directly from $\tilde{u}_{2j-1}^{k-1}$ through the detail d_j^k. The approximate value is obtained by interpolation, as shown in (2), from the adjacent level.

The storage memory, required for the multiresolution representation, can be reduced using the following procedure. Let us choose a certain threshold ε. Then, the details can be compared to the threshold imposed for the kth level ε_k, yielding a truncated details $\hat{d}_j^k$ defined as follows:

$$\hat{d}_j^k = \begin{cases} d_j^k & \text{if } |d_j^k| > \varepsilon_k, \\ 0 & \text{if } |d_j^k| \leqslant \varepsilon_k. \end{cases} \tag{7}$$

If the procedure is recursively followed for each level, a new multiresolution representation is obtained, with a large number of zero details, that do not need to be stored

$$\hat{u_M} = \operatorname{tr}(u_M) = \left(\hat{d}^1, \hat{d}^2, \ldots, \hat{d}^L, u^L\right).$$

If the result of the *decoding* procedure, after the truncation, is denoted as $\hat{u}^0$

$$\hat{u}^0 = M\hat{u_M} = M\operatorname{tr}(u_M)$$

the following estimation holds (see [29] for a proof), both in the L_1 and L_∞ norms

$$\left\|u^0 - \hat{u}^0\right\| \leqslant C\varepsilon, \tag{8}$$

where the constant C is independent on the coarsest level L and the local threshold ε_k is defined as

$$\varepsilon_k = \varepsilon/2^k.$$

As a result, one needs only to fix a threshold ε for the finest level $k = 0$ and, moving from the finest to the coarsest, the threshold for the other levels is directly obtained from the finer one.

Details in this framework are of strong importance not only to compress data, but because they can be associated to the local regularity of the function. This feature will be presented in more detail in Section 5, where the regularity of the function plays a fundamental role. However, thanks to the ability to capture the regularity properties of the function, the *wavelets* are employed in the refinement step, as presented in the following section.

3. A one-time truncated-encoding strategy

The aim of this work is to develop a more flexible strategy, in the context of uncertainty quantification for compressible flow problems, in which the classical multiresolution framework Section 2 is employed as a basis to build a non-intrusive technique for steady and non-steady problems. However in the following (Section 4) a procedure to extend the present strategy to unsteady problems is also presented, introducing a proper advancing procedure, and some numerical results will be reported in Section 6.

A representation of the solution on a finest grid is computed starting from a coarsest grid, with a lower number of evaluation of the function. This implies that only a reduced set of point values, on the finest grid, is evaluated, while the remaining set is obtained by interpolation. This procedure moves recursively, with a combination of interpolation and evaluation, from the coarsest level to the finest. The direction here (from the coarsest to the finest) is the opposite with respect the classical Harten's framework, where, as shown in the previous section, the algorithm can start with an *encoding* procedure on the initial condition and successively the original system of equation is transformed into an equivalent set of equations on the *wavelets* coefficients. Then at each time step the scheme allows computing the solution on the finest level by the *decoding* procedure, i.e. the computation is explicitly performed only on the coarsest level and only the significative coefficients are computed.

Let us consider a scalar function $u = u(\xi)$ with $\xi \in \Xi = [0, 1]$. The proposed strategy is constituted by the following steps (the notation is the same of the Harten's multiresolution framework, i.e. $k = 0$ for the finest level and $k = L$ for the coarsest):

- Parameters assignment (the procedure can start only if the condition $m_L < m_{\max}$ is satisfied)
  – Fix a threshold ε (the solution is assumed to be solved with this threshold on the finest grid[1]);
  – Fix an index $m_{\max} \in \mathbb{N}$ for the maximum allowed level ($N_{\max} = N_0 = 2^{m_{\max}}$);
  – Fix an index $m_L \in \mathbb{N}$ for the coarsest level ($N_L = 2^{m_L}$).

[1] This is the same hypothesis as in the classical MR framework.

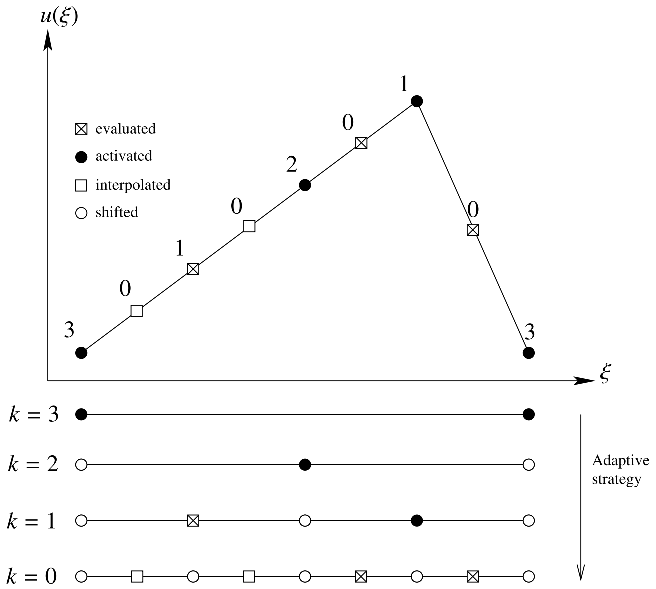


Fig. 2. Example of application of the adaptive strategy on a triangular function $f = f(\xi)$.

- Initialization of the function u at each location at the coarsest level $u(\xi_j^L) = u_j^L$ with $j = 0, \dots, N_L$ where

$$\mathcal{G}^L = \{\xi_j^L\}_{j=0}^{N_L}, \quad \xi_j^L = jh_L,\ h_L = 2^L h_0,\ N_L = N_0/2^L, \tag{9}$$

and $h_0 = 1/N_0$. Each level can be labeled computing the equivalent index k_{eq}

$$N_{k_{eq}} = \frac{N_0}{2^{k_{eq}}} \to 2^{k_{eq}} = \frac{N_0}{N_{k_{eq}}} \to \log_2 2^{k_{eq}} = k_{eq} = \log_2\left(\frac{N_0}{N_{k_{eq}}}\right),$$

and then the coarsest level can be labeled by k_L where $L = m_{\max} - m_L$.
Evaluation of the function at the subsequent level, with respect to the coarsest one, i.e. the function should be evaluated at the new set of mid points $\mathcal{G}^{L-1} \setminus \mathcal{G}^L$

$$\mathcal{G}^{L-1} = \{\xi_j^{L-1}\}_{j=0}^{N_{L-1}}, \quad \xi_j^{L-1} = jh_{L-1},\ h_{L-1} = 2^{L-1}h_0,\ N_{L-1} = N_0/2^{L-1}. \tag{10}$$

- Starting of the adaptive strategy by means of a recursive procedure ($k < L - 1$)
  (A) The *wavelets coefficients* are computed for the present level k as

$$d_j^k = u_j^k - \frac{1}{2}\left(u_{\frac{j+1}{2}}^{k+1} + u_{\frac{j-1}{2}}^{k+1}\right) \quad \text{for } 0 \leqslant j \leqslant N_k \text{ with } j \text{ odd;}$$

  (B) The wavelets coefficients are compared with the threshold $\varepsilon_k = \varepsilon/2^k$. If $|d_j^k| > \varepsilon_k$ then the two nodes ξ_{2j+1}^{k-1} and ξ_{2j-1}^{k-1} will be flagged as active on the next finer mesh $\mathcal{G}^{k-1}$. If $|d_j^k| < \varepsilon_k$ then the *wavelets* is truncated, i.e. its value is posed zero.
  (C) The new level $k - 1$ is generated if $k > 0$ and only on the activated points the function u is evaluated.
  (D) Moving from a level k to the finer adjacent one $k - 1$, three different cases are possible:
  $*$ If $\xi_j^k \in \mathcal{G}^k \cap \mathcal{G}^{k+1}$ then $u_j^k = u_{2j}^{k+1}$ (shifting).
  $*$ If $\xi_j^k \notin \mathcal{G}^k \cap \mathcal{G}^{k+1}$ and it is not flagged then interpolate

$$u_j^k = \frac{1}{2}\left(u_{\frac{j+1}{2}}^{k+1} + u_{\frac{j-1}{2}}^{k+1}\right).$$

  $*$ If $\xi_j^k \notin \mathcal{G}^k \cap \mathcal{G}^{k+1}$ and it is flagged as active (by the step B of the algorithm) then evaluate, i.e. call the model.
  (E) The algorithm stops when the maximum level is reached or when all the *wavelets* coefficients can be truncated (at a certain level $k > 0$).

In order to make things clearer, the application of the proposed strategy is illustrated on a triangular function $f = f(\xi)$ in Fig. 2. The following parameters, $m_L = 0$, $m_{\max} = 3$ then $L = m_{\max} - m_L = 3$, are chosen. Full circles indicate an activated evaluation point, performed during the first steps of the algorithm when levels $k = 3$ and $k = 2$ are evaluated. Remark the exclusion of the shifted points of the level 2 from 3. The interpolation operator is linear and the threshold is fixed. At the last (finest) level, the algorithm stops because the maximum level is reached, however no other points in level $k = 0$ are marked as active and then the procedure would be stopped in any case. Remark that, at the level $k = 1$, a square with a cross indicates an evaluated, but non-activated point, in fact in this position the interpolation error is zero (the function is linear). The two neighborhood points at the next levels are then only interpolated (squares) and not evaluated by reducing the global computational cost.

3.1. Application to the uncertainty quantification

In the present work the case of equations with a solution depending only from the stochastic (and not physical) space is addressed. Due to the presence of a weight function, i.e. the probability density function, this case could be seen as a generalization of the classical approach in which function defined in the physical space (with unitary weight function) are considered.

One of the aim of UQ is to quantify the statistic moments of a quantity of interest. Let us assume to have a solution $u(\xi, t)$ for which the aim is to compute the statistics, i.e. for example the expectancy $\mathcal{E}$:

$$\mathcal{E}(t) = \int_{\Xi} u(\xi, t) p(\xi, t)\, \mathrm{d}\xi,$$

where $\xi \in \Xi$ is the parameter (or vector of parameters) and $p(\xi)$ is the probability density function (pdf) related to ξ. Remark that the pdf could assume a whatever form, including discontinuity and time-dependent properties.

Remark that the normalization condition, i.e. $\int_{\Xi} p(\xi, t)\, \mathrm{d}\xi = 1$, must be always fulfilled for each pdf and time step.

The above integrals could be computed when the solution is already computed at the finest resolution level, by means of true evaluations of the model or interpolations. In particular, using the trapezoidal rule, the integrals can be computed. Moreover, variance can be obtained applying the following relations

$$\begin{aligned} \mathcal{E}(t) &= \int_{\Xi} u(\xi, t) p(\xi, t)\, \mathrm{d}\xi, \\ \mathrm{Var}(t) &= \int_{\Xi} \bigl(u(\xi, t) - \mathcal{E}(t)\bigr)^2 p(\xi, t)\, \mathrm{d}\xi = \int_{\Xi} u(\xi, t)^2 p(\xi, t)\, \mathrm{d}\xi - \mathcal{E}^2(t). \end{aligned} \tag{11}$$

Note that the integrals are evaluated with the solution discretized at the finest level. In the case of the variance, the expectancy of the squared function $u(\xi, t)$ is computed applying the quadrature rule to the square of the values $u(\xi, t)$, that are already stored.

Now, let us show how the use of TE algorithm allows handling pdf's of whatever form with, eventually, traveling unsteady discontinuities in the stochastic space. Let us consider a solution $u(\xi)$ defined on the domain $\Xi = [-1, 1]$ with a jump discontinuity located in $\xi = -1/2$ and smooth in the remaining part of the stochastic domain. If the pdf is uniform, the correct representation of the function, also in the neighborhood of its discontinuity becomes fundamental. However, let us imagine to have a different pdf distribution, for instance still a uniform pdf, but now defined only on the positive sub-domain $\Xi \supset \bar{\Xi} = [0, 1]$, i.e. the pdf would be uniform and unitary in $[0, 1]$ and equal to zero in $[-1, 0]$. Despite the possibility to solve accurately the solution in the neighborhood of the discontinuity, this effort it is not motivated, because the weight function, i.e. the pdf, would lead to zero the contribution to Eqs. (11) in each point of the domain $[-1, 0]$. Therefore, in the general case of a non-smooth time-dependent pdf, it is more convenient to apply the TE algorithm to the product between the solution $u(\xi, t)$ and the pdf $p(\xi)$ considering a unitary measure $\mathrm{d}\mu = 1\, \mathrm{d}\xi$. In particular, the *wavelet*, for the level k, is computed as follows

$$d_j^k = u_j^k p\bigl(\xi_j^k\bigr) - \widetilde{u_j^k p\bigl(\xi_j^k\bigr)} \quad \text{for } 0 \leqslant j \leqslant N_k \text{ with } j \text{ odd}, \tag{12}$$

where $\widetilde{u_j^k p(\xi_j^k)}$ for the point-value setting is

$$\widetilde{u_j^k p\bigl(\xi_j^k\bigr)} = \frac{1}{2}\bigl(u_{\frac{j+1}{2}}^{k+1} p\bigl(\xi_{\frac{j+1}{2}}^{k+1}\bigr) + u_{\frac{j-1}{2}}^{k+1} p\bigl(\xi_{\frac{j-1}{2}}^{k+1}\bigr)\bigr) \quad \text{for } 0 \leqslant j \leqslant N_k \text{ with } j \text{ odd}. \tag{13}$$

The scheme is modified only when *wavelets* are computed, i.e. in the significant points where the solution needs to be explicitly evaluated. However, the solution $u(\xi)$ is the only quantity retained, so computation of the statistics is not affected by the presence of a non-classical pdf. The present approach is justified because an analytical form of the probability density function $p = p(\xi, t)$ is assumed. Anyway, if the pdf itself is governed by an evolutionary equations, also the value of the pdf should be stored in the activated points, following the *wavelets* computed by (12). As a consequence, the evaluation of the statistics remain the same and relations (11) hold.

The computation of the *wavelets* should take into account the evolution, both in time and stochastic space, of the pdf and also of each component of a vectorial solution (if the system has several outputs). This can be performed by means of slight modifications of the TE algorithm, as presented in the next section.

3.1.1. *Application to vectorial solutions*

In this section, we focus our attention to a vectorial solution problem. Let us study the stochastic response of a system that has many outputs. In this case, the application of the algorithm is straightforward if different runs of the algorithm are performed, one on each component of the response vector. The final adding step would be to merge together the different multiresolution representations of each scalar components of the solution. Let us consider a vector of responses $u \in \mathbb{R}^n$ defined as $\mathbf{u}(\xi) = \{u_1(\xi), u_2(\xi), \dots, u_n(\xi)\}^{\mathrm{T}}$. The final MR representation should be, in this case, $\mathcal{G}^k = \bigcup_{i=1}^n \mathcal{G}_i^k$, where $\mathcal{G}_i^k$ is used to indicate the multiresolution representation associated to $u_i(\xi)$. The union of the different representations of the scalar components can be obtained in a very efficient way computing only the solution at the smallest possible set of points. This set of points is identified as the union of all the points, where at least one of the scalar components $u_i(\xi)$ should be explicitly evaluated. The consequence of such approach is to compute explicitly the entire vectorial solution in a point $\xi \in \mathcal{G}^k$, if and only if one component $u_i(\xi)$ cannot be reconstructed with the prescribed accuracy ε. In practice, this can be done performing only two modifications to the algorithm described in Section 3 for a scalar function.

Step A is formulated as follows: evaluate the *wavelets* coefficients d_j^k for each element of $\mathbf{u}$ by introducing a new index i for the component u_i:

$$d_{i,j}^k = u_{i,j}^k - u_{i,j}^{\tilde{k}} \quad \text{for } i = 1, \dots, n \text{ and } 0 \leqslant j \leqslant N_k \ (j \text{ odd}).$$

The other modification concerns the step B where the criterion for the truncation is computed for all $d_{i,j}^k$: $|d_{i,j}^k| > \varepsilon_k$. If almost one of $d_{i,j}^k$ (for $i = 1, \dots, n$) cannot be truncated, i.e. $|d_{i,j}^k| > \varepsilon_k$, the two successive nodes (ξ_{2j+1}^{k-1} and ξ_{2j-1}^{k-1}) are activated at the subsequent level.

Remark that all the other steps remain the same presented for the scalar function and that for non-uniform distributed parameter Eq. (12) should be employed.

At the end of the algorithm, the mesh that allows recovering the finest one with the prescribed accuracy regarding each component u_i is obtained. Remark that in practical cases, this procedure, even if general, could not be very efficient. For example, when compressible fluid dynamics problem are considered (under the hypothesis of ideal gas), the density could be used as the unique parameter in order to detect smooth region, contact discontinuity and shock waves. In Section 6, this procedure has been applied to the 1D Kraichnan–Orszag problem and to an aeroelastic problem where the transverse deflection (that models the bending of the wing) and the torsional angle should be computed to recover the aerodynamic loads. In these cases, the TE algorithm has been applied on all the components following the procedure described in this section.

4. An accuracy preserving time-advancement strategy

In this section we focus on a procedure able to deal with time dependent probability density functions employing the TE algorithm.

4.1. *Harten framework*

The original Harten framework provides a procedure based on updating the solution based on the CFL condition. The set of important coefficients is modified, at each time step, in order to capture the correct evolution of the solution. Then, the original numerical scheme can be reduced to a scheme on the wavelets coefficients [29].

Let us consider the case of linear interpolation. In this case, the multiresolution representation $u_M = \{d^1, d^2, \dots, d^L, u_L\}^{\mathrm{T}}$ can be obtained multiplying the matrix M, of dimension $(N_0 + 1) \times (N_0 + 1)$ with the solution at the finest level u_0

$$u_M(t) = M u_0(t).$$

The time advancing procedure, used to obtain u_0 at the successive time steps, is based on the computation of the coarsest level u_L and the wavelets coefficients d^k with $k = 1, \dots, L$. These terms recasted in the multiresolution vector can be used to compute the solution at the finest level

$$u_0(t) = M^{-1} u_M(t). \tag{14}$$

In the case of compressed values, i.e. the multiresolution representation after the *truncation* procedure, the procedure remains the same, but the multiresolution vector u_M is truncated, then $\hat{u_M}$ is obtained from the truncated counterparts of the details (see Eq. (7)).

As a result, the key part of the Harten algorithm is to identify and compute the important coefficients at each time step. This is made by a CFL based algorithm described in detail in [29]. The general idea is to identify the nodes in which an information can propagate and enlarge this stencil for stability reasons applying some empirical rules, the so-called *safety set* [29].

Many authors showed the efficiency of this procedure, but in the context of stochastic equations, finding a CFL condition can be hard. Then, it is not possible to apply directly this technique. In particular, the evolution of the probability density function is not governed by evolutionary equations. In the next section, we show how to extend the TE algorithm in order to follow the evolution of an unsteady pdf.

4.2. Extension of TE algorithm to unsteady problems

Let us assume an unsteady differential equation $\mathcal{L}(f(\xi)) = 0$ in which the random parameter has a probability distribution $p = p(\xi)$. The statistics of the output f will be time-dependent according to the solution of the differential equation governed by the operator $\mathcal{L}$. In this case, the expectancy $\mathcal{E}$ should be computed as $\mathcal{E}(t) = \int_{\Xi} f(\xi, t)p(\xi)\,d\xi$, then the multiresolution approach can be applied directly to $f(\xi, t)$ using a CFL-like condition. What happens if the output f of the system is steady but the pdf is unsteady, i.e. $f = f(\xi)$ with $p = p(\xi, t)$? In this case the expectancy should be computed as $\mathcal{E}(t) = \int_{\Xi} f(\xi)p(\xi, t)\,d\xi$. The pdf of the input could be imposed, but is not governed by a differential equation from which a CFL-like condition can be derived. In this case, the classical Harten framework cannot be used. However, if the evolution of the product function $\tilde{f}(\xi, t) = f(\xi)p(\xi, t)$ is considered, an extended TE algorithm can be introduced.

The idea is to perform, for each time step, the TE algorithm on the product function $\tilde{f}(\xi, t)$. This allows minimizing the number of points to recover the finest mesh set of points, with an *ad hoc* interpolation technique, within a prescribed accuracy. At the successive steps, the integration in time could be performed starting from the knowledge of the solution at the previous step known with the prescribed tolerance.

Let us consider the solution of a Cauchy problem, i.e. find statistics for $y \in \mathcal{C}(0, T)$

$$\begin{cases} \dot{y}\big(t, \xi(t)\big) = f\big(t, y(t), \xi\big) & t \in [0, T] \text{ and } \xi \in \Xi = [0, 1], \\ y(0) = y_0. \end{cases} \tag{15}$$

The first step is to identify, as in the pure deterministic case, the time-integration technique. If the time space is discretized as $t_n = n\Delta t$ with a constant step Δt according to the stability condition, and if an explicit Euler scheme is considered, Eq. (15) becomes

$$y(t_{n+1}, \xi) = y(t_n, \xi) + \Delta t f\big(t_n, y(t_n), \xi\big) \tag{16}$$

where each time integration is performed at a fixed location in the stochastic space. The TE algorithm must be, in the unsteady case, applied at each time step starting from the coarsest level to the finest one. According to the time-integration technique, the choice of a stencil in time is needed in order to evaluate a solution in the point of the space Ξ–t. In the case of the explicit Euler scheme shown before, only the previous time step t_n (at the same stochastic location ξ) is needed to evaluate the solution $y(t_{n+1}, \xi)$. For each location in the stochastic space $\xi_j \in \mathcal{G}^k = \{\xi_j^k\}_{j=0}^{N_k}$, corresponding to the stochastic mesh at the kth level[2] of the MR representation by the TE algorithm, two cases could occur:

1. At the previous time step the TE algorithm have reached a level equal or finer than the level k.
2. At the previous level a level coarser than k has been reached.

In the first case the solution can be computed as

$$y\big(t_{n+1}, \xi_j^k\big) = y\big(t_n, \xi_j^k\big) + \Delta t f\big(t_n, y(t_n), \xi_j^k\big), \tag{17}$$

while in the second case the value $y(t_n, \xi_j^k)$ was not computed at the previous time step and it is not available to evaluate $y(t_{n+1}, \xi_j^k)$. However at each time step the representation of the solution, thanks to the TE algorithm, is known with a prescribed accuracy with respect the solution on the finest level as reported in Eq. (8). The relation (8) justifies the possibility to interpolate the value of the function y at the previous time step $y(t_n, \xi_j^k)$ as

$$y\big(t_n, \xi_j^k\big) \approx \mathcal{I}\big(\xi_j^k; y^k(t_n)\big), \tag{18}$$

where, using the same notation of Section 2, the solution discretized at level kth at the time t_n, $y(t_n, \xi_j^k)$, is indicated as $y^k(t_n)$.

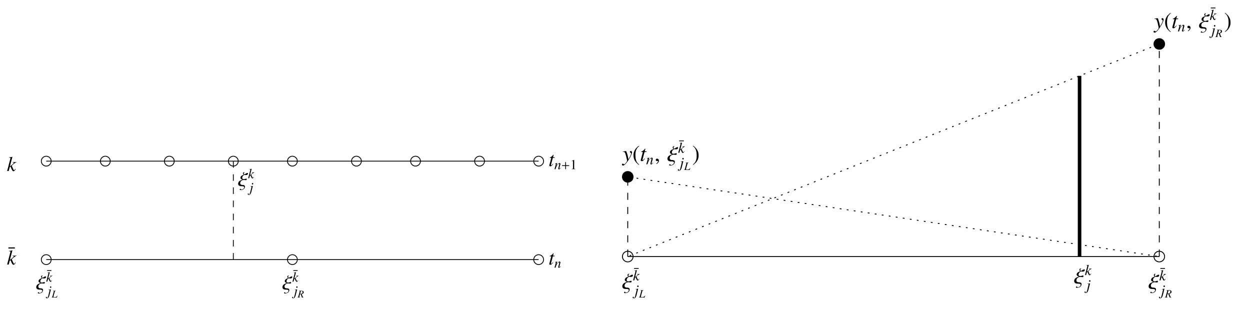
In the case of linear interpolation, considering the maximum level $\bar{k}$ reached at the time step t_n, the algorithm that can be used in order to compute the value of y at the position ξ_j^k, is the following (see Fig. 3):

- Determination of the left value $\xi_{j_L}^{\bar{k}}$ as the maximum of the set $\{\xi_j^{\bar{k}} \mid \xi_j^{\bar{k}} < \xi_j^k\}$;
- Determination of the right value $\xi_{j_R}^{\bar{k}}$ as $\xi_{j_R}^{\bar{k}} = \xi_{j_L}^{\bar{k}} + h_{\bar{k}}$ where $h_{\bar{k}} = 1/2^{\bar{k}}$.

The value of the function y at the time step t_n, interpolated by means of the values $y(t_n, \xi_{j_L})$ and $y(t_n, \xi_{j_R})$ already known, is then recovered by linear interpolation as

$$y\big(t_n, \xi_j^k\big) \approx \mathcal{I}\big(\xi_j^k; y^k(t_n)\big) = \frac{y(t_n, \xi_{j_R}^{\bar{k}}) - y(t_n, \xi_{j_L}^{\bar{k}})}{h_{\bar{k}}} \xi_j^k + y\big(t_n, \xi_{j_L}^{\bar{k}}\big) - j_L\big(y\big(t_n, \xi_{j_R}^{\bar{k}}\big) - y\big(t_n, \xi_{j_L}^{\bar{k}}\big)\big), \tag{19}$$

[2] The index of the level hereafter must be intended as equivalent indexes as presented in Section 3.



(a) (b)

Fig. 3. Linear interpolation stencil and convention adopted to perform the time integration. Stencil identification (a) and linear interpolation (b).

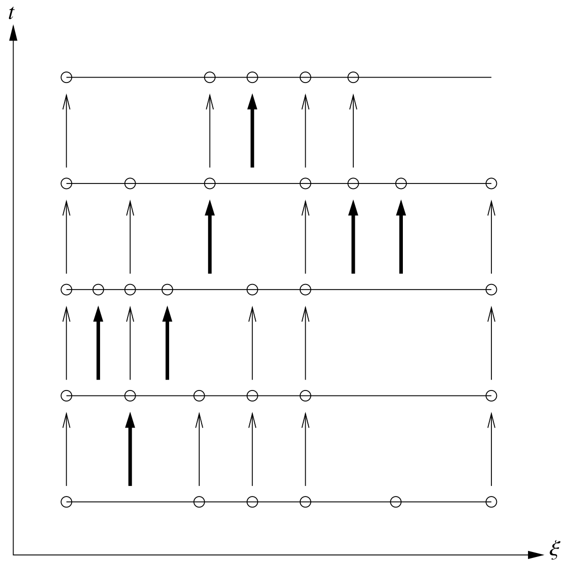


Fig. 4. Time advancing. In bold the arrows for the advancing in time from interpolated value (Eq. (20)).

where $j_L = \xi^{\bar{k}}_{j_L}/h_{\bar{k}}$. In the case of evaluation of the value $y(t_{n+1}, \xi^k_j)$, starting from the interpolated value at the previous time step, the scheme can be written as

$$y\left(t_{n+1}, \xi^k_j\right) = \mathcal{I}\left(\xi^k_j; y^k(t_n)\right) + \Delta t\, f\left(t_n, \mathcal{I}\left(\xi^k_j; y^k(t_n)\right), \xi^k_j\right). \tag{20}$$

After some time steps the pattern of the different paths, corresponding to different stochastic location ξ, appear as a discontinuous succession of time integration. A typical pattern, for a Cauchy problem as the one considered in this section, employing an explicit Euler formula, is sketched in Fig. 4. An arrow means a time-advancement from a known, i.e. already computed point, while a bold arrow means an evaluation from an interpolated value.

We point out that this procedure permits to solve the long-time integration problem if a proper threshold is provided and if the number of refinements is large enough, according to the regularity of the solution. In the following, some results concerning the long-time integration problem will be illustrated for the 1D Kraichnan–Orszag problem.

The proposed time-advancement strategy can be easily used to build numerical scheme with automatic refinement/derefinement. The refinement/derefinement technique is shown in Fig. 4 where the time evolution, once fixed a certain position ξ, could not appear as a continuous sequence of advancing operations (arrows) on computed points (circles). It, instead, looks like a sequence of arrows and blank spaces (interpolation operations). This property is very promising for systems where unsteady discontinuities can appear only in a limited part of the domain, as for example in some compressible fluid dynamics problems.

5. Some remarks on the difference between the classical and adaptive MR approach

In this section, some differences between the classical approach discussed in Section 2 and the adaptive procedure presented in Section 3 are illustrated. The TE algorithm is an encoding procedure with an embedded truncation capability. The application of the TE strategy produces a multiresolution representation $\hat{u_M} = (\hat{d}^1, \hat{d}^2, \ldots, \hat{d}^L, u^L)$, already truncated with respect to the accuracy governed by the threshold ε. This structure allows reconstructing the solution at the finer level u^0 starting from the coarsest level with the classical *decoding* procedure. Remark that in this case the *encoding* and *decoding* procedures move from the coarsest to the finest level, while in the classical MR approach the *encoding* procedure is performed starting from the finest level to the coarsest one.

Provided the differences between the two representations, one key question is if the multiresolution structures u_M are or not the same.

To make the analysis, we assume in this section to know the exact solution f. The problem is: given the L^∞ norm, what is the behavior of the quantity $\|u^0 - \hat{u}_k\|_\infty$. The next question is: given $\varepsilon > 0$, is there a coarsest level k for which, for any $l \geqslant k$, we have $\|u^0 - \hat{u}_k\|_\infty \leqslant \varepsilon$. For this reason, let us focus on the stability of the reconstruction procedure in the context of the classical framework. In Harten's framework, the concept of stability is related to the possibility of controlling the norm $\|u^0 - \hat{u}^0\|$. Here, we address explicitly the case of a linear operator of reconstruction, i.e. the case in which the stencil is fixed, in particular focused on Lagrangian interpolation. However, similar results hold also for an average reconstruction procedure, as explained in Section 2. The aim is to quantify the L_∞ norm for $\|u^0 - \hat{u}^0\|_\infty$, where the hat solution $\hat{u}^k$ represent the reconstruction of the function u at the level of resolution k obtained with the application of the TE algorithm. Starting from the coarsest level $k = L$ until to the finest $k = 0$ by means of the TE strategy, the solutions discretized on the first two levels $k = L$ and $k = L - 1$ is computed explicitly. The following relation must hold: $u^L = \hat{u}^L$ and $u^{L-1} = \hat{u}^{L-1}$. By means of the TE strategy, the solution in a point $\xi_{2j-1}^k \in \mathcal{G}^k \setminus \mathcal{G}^{k+1}$ is computed as follows

$$\hat{u}_{2j-1}^k = \begin{cases} u_{2j-1}^k & \text{if } |d_{j^\star}^{k+2}| > \varepsilon_{k+2}, \\ \mathcal{I}(\xi_{2j-1}^k; \hat{u}^{k+1}) & \text{if } |d_{j^\star}^{k+2}| \leqslant \varepsilon_{k+2}. \end{cases} \tag{21}$$

The generic notation $d_{j^\star}^k$ indicates that the *wavelet* at level k, if marked as active, generates the two successive evaluations at level $k - 2$.[3]

For all the points, it holds $\xi_{2j}^k \in \mathcal{G}^k \cap \mathcal{G}^{k+1}$, due to the shifting of data $\hat{u}_{2j}^k = \hat{u}_j^{k+1}$. Obviously, the last relation must hold for the original function u discretized between two consecutive resolution levels $u_{2j}^k = u_j^{k+1}$.

The difference between the original function u and its TE counterpart $\hat{u}$ is then as follows

$$u^k - \hat{u}^k = \begin{cases} 0 & \xi_{2j-1}^k \in \mathcal{G}^k \setminus \mathcal{G}^{k+1} \text{ and } |d_{j^\star}^{k+2}| > \varepsilon_{k+2}, \\ d_j^{k+1} & \xi_{2j-1}^k \in \mathcal{G}^k \setminus \mathcal{G}^{k+1} \text{ and } |d_{j^\star}^{k+2}| \leqslant \varepsilon_{k+2}, \\ u_j^{k+1} - \hat{u}_j^{k+1} & \xi_{2j}^k \in \mathcal{G}^k \cap \mathcal{G}^{k+1}. \end{cases} \tag{22}$$

Therefore, the norm $\|u^k - \hat{u}^k\|$ in the L_∞ space holds

$$\begin{aligned} \|u^k - \hat{u}^k\|_\infty &\leqslant \max\{|d_j^{k+1}|, \|u^{k+1} - \hat{u}^{k+1}\|_\infty\} \leqslant \{|d_j^{k+1}|, |d_j^{k+2}|, \ldots, |d_j^{L-1}|, \|u^{L-1} - \hat{u}^{L-1}\|_\infty\} \\ &= \max\{|d_j^{k+1}|, |d_j^{k+2}|, \ldots, |d_j^{L-1}|\}. \end{aligned} \tag{23}$$

Previous *wavelets* should be dependent on a *wavelets* $d_{j^\star}^k$, that is bounded by the threshold ε_k, as already reported in (22).

Depending on the degree of polynomial reconstruction employed, each wavelets can be expressed, following the classical interpolation results (see for instance [30]), as follows

$$d_j^k = u_{2j-1}^{k-1} - \mathcal{I}(\xi_{2j-1}^{k-1}; u^k) = u[\mathcal{S}_j^k, \xi_{2j-1}^{k-1}]\omega_{r+1}(\xi_{2j-1}^{k-1}), \tag{24}$$

where $u[\mathcal{S}_j^k, \xi_{2j-1}^{k-1}]$ is the $r + 1$th divided difference on the stencil $\mathcal{S}_j^k$, evaluated at the point $\xi_{2j-1}^k - 1$ and the nodal polynomial ω_{r+1}. These quantities are

$$\begin{aligned} u[\mathcal{S}_j^k, \xi_{2j-1}^{k-1}] &= \sum_{\xi_m \in \mathcal{S}_j} \frac{u(\xi_m)}{\omega'_{r+1}(\xi_m)}, \\ \omega'_{r+1}(\xi_j) &= \prod_{\substack{\xi_m \in \mathcal{S}_j \\ \xi_m \neq \xi_j}} (\xi_j - \xi_m), \\ \omega_{r+1}(\xi) &= \prod_{\xi \in \mathcal{S}_|} (\xi - \xi_m). \end{aligned} \tag{25}$$

It is important to remark that the stencil $\mathcal{S}_j$ must contain $r + 1$ points. Supposing the function to be regular on the stencil $\mathcal{S}_j^k$, i.e. $u \in \mathcal{C}^{r+1}(\mathcal{S}_j)$, following relations hold

[3] Note that the step between k and $k - 2$ is only a matter of notation in which the generic *wavelet* d_j^k can be associated directly to the point ξ_{2j-1}^{k-1} and stored on this level, or associated directly to the interval containing the point ξ_{2j-1}^{k-1} at level k.

$$\begin{aligned} &u\big[\mathcal{S}_j^k, \xi_{2j-1}^{k-1}\big] = \frac{u^{(r+1)(\xi')}}{(r+1)!} \quad \text{with } \xi' \in \text{convex hull}\big\{S_j^k, \xi_{2j-1}^{k-1}\big\},\\ &\omega_{r+1}\big(\xi_{2j-1}^{k-1}\big) \sim \mathcal{O}\big(h_{k-1}^{r+1}\big). \end{aligned} \tag{26}$$

The same should hold for the *wavelet* $d_{j^\star}^{k+1}$ which generates d_j^k, i.e.

$$d_{j^\star}^{k+1} \sim \frac{u^{(r+1)}(\xi'')}{(r+1)!}\mathcal{O}\big(h_k^{r+1}\big) \quad \text{with } \xi'' \in \text{convex hull}\big\{S_j^{k+1}, \xi_j^k\big\}. \tag{27}$$

The ratio between the *wavelets* is (note that this result is valid only if the resolution level is not too coarse, this point is discussed in the following)

$$\frac{|d_j^k|}{|d_{j^\star}^{k+1}|} \approx \frac{|u^{(r+1)}(\xi')|}{|u^{(r+1)}(\xi'')|}\frac{h_{k-1}^{r+1}}{h_k^{r+1}}. \tag{28}$$

Remembering that $h_k = 2h_{k-1}$, the *wavelet* d_j^k can be generated from the generative *wavelet* $d_{j^\star}^{k+1}$ as follows

$$\big|d_j^k\big| \approx C\frac{|d_{j^\star}^{k+1}|}{2^{r+1}} \leqslant C\frac{\varepsilon_{k+1}}{2^{r+1}}. \tag{29}$$

As a consequence, norm $\|u^0 - \hat{u}^0\|_\infty$ is bounded by

$$\big\|u^0 - \hat{u}^0\big\|_\infty \leqslant \max\big\{\big|d_j^1\big|, \dots, \big|d_j^{L-1}\big|\big\} = \big|d_j^1\big| \leqslant C\frac{\varepsilon_2}{2^{r+1}} = \frac{C}{2^{r+4}}\varepsilon. \tag{30}$$

In the case of non-smooth function u, the $r+1$th divided difference can be related to the jump in the pth derivative of the function $[u^{(p)}]$:

$$u\big[\mathcal{S}_j^k; \xi_{2j-1}^{k-1}\big] \approx \begin{cases} \mathcal{O}([u^{(p)}])/h_{k-1}^{r+1-p} & \text{if } r+1 > p,\\ \mathcal{O}(\|u\|_\infty) & \text{if } r+1 < p. \end{cases} \tag{31}$$

In this case, the norm $\|u^0 - \hat{u}^0\|$ remain bounded using the jump in the pth derivative.

This means that, in a smooth region, the *wavelets* decreases with a rate determined by the local regularity and by the order of the interpolation, while in a region near the discontinuity the *wavelet* remains of the same order for all the levels of refinement. This information could permit to build adaptive procedures for evaluating the numerical fluxes in the framework of finite volumes schemes. One could shift from the centered to non-oscillatory schemes (ENO, WENO) in the neighborhood of a discontinuity [29]. In this work, this property is not exploited but it could be used to extend adaptivity of the stochastic space in the semi-intrusive method [21–23] recently proposed. However, in this case the procedure could become highly non-linear.

The estimation of a bounded norm, although with different bounds, guarantees the stability of the algorithm in the sense of the classical MR framework. The prominent difference with the classical framework is that the function is not known at the finest level. This property allows to gain in terms of memory and computational resources. Anyway, it is clear that a proper coarsest level must be chosen, i.e. the parameter m_L must be chosen in order to guarantee the reproduction of the function without the presence of *aliasing* effects, i.e. the discrete values u^{L-1} should contain enough information to recover the original function u^0 without loss of frequency information. From a direct application of the classical Nyquist–Shannon sampling theorem, this should display that the spatial frequency of the level $L-1$ (the second level of the TE algorithm and the last fully evaluated by default) should be sampled with a frequency doubled with respect to the maximum frequency of the signal. In practice, the TE algorithm is designed to relax this condition capturing automatically the regions where the maximum spatial frequencies occur. In particular, this can be valid only if the function reproduced at the level $L-1$ is not aliased, i.e. confused, with its zero frequency counterpart. In this last case, none of the *wavelets* at level L can be activated, thus stopping the algorithm. Different cures could be applied: for example, forcing the activation of *wavelets* if the frequency is zero or introducing a randomization in the process of generating new 'mid' points $xi_j^{k-1} \in GGk-1 \setminus \mathcal{G}^k$. In practice, good results are already obtained with the proposed algorithm and we left to a future work further investigations in this direction. Of course, the estimation proposed in this section is valid only in the case of not aliased representation of the function. Let us focus on the parameter m_0. This parameter fix the maximum resolution that can be reached, when the function is fully discretized at the level $k=0$. This is the same role played by the resolution level in Harten's framework. In practice, the quality of the TE algorithm is limited by the resolution $k=0$ as the solution cannot be improved over this level. At the level $k=0$ is also associated the value of the threshold ε that can be interpreted as the desired accuracy related to the representation of the function at the finest level ε. The same parameter play a role in the bound estimation of both algorithms, the classical Harten MR and the TE.

An important aspect to clarify remains why a novel and different MR approach is required in the context of stochastic differential equations. Let us consider the response of a system dependent from a random parameter (or a vector of parameter) in a steady configuration. The aim of this kind of analysis is the computation of statistics, for example the expectancy.

To compute statistics it is necessary to (numerically) integrate the solution in the space of the parameters. Seeing the cost that can be associated to the computation of the solution in the space of the parameters (imagine the response of a complex numerical code), one issue of UQ is to reduce the computational cost of the global algorithm. In this case, the entire classical MR framework cannot be used. In fact, from a theoretical point of view, it is always possible to compress a well-resolved solution, as for example in image processing applications, but this could be done only after the complete calculation at the finest level. For this reason, the TE algorithm is formulated, because it allows increasing computational efficiency preserving the same saving in term of memory requirement with respect the classical MR approach. In fact, the computation starts with the coarsest level and, only where required, the solution can be refined. At the end of the algorithm, the multiresolution representation will be the same as the classical MR framework, but the number of points in the space of parameters would be, hopefully, less.

In this case, such a strategy should be intended as a non-intrusive UQ method, where the TE algorithm allows the reconstruction of a response surface of a scalar or vectorial outputs. For instance, the output could be a physical quantity related to a complex CFD simulation performed by a code where any modification is forbidden.

This advantage is shown in Fig. 2. The classical framework would consist in nine computations, the entire finest level $k=0$, in order to obtain a multiresolution representation of four point (the two points at the coarsest level $k=3$ and the two *wavelets* activated at the levels $k=2,3$). The same results, in term of multiresolution representation, would be obtained by means of the TE algorithm with only seven computations (the full circles and crossed squares).

Classically the efficiency of the MR approach is measured by a compression ratio μ_{cr} that is computed as the ratio between the number of points in the finest level (N_0+1) and the number of significative *wavelets* ($N_w=\mathrm{card}\{d_j^k\colon |d_j^k|>\varepsilon_k,\ 0\leqslant k\leqslant L-1\}$),[4] i.e. the number of active *wavelets* coefficients

$$\mu_{\mathrm{cr}}=\frac{(N_0+1)}{N_w+(N_L+1)}. \tag{32}$$

For the case reported in Fig. 2 the compression ratio is $\mu_{\mathrm{cr}}=9/4$.

This ratio is always the same for the classical approach and the present strategy as the MR representation is the same in both cases (see the discussion above). However, another ratio can be introduced, i.e. an evaluation compression ratio τ, that measures the computational saving to obtain the MR strategy defined as the ratio between the number of points in the finer level and the number of evaluations needed ($N_{\mathrm{eval}}=\mathrm{card}\{\xi_j^k\colon\ u(\xi_j^k)\ \text{evaluated},\ 0\leqslant k\leqslant L\}$) to construct the MR representation

$$\tau=\frac{(N_0+1)}{N_{\mathrm{eval}}}. \tag{33}$$

In the previous example, reported in Fig. 2, the evaluation compression ratio is $\tau=1.29$. Of course, for the classical MR approach this ratio is always one because the solution must be known at the finest resolution level according to Harten's framework. Even if the compression capabilities of this strategy is the same than the classical approach, in this way it is possible to reduce computational cost for non-necessary functional evaluations. This is a very important property for treating UQ problems where a functional evaluation can be associated to a high computational cost.

For solving efficiently conservation laws systems, some techniques have been proposed ([31] and Cohen et al. [32]) with a high CPU and memory efficiency.

It could be useful to remark here that the TE algorithm differs significantly from other adaptive refinement techniques like, for example, the automatic mesh refinement (AMR) techniques. In the MR context, the solution is refined and an accuracy requirement is fulfilled, not only locally as in AMR, but with respect to the representation at a specific level (the finest). This is the key idea that allows building the time-advancement algorithm described in Section 4. In this context, remark that the strategy, even if weakly, is considered intrusive and the deterministic code itself becomes a part of the TE strategy, when associated to the accuracy preserving time integration strategy described in Section 4.

Let us consider now unsteady problems. In this case, the MR classical framework seems to work properly. At the first time step, the MR algorithm could obtain a multiresolution representation of the initial solution and, then, the adding cost related to non-necessary evaluations would be very small if the initial solution is known analytically. For the subsequent time steps, the first to be really computed, some differences could exist between the MR classical approach and the present one. In fact, as already explained in the previous section, the MR strategy is based on a CFL approach for moving the significative wavelets. Then, if the problem is dominated by an unsteady pdf, the algorithm could move points basing only on the temporal evolution of the solution and not on the evolution of the product with the pdf. For this reason, the classical MR algorithm could totally fail having no CFL condition to follow, i.e. it is not possible to predict the grid from one step to the next one as well as it can be accomplished in the physical space according to the direction of information propagation.

For this reason the time-advancement algorithm, presented in the previous section, is of great interest for unsteady pdf. A specific test-case is presented in Section 6.

[4] Note that in the TE framework here presented, a *wavelet* d_j^k is associated directly to the point ξ_j^k at the finer level, i.e. the level at which the *missing point* is located, instead of the coarser level of the classical MR framework. This reflects on a different *threshold* ε_k definition.

6. Numerical results

In this section, the TE algorithm is applied to several numerical test-cases in order to check its accuracy and convergence rate with respect to some classical stochastic methods, such as quasi-Monte Carlo and Polynomial Chaos [6]. In the case of probability distribution not belonging to the so-called Wiener–Askey scheme [3], a PC method is used in order to evaluate the statistics as in a collocation method (the function is multiplied by the pdf). This correspond in practice to the computation of the first coefficient of the polynomial expansion that is equal to the expectation. Let us consider a truncated polynomial expansion in terms of Legendre polynomials Ψ_k:

$$f(\xi) = \sum_{i=0}^{P} \beta_k \Psi_k(\xi), \tag{34}$$

where, in the case of full tensorization, the number of term $P+1 = (n_0+d)!/(n_0!d!)$, in which n_0 is the total polynomial degree and d is the number stochastic dimensions. In the 1D case, $P = n_0$ and in all the numerical results presented in this work the number of simulations N relative to the PC is $N = n_0 + 1$. The coefficients β_k for the expansion are computed exploiting the orthogonality of the Legendre basis with respect to a uniform pdf. The numerical integration is performed with a Gauss–Legendre quadrature with the integrand function evaluated at the zeros of the polynomial basis.

In the case of probability distribution not belonging to the Wiener–Askey scheme, the polynomial expansion (34) is used to represent the product between the function and the probability density function. Computing the variance requires the expectancy of the squared function times the pdf; this is accomplished expanding the function as follows

$$f^2(\xi)p(\xi) = \sum_{i=0}^{P} \bar{\beta}_k \Psi_k(\xi), \tag{35}$$

from which the $\bar{\beta}_0$ coefficient, i.e. the expectancy, can be extracted.

Finally, the following relations allow to compute the expectancy and the variance of the function $f = f(\xi)$ with ξ described by a non-classical pdf $p(\xi)$:

$$\begin{aligned} \mathcal{E} &= \int_{\Xi} f(\xi)p(\xi)\,\mathrm{d}\xi = \beta_0, \\ \mathrm{Var} &= \int_{\Xi} \bigl(f(\xi) - \mathcal{E}\bigr)^2 p(\xi)\,\mathrm{d}\xi = \int_{\Xi} f^2(\xi)p(\xi) - \mathcal{E}^2 = \bar{\beta}_0 - \beta_0^2. \end{aligned} \tag{36}$$

This procedure is indicated generally as Polynomial Chaos Method in the following.

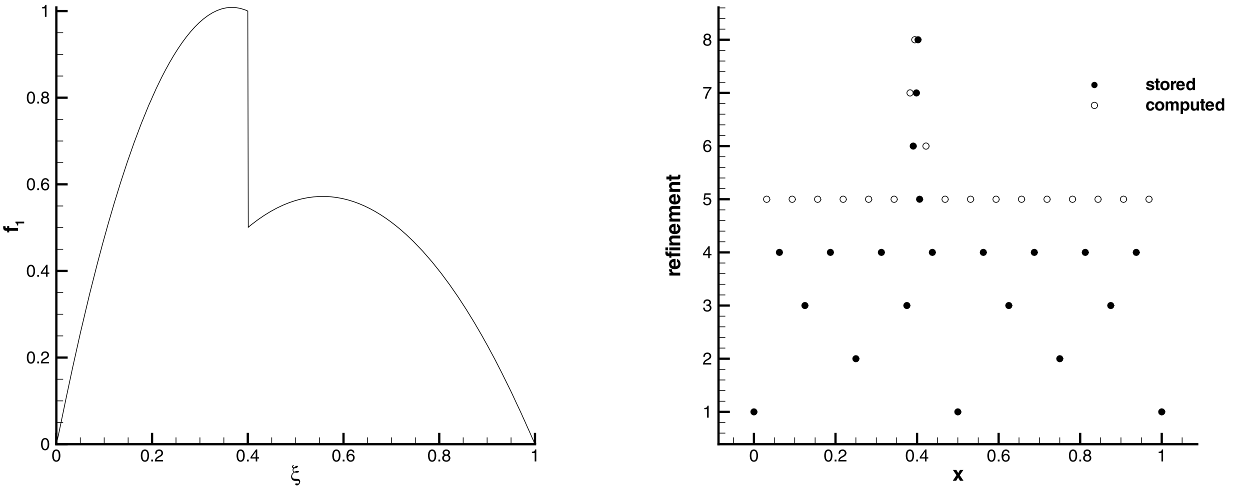
First, some steady algebraic problems (Section 6.1) are considered where analytical discontinuous functions are evaluated in terms of their expectancy and variance; for this case the performances in terms of compression and evaluation ratios are also evaluated. The capability of the TE algorithm to preserve accuracy for time-evolving solutions is displayed in two test-cases. The first-one (see Section 6.2) deals with an ordinary differential equation (ODE), i.e. taken from [6] but with some modifications in order to obtain a more stiff ξ–t pattern of the function. The second test-case (see Section 6.3) for checking the convergence properties in long-time integration problem is the so-called Kraichnan–Orszag 1D ODE, a well-known problem in literature for testing UQ methods properties. Different kinds of pdf, i.e. uniform and discontinuous, are considered. Finally, a simplified model for aeroelastic study (Section 6.6), a two degree-of-freedom typical wing section coupled with a quasi-steady strip theory model for aerodynamics, is proposed to compute the statistics of the motion considering uncertainties on mass properties for discontinuous probability distribution. For all the examples, exhaustive comparisons with quasi-Monte Carlo and Polynomial Chaos are performed.

6.1. Steady problems

Let us consider a function of the form $f = f(\xi)$, where the parameter $\xi \in \Xi$ takes uniform values between 0 and 1, i.e. $\xi \sim \mathcal{U}[0,1]$. The aim is to compute expectancy $\mathcal{E}$ and variance Var for f according to the following definitions

$$\begin{aligned} \mathcal{E} &= \int_{\Xi} f(\xi)p(\xi)\,\mathrm{d}\xi, \\ \mathrm{Var} &= \int_{\Xi} \bigl(f(\xi) - \mathcal{E}\bigr)^2 p(\xi)\,\mathrm{d}\xi. \end{aligned} \tag{37}$$

All the numerical integrals are computed with the trapezoidal rule on the points distribution generated by means of the TE algorithm. Results are compared with respect to the reference-solutions obtained from Monte Carlo and Polynomial Chaos methods.



(a)

(b)

Fig. 5. Model function f_1 (a) with its MR representation on the left (b).

Two model functions, with one or two discontinuities in the stochastic space, are considered. The function f_1 (see Fig. 5(a)), is a piecewise smooth function with one discontinuity located at $\xi = 2/5$

$$f_1(\xi) = \begin{cases} -\frac{15}{2}\xi^2 + \frac{11}{2}\xi & 0 \leqslant \xi \leqslant 2/5, \\ -\frac{35}{12}\xi^2 + \frac{13}{4}\xi - \frac{1}{3} & 2/5 < \xi \leqslant 1. \end{cases}$$

The second function f_2 (see Fig. 7(a)) is constituted by a sinus function and a fourth-order polynomial form, then contains two discontinuities at $\xi = 1/5$ and $\xi = 3/4$:

$$f_2(\xi) = \begin{cases} 10\sin(\xi\pi) & \xi \leqslant 1/2 \text{ or } \xi \geqslant 3/4, \\ 10\xi^4 + 79/4 & \text{otherwise}. \end{cases} \tag{38}$$

Starting with the coarsest level $m_L = 1$ and with a maximum level between 3 and 12, using a threshold equal to $\varepsilon = 10^{-1}$, the TE algorithm is used to generate the points distribution. The MC method is used to generate reference-values for the expectancy and the variance with a number of points bounded between 5 and 101, while a PC approach is employed to evaluate the integrals with a degree between 5 and 100 with a step equal to 5. The results are evaluated in terms of percentage errors, computed as follows

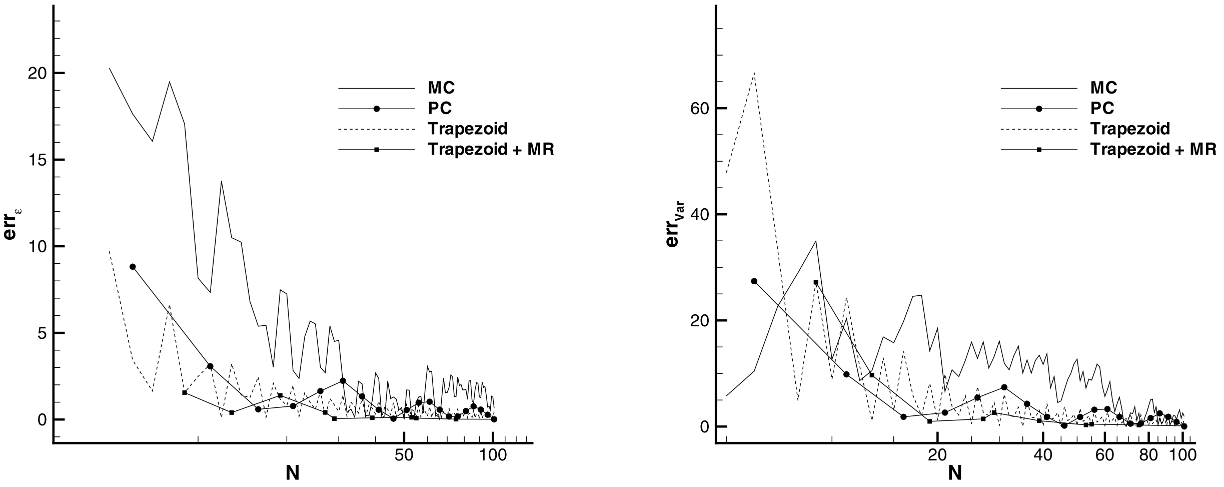
$$\begin{aligned} \mathrm{err}_{\mathcal{E}} &= \frac{|\mathcal{E} - \mathcal{E}_{\mathrm{exact}}|}{\mathcal{E}_{\mathrm{exact}}} 100, \\ \mathrm{err}_{\mathrm{Var}} &= \frac{|\mathrm{Var} - \mathrm{Var}_{\mathrm{exact}}|}{\mathrm{Var}_{\mathrm{exact}}} 100. \end{aligned} \tag{39}$$

In Fig. 6, we reported the errors for the mean $\mathcal{E}$ and for the variance Var with respect to the total amount of evaluations required N to compute statistics on the left and on the right, respectively.

All the methods exhibit a non-monotone convergence to the exact value both for $\mathcal{E}$ and Var. The present MR algorithm allows reaching lower levels of error with a fixed number of evaluations N and, moreover, allows obtaining the smoothest decrease of the error with the number of evaluations both for mean and variance. The MC method displays strong oscillations with respect to all the other methods while the PC exhibit a faster convergence with respect to the MC, but is affected by numerous oscillations.

In Fig. 6, the results of applying the trapezoidal rule without the TE algorithm, are also reported. It is possible to appreciate that the good performances obtained with TE are not related to the choice of quadrature formula (seeing that trapezoidal rule is not so accurate), but the efficiency in terms of computational cost can be attributed globally to the proposed algorithm. The trapezoidal rule results display several oscillations even if the errors are lower with respect to both the MC and PC methods.

To better understand the TE algorithm, in figure (b), the pattern of the computed points and the activated *wavelets* are reported. The maximum level was fixed, in this case, to $m_{\max} = 8$ equal to 257 evaluations. As shown in figure (b), the adaptive algorithm do activate all the points until to the third refinement. At the fourth refinement, the accuracy requirement ($\varepsilon = 10^{-1}$) is reached for all the points but not for the point closer to the discontinuity. Then, the algorithm stops after that the maximum level is reached. It is remarkable that the stored points in Fig. 5(b) are the set of wavelets of the Harten framework after the truncation procedure. However, the TE algorithm needs some extra points to generate



(a) (b)

Fig. 6. Errors for the mean (a) and variance (b) for the first function f_1.

Table 1
Results for the function f_1. The compression ratio μ_{cr} and the evaluation compression ratio τ are computed as shown in Eqs. (32) and (33), respectively.

$m_{\max}$	N_{w}	N_{eval}	μ_{cr}	τ	err_{L_1}	err_{L_∞}
3	7	9	0.1285E+01	0.1000E+01	0.3331E−15	0.9252E−16
4	8	13	0.2125E+01	0.1308E+01	0.1139E−01	0.2681E−02
5	11	19	0.3000E+01	0.1737E+01	0.7324E−02	0.2022E−02
6	15	27	0.4333E+01	0.2407E+01	0.2848E−02	0.1236E−02
7	16	29	0.8063E+01	0.4448E+01	0.2848E−02	0.1384E−02
8	21	39	0.1224E+02	0.6590E+01	0.1831E−02	0.7150E−03
9	28	53	0.1832E+02	0.9679E+01	0.7121E−03	0.3828E−03
10	29	55	0.3534E+02	0.1863E+02	0.7121E−03	0.3854E−03
11	39	75	0.5254E+02	0.2732E+02	0.4578E−03	0.1860E−03
12	52	101	0.7879E+02	0.4056E+02	0.1780E−03	0.1004E−03

the MR representation, i.e. the wavelets not activated (the withe circles). At the end of the procedure, the compression ratio μ_{cr} is equal to $\mu_{\text{cr}} = 257/21 = 12.24$. Moreover, the TE algorithm allows saving computational cost reaching the MR representation without knowing all the points at the maximum level ($2^8 + 1 = 257$ in this case). The evaluation ratio τ is equal to $\tau = 257/39 = 6.59$.

Table 1 summarizes the results obtained for the function f_1 starting with a coarsest level equal to 1 and with a maximum level between 3 and 12. Errors in the norms L_1 and L_∞ are reported in terms of the number of the activated *wavelets* N_w, the number of the evaluated points N_{eval} and the compression μ_{cr} and evaluation τ ratios. Errors in the norms L_1 and L_∞ are computed as follows

$$\begin{aligned} \text{err}_{L_1} &= \|f^0 - \hat{f}\|_{L_1} = \frac{1}{N}|f_i^0 - \hat{f}_i|, \\ \text{err}_{L_\infty} &= \|f^0 - \hat{f}\|_{L_\infty} = \max_i |f_i^0 - \hat{f}_i|, \end{aligned} \tag{40}$$

where f^0 is the function at the finest level $k = 0$ and $\hat{f}$ is the truncated function, i.e. the function evaluated only in the set of points corresponding to the activated wavelets.

For the function f_2 (see Eq. (38)), the pattern of the MR representation is reported in Fig. 7(b) using a threshold equal to $\varepsilon = 10^{-1}$ and the coarsest and the maximum level equal to 1 and 8, respectively. Remark that the TE algorithm follows the two discontinuities, where more points are generated until reaching the maximum level. The advantage of such a distribution of points can be clearly seen in Fig. 8, where the percentage errors for the mean $\mathcal{E}$ and variance Var are reported.

Also in this case the error displays strong oscillations, but the TE algorithm shows a monotone decrease of the errors with respect to the number of evaluations. For the other methods, convergence is attained only in terms of mean values. Even in this case, the effect of the TE algorithm in the points distribution allows obtaining better results with respect to the same trapezoidal quadrature technique. The detailed results in term of compression μ_{cr} and evaluation τ ratios and the corresponding errors in norm L_1 and L_∞ are reported in Table 2.

In the next section some unsteady problems with continuous and discontinuous pdf are taken into account.

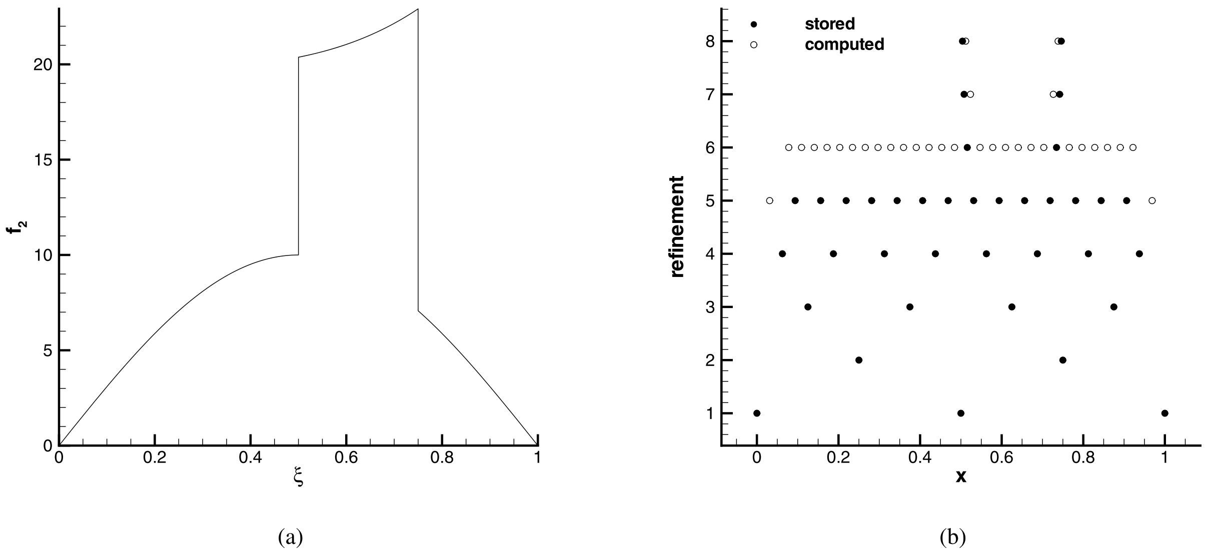


Fig. 7. Function f_2 (a) and its MR representation (b).

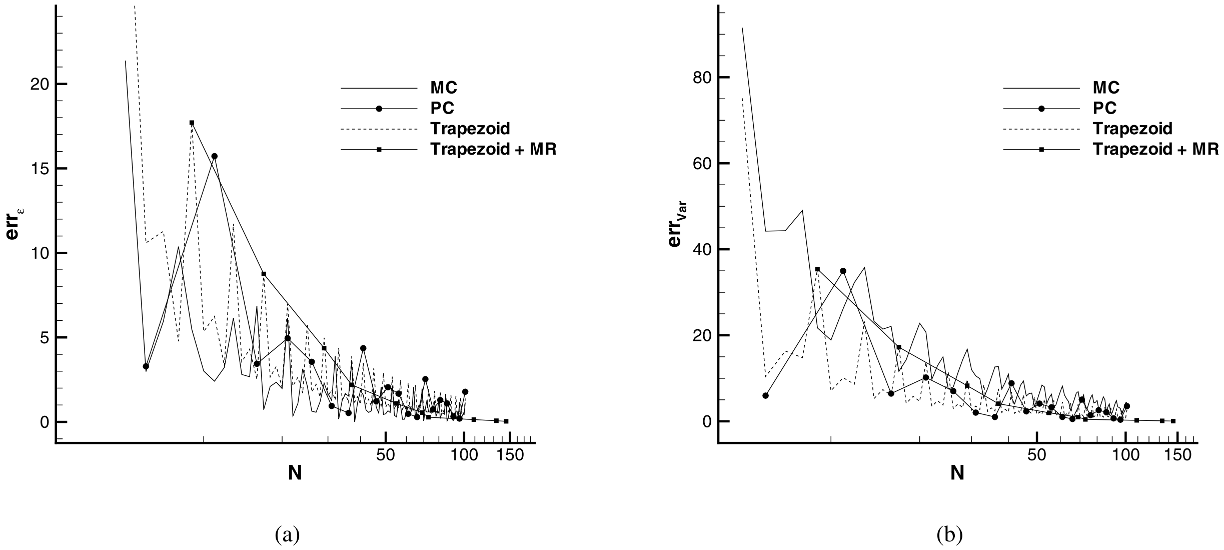


Fig. 8. Errors of the mean (a) and variance (b) for the function f_2.

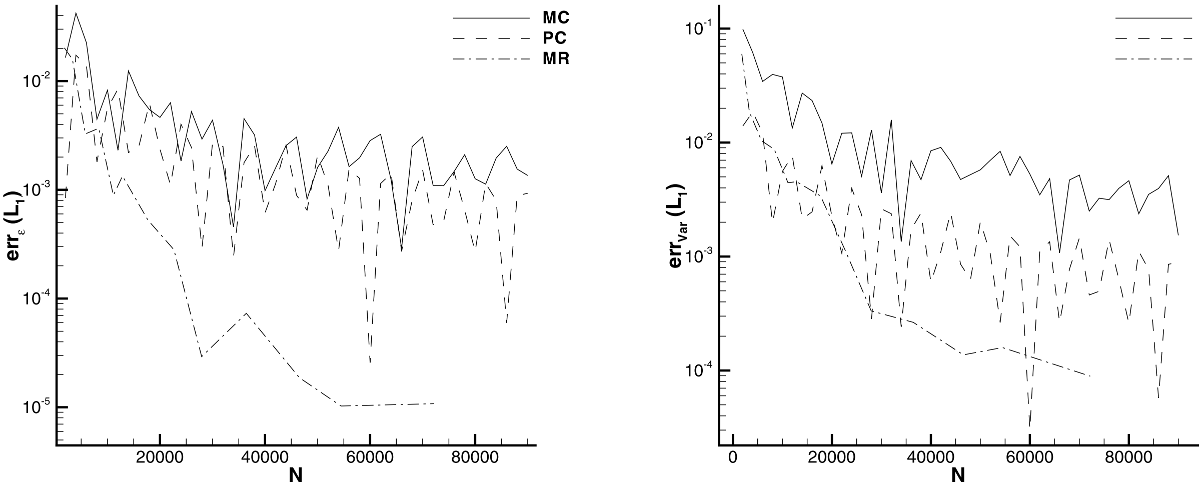
Table 2
Results for the function f_2. The compression ratio μ_{cr} and the evaluation compression ratio τ are computed as shown in Eqs. (32) and (33), respectively.

$m_{\max}$	N_w	N_{eval}	μ_{cr}	τ	err_{L_1}	err_{L_∞}
3	9	9	0.1000E+01	0.1000E+01	0.3997E−14	0.1283E−14
4	15	17	0.1133E+01	0.1000E+01	0.2487E−13	0.4415E−14
5	17	29	0.1941E+01	0.1138E+01	0.1398E−01	0.1133E−02
6	21	37	0.3095E+01	0.1757E+01	0.1203E−01	0.2795E−02
7	30	55	0.4300E+01	0.2345E+01	0.6193E−02	0.1551E−02
8	37	69	0.6946E+01	0.3725E+01	0.3011E−02	0.1019E−02
9	39	73	0.1315E+02	0.7027E+01	0.3011E−02	0.1070E−02
10	57	109	0.1798E+02	0.9404E+01	0.1611E−02	0.4983E−03
11	69	133	0.2970E+02	0.1541E+02	0.8033E−03	0.3005E−03
12	75	145	0.5463E+02	0.2826E+02	0.7529E−03	0.2647E−03

6.2. A scalar ordinary differential equation

In this section, the time-advancing strategy presented in Section 4 is applied to some ordinary differential equations. In this section, the scalar case is analyzed, while the results for the vectorial cases are reported in Sections 6.3 and 6.6.

The first ODE example is extracted from [6] with some slight modifications in order to achieve a variable final state, as follows



(a) (b)

Fig. 9. Errors for the mean (a) and variance (b) in the L_1 norm for Eq. (41) with $m_L = 1$ and $\varepsilon = 10^{-2}$.

$$\begin{cases} \dfrac{\mathrm{d}\rho}{\mathrm{d}t} = \alpha(\bar{\rho} - \rho) - \gamma\rho - \beta(\rho - \bar{\rho})\rho^2, \\ \bar{\rho} = 1 + \dfrac{1}{2}\sin(5\omega + 8/5), \\ \beta = 20\omega, \end{cases} \tag{41}$$

where $\alpha = 1$, $\gamma = 0.01$ and $\omega \sim \mathcal{U}[0,1]$. The original problem [6] is related to the evolution of the surface coverage ρ for a given species. A discontinuous initial solution in the stochastic space is chosen in order to obtain a discontinuous response

$$\rho(t = 0) = \begin{cases} 3/4 & \text{if } 0.3 < \omega < 0.7, \\ 0 & \text{otherwise}. \end{cases} \tag{42}$$

In this case of unsteady problems, the aim is to compute the temporal evolution of the mean and the variance following

$$\begin{aligned} \mathcal{E}(t) &= \int_{\Xi} f(\xi, t) p(\xi)\,\mathrm{d}\xi, \\ \mathrm{Var}(t) &= \int_{\Xi} \bigl(f(\xi, t) - \mathcal{E}(t)\bigr)^2 p(\xi)\,\mathrm{d}\xi. \end{aligned} \tag{43}$$

The MC converged solution, $\rho_{\mathrm{ref}}(t)$, is retained as reference for mean and variance solutions. The errors are computed as follows

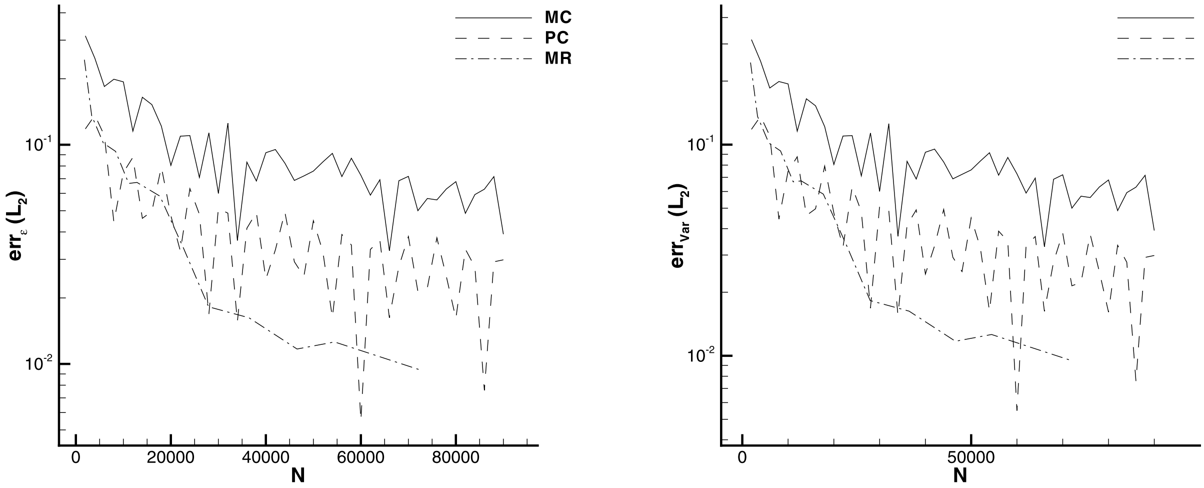
$$\mathrm{err}_{\mu^m}|_{L_p} = \bigl\|\mu^m(\rho, t) - \mu^m(\rho_{\mathrm{ref}}, t)\bigr\|_{L_p} = \left(\frac{1}{N_t}\sum_{i=1}^{N_t}\left|\frac{\mu_i^m(\rho, t) - \mu_i^m(\rho_{\mathrm{ref}}, t)}{\mu_i^m(\rho_{\mathrm{ref}}, t)}\right|^p\right)^{1/p}, \tag{44}$$

where μ^m are the statistic moments, i.e. mean $\mathcal{E}$ or variance Var, and $p = 1, 2$ is referred to the spaces L_1 or L_2, respectively. The number of time steps is indicated with N_t and the total time of the simulation is equal to $T = 2$ where the time step is equal to $\Delta t = 0.01$ ($N_t = 200$). Moreover, the L_∞ norm is computed as

$$\mathrm{err}_{\mu^m}|_{L_\infty} = \bigl\|\mu^m(\rho, t) - \mu^m(\bar{\rho}, t)\bigr\|_{L_\infty} = \max_i\left|\frac{\mu_i^m(\rho, t) - \mu_i^m(\rho_{\mathrm{ref}}, t)}{\mu_i^m(\rho_{\mathrm{ref}}, t)}\right|, \tag{45}$$

for both mean and variance.

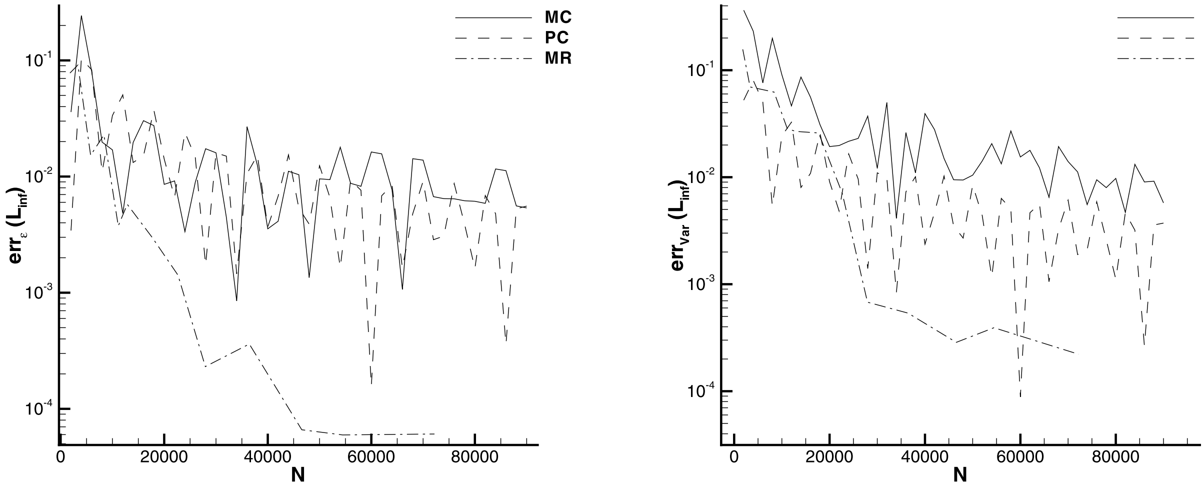
In Figs. 9, 10 and 11, the results for the errors in L_1, L_2 and L_∞ are reported for MC, PC and TE. The number of points N in this case is equal to the total number of points in the grid $\omega - t$, i.e. the product of the number of points in the stochastic space N_ξ and the number of time intervals $N_t = 200$ employed $N = N_\xi \times N_t$. The reference solution is obtained with a number of stochastic points equal to $N_\xi = 2.5 \times 10^6$. Several set of points are chosen in order to study the convergence of the different methods, in particular N_ξ is varied between 10 and 450 with step of 10 for both MC and PC, while computations are performed with $\varepsilon = 10^{-2}$ and a maximum levels between 3 and 15. The integration in time was performed by means of an explicit Euler formula.



(a)

(b)

Fig. 10. Errors for the mean (a) and variance (b) in the L_2 norm for Eq. (41) with $m_L = 1$ and $\varepsilon = 10^{-2}$.



(a)

(b)

Fig. 11. Errors for the mean (a) and variance (b) in the L_∞ norm for Eq. (41) with $m_L = 1$ and $\varepsilon = 10^{-2}$.

The TE algorithm allows reaching a lower level of error for the same number of points. In the case of L_∞ norm for the variance, at the same number of points it corresponds an order of magnitude of the error inferior up to three degrees of magnitude. The convergences plots exhibit the smoothest behaviors with respect to MC and PC despite to the presence of two discontinuities in the function $\rho(\xi, t)$.

A plot of the function $\rho(\omega, t)$ is reported in Fig. 12(a) displaying only the evaluated points. The same plot is reported in two dimensions (Fig. 12(b)), where the time evolution of the evaluations is clearer. In both cases, the maximum level was fixed to 8 with $m_l = 1$ and $\varepsilon = 10^{-2}$. It is possible to recognize the two discontinuities located in $\omega = 0.3$ and $\omega = 0.7$ where more points are collocated by the TE algorithm. One discontinuity at $\omega = 0.3$, disappears before the second one. At $t = 1$, the TE algorithm is capable to respect the accuracy requirements without collocating more points in these ω stations because the function reaches a smoother state.

The behavior shown in Fig. 12 displays the refinement/derefinement capability of the algorithm to move and collocate points in the stochastic space with respect to the time evolution. For example, in Fig. 13, the evolution of the evaluated points in time is reported employing a maximum level equal to 8, corresponding to a total number of evaluations equal to 257, with $m_l = 1$ and $\varepsilon = 10^{-2}$. It is evident that in proximity of $t = 1$, when both the discontinuity disappear, the number of points changes abruptly. However, after $t = 1$, the development of a region characterized by higher gradient (see Fig. 12(a)), requires a greater number of points as it is evident in Fig. 13.

The pattern of the evaluated points and the activated ones, with respect to the time evolution, is reported in Fig. 14. The parameters are the same of the previous cases, i.e. $m_l = 1$, $m_{\max} = 8$ and $\varepsilon = 10^{-2}$. Obviously, due to the discontinuities, the TE algorithm generates all the levels up to $m_{\max}$.

Now, let us apply the TE algorithm to a vectorial case, i.e. the stiff Kraichnan–Orszag problem in one stochastic dimension.

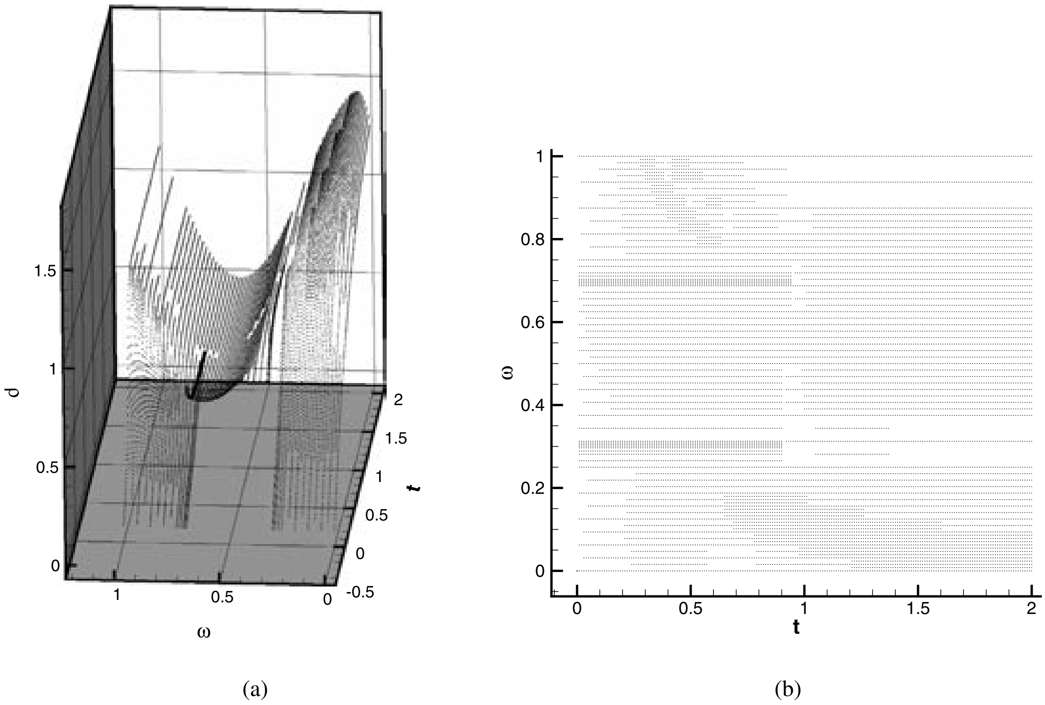


Fig. 12. Patterns of the evaluations of $\rho(\omega, t)$ in the space ω–t in a 3D (a) and 2D (b) frames.

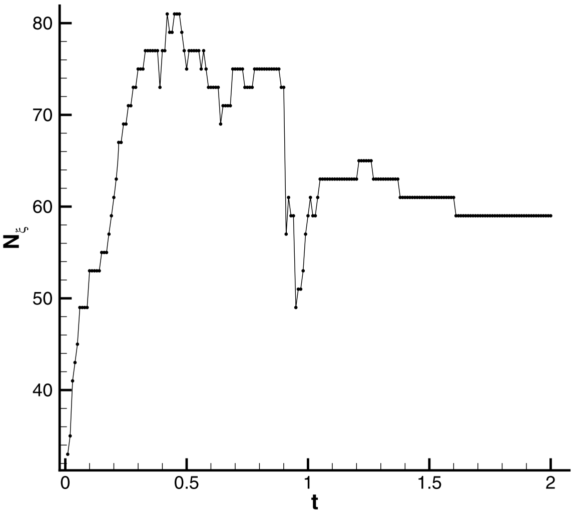


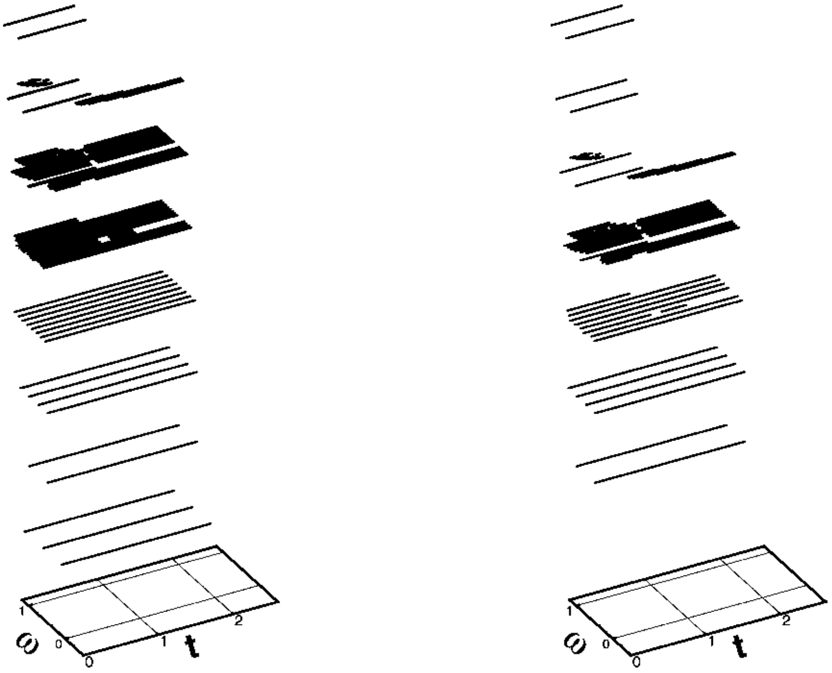
Fig. 13. Time evolutions of the evaluations.

6.3. A time dependent vectorial ODE: the Kraichnan–Orszag problem

The so-called Kraichnan–Orszag problem was proposed in 1967 [33] by Orszag as a three mode problem which can be seen as an inviscid turbulence model given by a set of ordinary differential equations. Actually this differential model is considered a stiff problem for the solution of stochastic problems, due to its high non-linearity. The original system, rotated by $\pi/4$ around the axis y_3 can be written following [4] as

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}t} = y_1 y_3, \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}t} = -y_2 y_3, \\ \dfrac{\mathrm{d}y_3}{\mathrm{d}t} = -y_1^2 + y_2^2. \end{cases} \tag{46}$$

To formulate the 1D problem, Eq. (46) must be correlated with the following (uncertain) initial condition $\mathbf{y}(t = 0) = (1, 0.1\xi, 0)^{\mathrm{T}}$, in which the parameter ξ is uniformly distributed between -1 and 1, $\xi = 2\omega - 1$, where $\omega \sim \mathcal{U}[0, 1]$. The



(a) (b)

Fig. 14. Pattern of the computed (a) and activated (b) points in time for the application of the TE algorithm on Eq. (41).

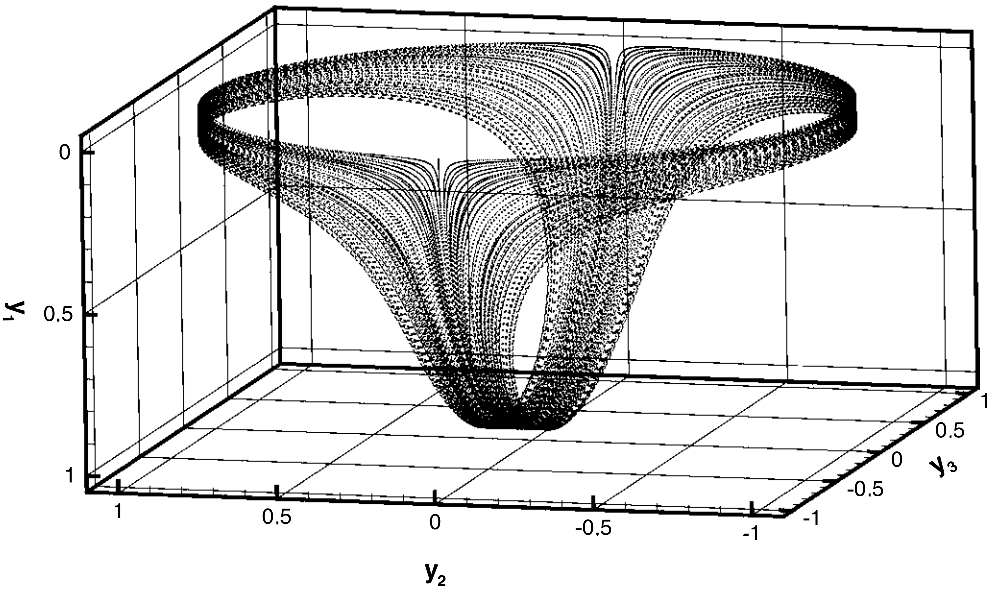


Fig. 15. Pattern of the computed points for the solution of Eq. (41).

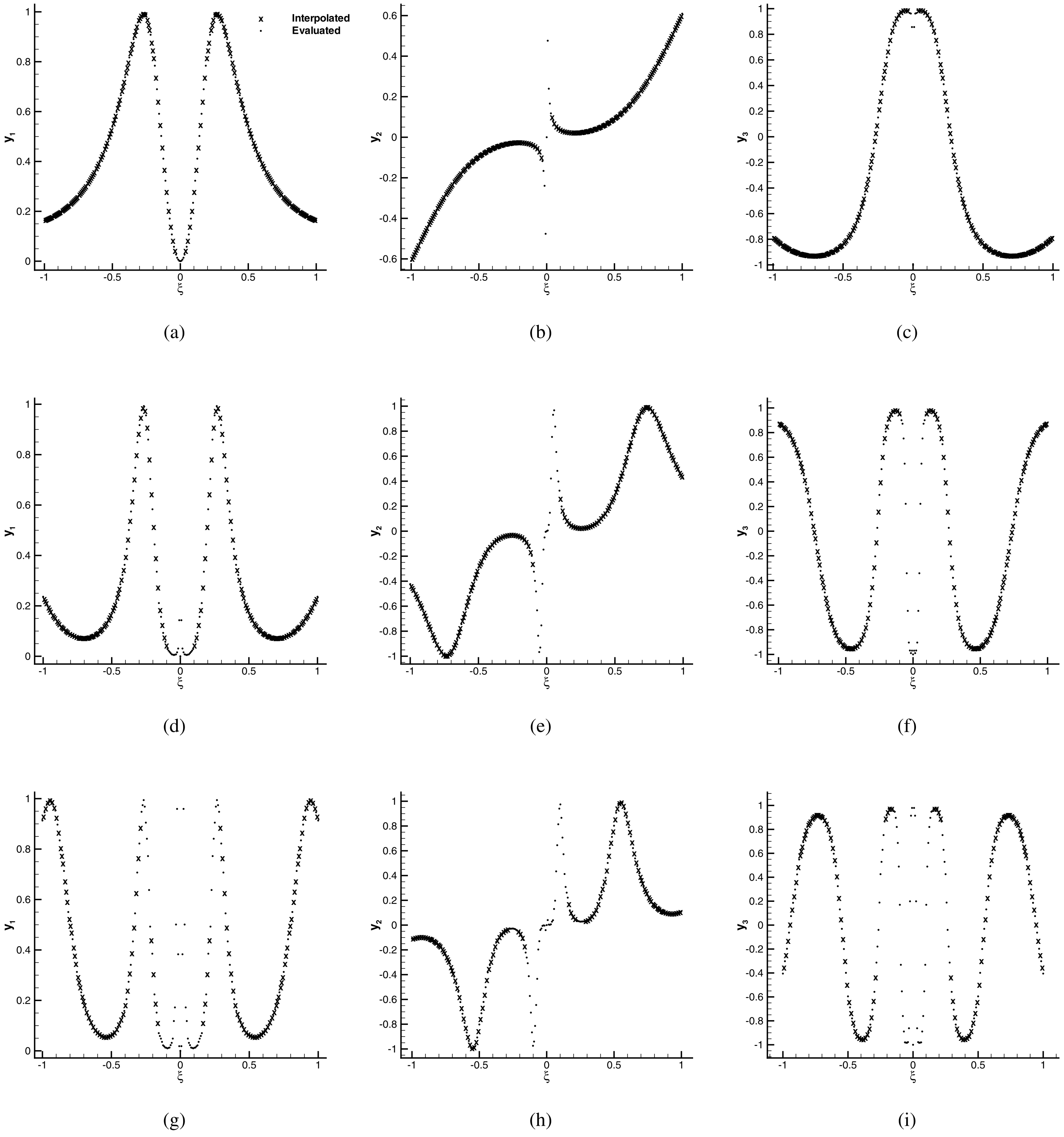
numerical integration scheme used in this work is the classical Runge–Kutta (RK4) with a time step of 0.05, chosen after a convergence study not reported here for brevity.

The pattern of the solution is reported (in the state space) in Fig. 15, where the TE algorithm performed with a maximal level equal to 8 starting with $m_L = 4$ and a threshold $\varepsilon = 10^{-1}$.

The evolution of the three variables in time is reported in Fig. 16, where the TE algorithm is applied with $m_{\max} = 8$, $m_L = 4$ and $\varepsilon = 10^{-1}$. Remark that this set of parameter is chosen only to show qualitatively the pattern of points even if better results could be obtained using a higher maximal level $m_{\max}$.

Observing Fig. 16, it is evident that discontinuities occur crossing the plane at $\xi = 0$. Moreover, the variables distributions are even with respect to the axis $\xi = 0$ for the variables y_1 and y_3 while is odd for the variable y_2. This behavior produces a zero mean for the variable y_2.

The global behavior of the mean and variance in time for the three variables has been reported in Fig. 17. The number of simulations used by the TE algorithm with $\varepsilon = 10^{-1}$, $m_L = 4$ and $m_{\max} = 8$ is equal to $N_\xi = 89$. In order to assess the



(a) (b) (c)

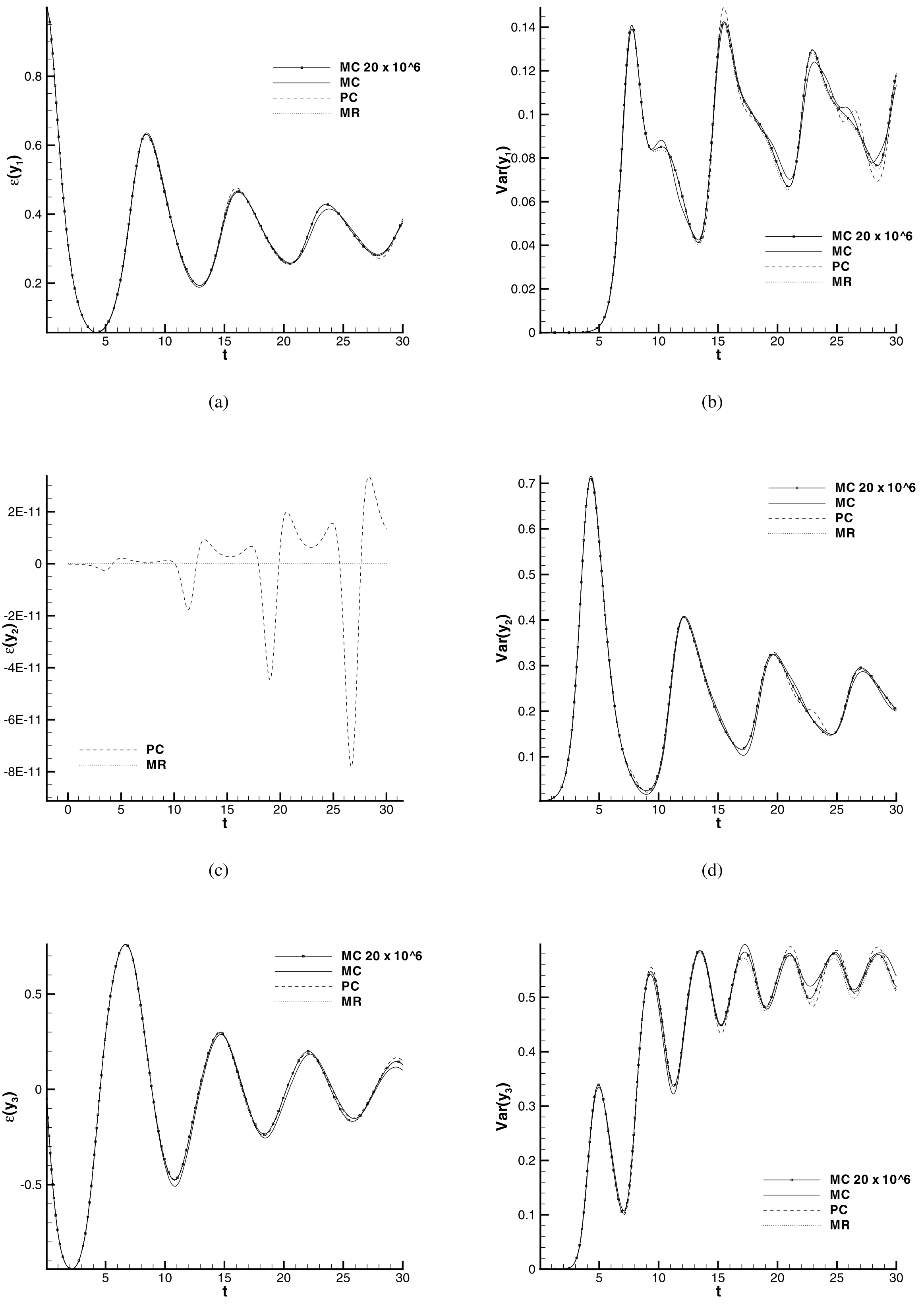
(d) (e) (f)

(g) (h) (i)

Fig. 16. Time evolution of the variables y_1 (left), y_2 (middle) and y_3 (right) at different time steps: $t = 10$ ((a), (b) and (c)), $t = 20$ ((d), (e) and (f)) and $t = 30$ ((g), (h) and (i)). x interpolated points and $\cdot$ evaluated ones.

results obtained for the mean and variance of the three variables, the PC and MC methods are applied by using the same number of points. The statistics for the three variables are reported in Fig. 17, where the reference solution obtained with $N_\xi = 20 \times 10^6$ is also reported. Concerning the variable y_2, the MC is not capable to predict a zero-value at the machine accuracy even in the case of the reference solution ($N_\xi = 20 \times 10^6$); in particular, the solution for the reference case is of the order of 10^{-8}, while for the case $N_\xi = 89$ is of the order of 10^{-3}. For this reason, only the PC and MR results are reported in Fig. 17(c). Remark that only the TE algorithm is capable to compute zero-values, accurate at the machine accuracy.

For this vectorial ODE case, the error in norms are computed for the three variables according to the definitions (44) and (45). We remark that normalizing the norms for the mean is not easy because the reference value could be equal, or almost equal to zero. In particular $\mathcal{E}(y_2)$ should be exactly zero, while $\mathcal{E}(y_3)$ cross periodically zero. Moreover, the variances for all



(a)

(b)

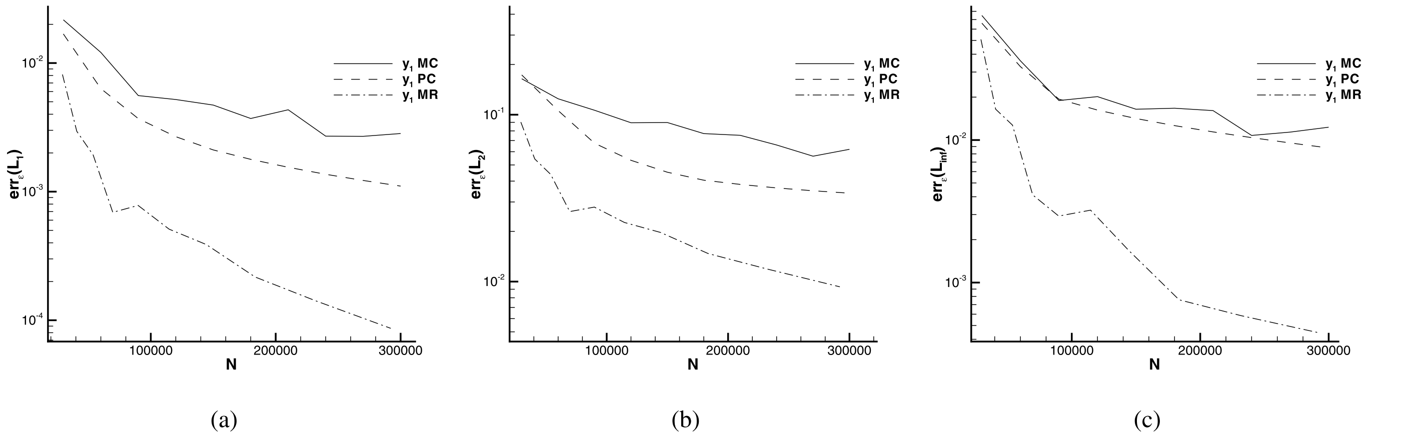
(c)

(d)

(e)

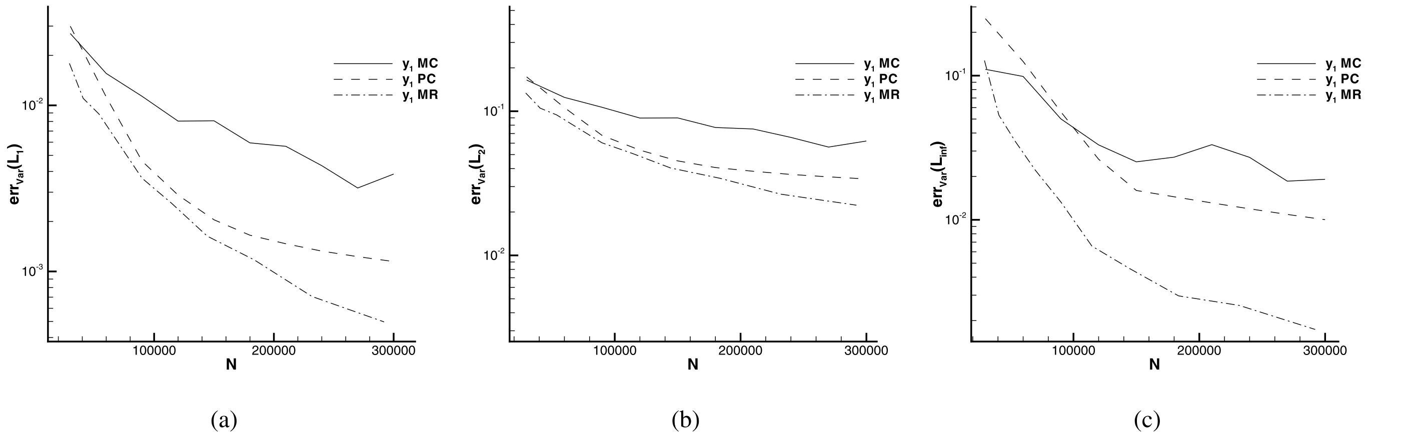
(f)

Fig. 17. Time evolution of the mean (left) and variance (right) for the three variables y_1 (up), y_2 (middle) and y_3 (bottom).



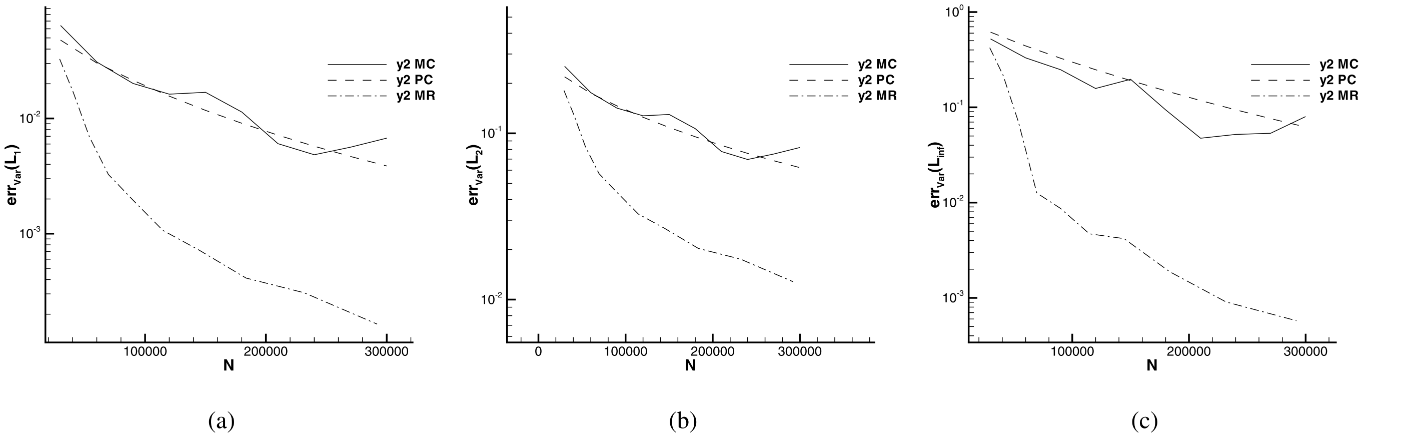
(a) (b) (c)

Fig. 18. Error norms of the mean of the variable y_1 for the 1D Kraichnan–Orszag problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Fig. 19. Error norms of the variance of the variable y_1 for the 1D Kraichnan–Orszag problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Fig. 20. Error norms of the variance of the variable y_2 for the 1D Kraichnan–Orszag problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

the variables start from zero. For these reasons, we report in Fig. 18 the results for the mean $\mathcal{E}(y_1)$ and all the variances, but in this last case the norms are computed in a bound time interval $[8, 30]$ (as chosen by [4]). Note that the value N_t considered in Eqs. (8) is, nevertheless, equivalent to the overall time interval $[0, 30]$.

In Fig. 18, the errors for the mean of the variable y_1 are reported measured in the L_1, L_2 and L_∞ spaces.

In Figs. 19, 20 and 21, the error norms of the variance in the time interval $[8, 30]$ (and in the L_1, L_2 and L_∞ spaces) are reported for the variables y_1, y_2 and y_3, respectively. As it can be observed, MR results display best performances with respect to PC and MC.

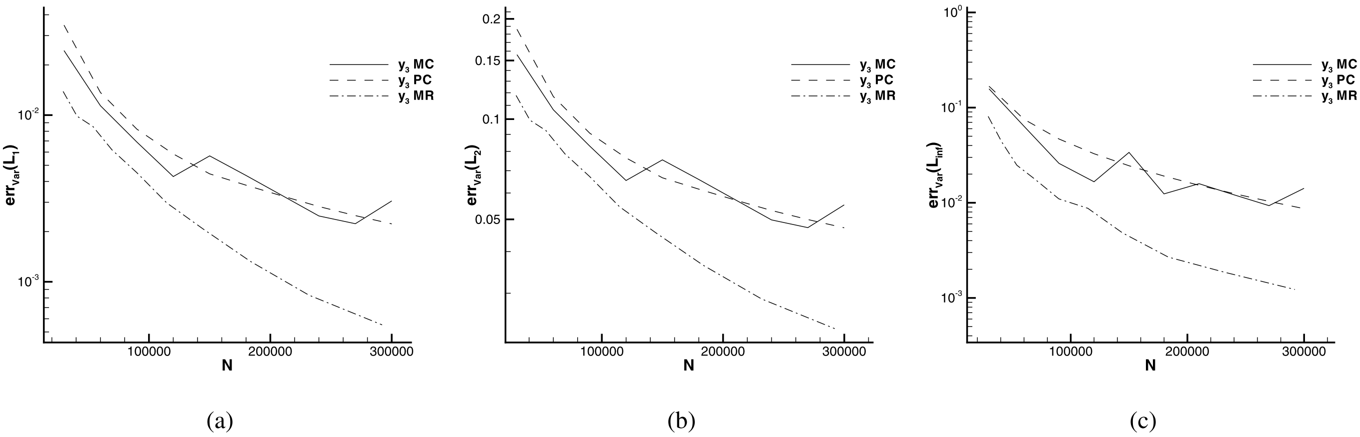


Fig. 21. Error norms of the variance of the variable y_3 for the 1D Kraichnan–Orszag problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

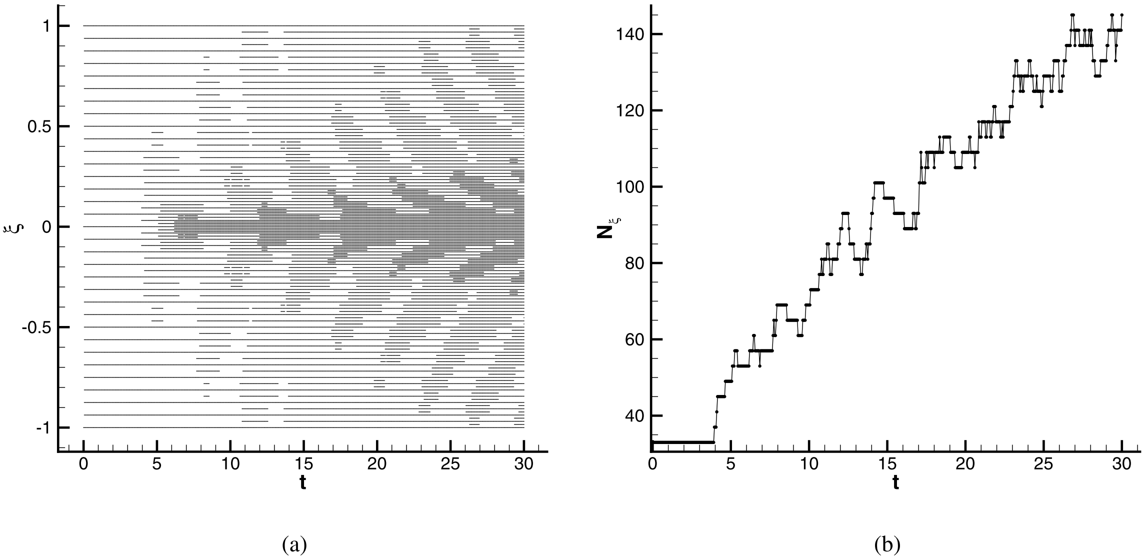


Fig. 22. Patterns of the evaluations in the space t–ω (a) and number of points in the stochastic space N_ξ employed by the TE algorithm with $m_{\max} = 8$ (b).

The points distribution, where the function is evaluated, are reported in Fig. 22. Even in this vectorial case, the TE algorithm displays good derefinement properties. Therefore, the points are distributed in the regions of higher gradients, or discontinuity and locally, between successive time levels (see Fig. 22(b)).

All the results of this section were obtained for stochastic parameters characterized by a uniform distribution. A more challenging problem, where a time varying and discontinuous pdf is considered, is reported in Section 6.5.

6.4. Some remarks on the long-time integration problem

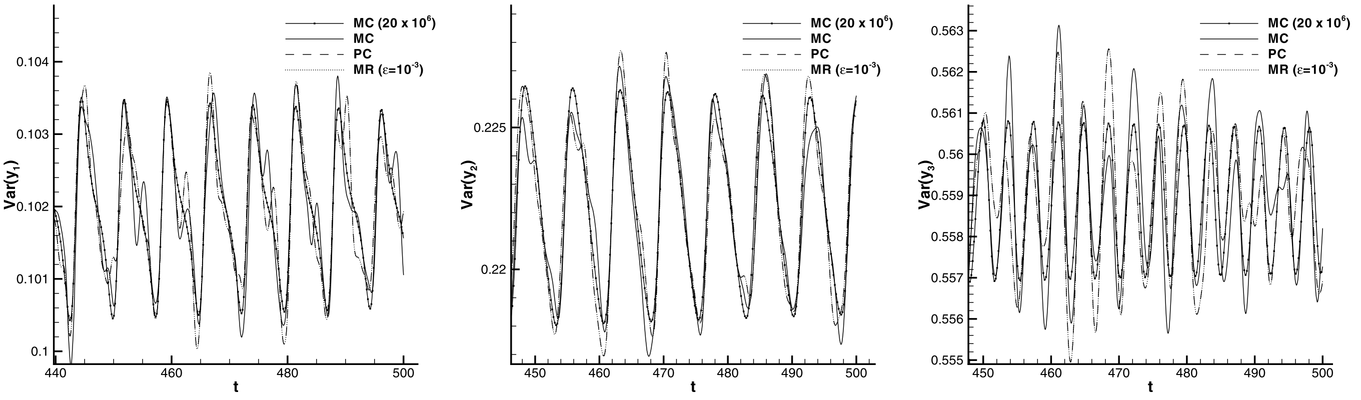
The aim of this section is to show how the TE algorithm can tackle the long-time integration problem. For this reason, the 1D Kraichnan–Orszag problem, with uniform probability density function for the random parameter ξ, is been solved for a time up to 500. Remark that intrusive techniques like the PC, as demonstrated in [4], fails to perform this kind of computation, even for smaller time. The proposed approach can lead to the correct evolution in time of the statistics only if the error cumulated at each time step remains bounded. As described in Section 4, the time integration procedure to advance in time allows moving a compressed solution with a prescribed accuracy requirement. These accuracy requirements allow bounding the error that can accumulate during several time integrations: the final results is a strong decreasing of the number of total evaluations employed to reproduce the exact solution even for long time.

As in the previous case, the MC method, with $N_\xi = 2 \times 10^6$ points, is used as reference solution.

In Fig. 23, the time evolution of the variance for the three variables y_1, y_2 and y_3 is reported for the last part of the simulation, i.e. $t \in [450, 500]$. The results in Fig. 23 are obtained employing the TE algorithm with $m_L = 4$, a maximum level with $m_{\max} = 12$ and a threshold $\varepsilon = 10^{-3}$.

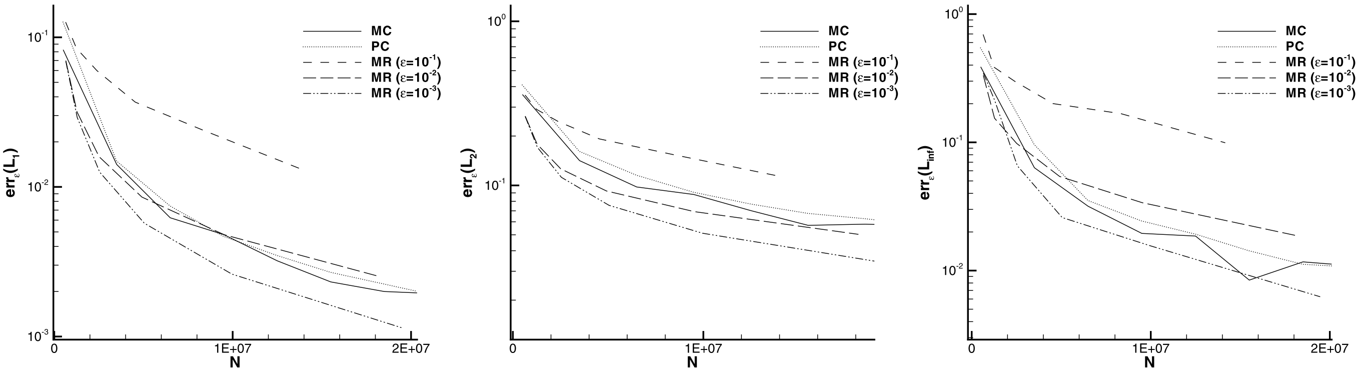
In Fig. 24, the errors for the mean of the variable y_1 are reported measured in the L_1, L_2 and L_∞ spaces.

The error norms of the variances computed in the time interval $[8, 500]$ (in the L_1, L_2 and L_∞ spaces) are reported for the variable y_1 in Fig. 25.



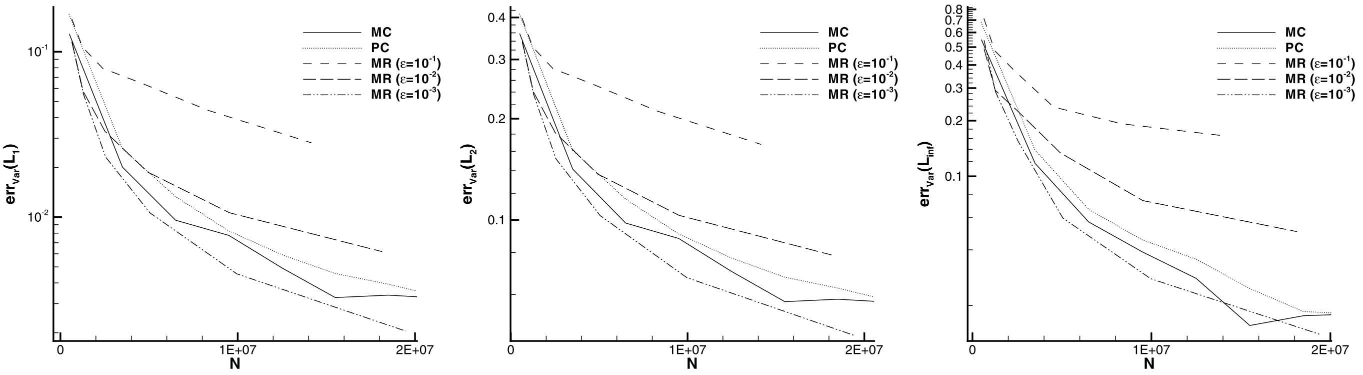
(a) (b) (c)

Fig. 23. Comparison of the time evolution for the variance of the variables y_1 (a), y_2 (b) and y_3 (c) in the Kraichnan–Orszag problem with uniform distribution ($T = 500$). The last time steps are computed with the Monte Carlo, Polynomial Chaos and TE algorithm, $\varepsilon = 10^{-3}$, $m_L = 4$ and $m_{\max} = 12$ ($N_\xi = 3760$), and compared to the Monte Carlo reference solution $N_\xi = 20 \times 10^6$.



(a) (b) (c)

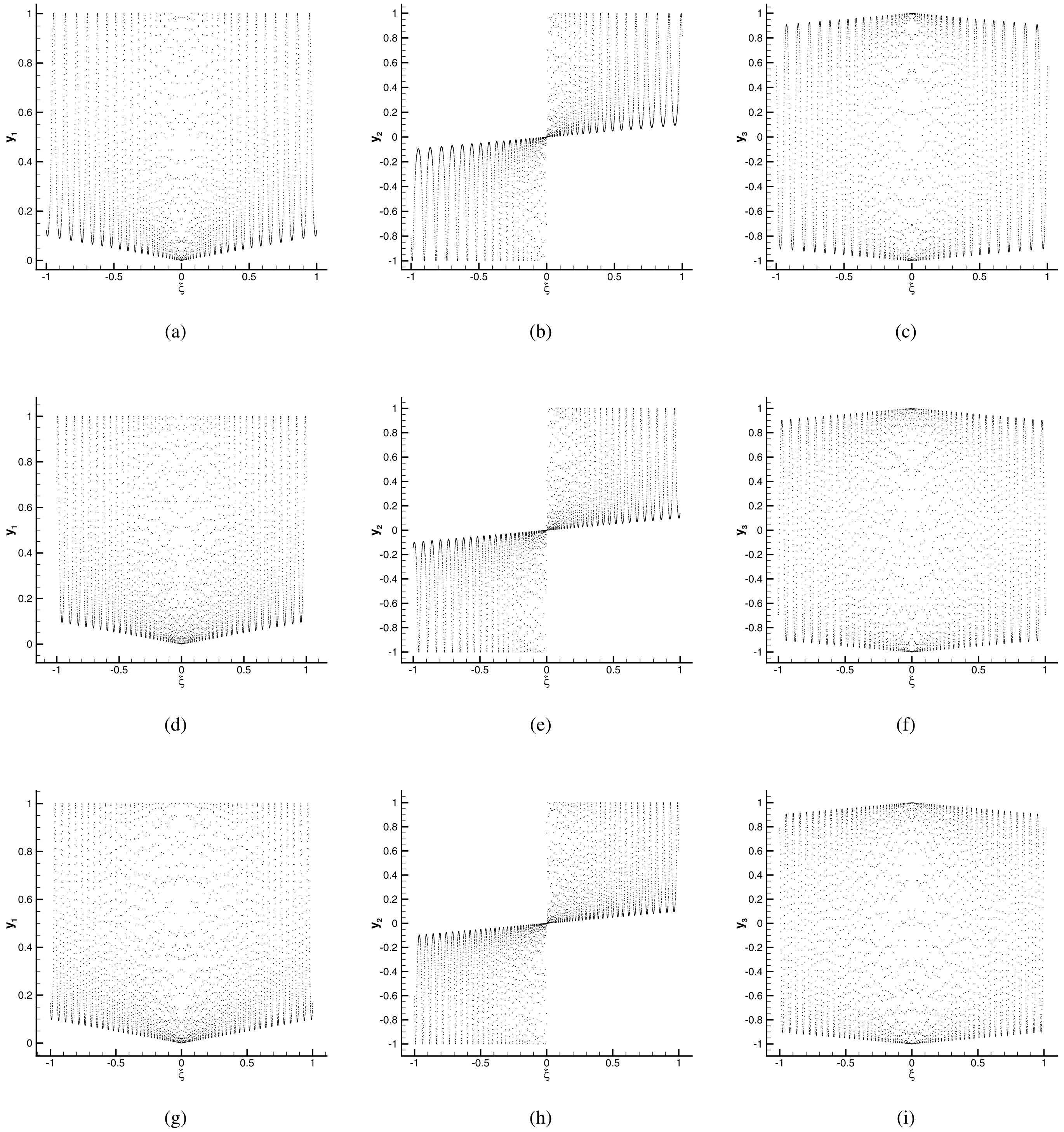
Fig. 24. Error norms for the mean of the variable y_1 of the 1D Kraichnan–Orszag (for long time $T = 500$) problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Fig. 25. Error norms of the variance of the variable y_1 for the 1D Kraichnan–Orszag (for long time $T = 500$) problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

Remark that, in the case of long-time integration, the advantages of the application of the TE algorithm appear very poor with respect to the short-time case reported in Section 6.3 or the unsteady pdf reported in Section 6.5. Nevertheless despite to the very high complexity of the solution (see Fig. 26), all the norms of the errors computed for the three variables are smaller in the case of the application of the TE algorithm.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 26. Time evolution of the variables y_1 (left), y_2 (middle) and y_3 (right) at different time steps: $t = 300$ ((a), (b) and (c)), $t = 400$ ((d), (e) and (f)) and $t = 500$ ((g), (h) and (i)).

In Figs. 24, 25, the TE algorithm is applied with different thresholds in order to display the different convergence properties with a fixed coarsest level correspondent to $m_L = 4$ and a maximal level equal to $m_{\max}$ between 6 and 11. According to the theoretical prediction, a higher threshold correspond to a higher error. However, we remark that, in the long time case reported in this section, a lower threshold is needed in order to obtain better results with respect to a PC method. Instead, for the other numerical experiences reported in Sections 6.3 and 6.4, a higher threshold ($\varepsilon = 10^{-1}$) is sufficient to ensure the best trade-off between a reduced number of points and the lowest error. In general, we can say that long-time highly non-linear problems require a lower threshold due to control the interpolation error cumulated during the time integration.

The evolution of the number of evaluations in time, for the TE algorithm, is reported in Fig. 27, where the refinement/derefinement properties of the algorithm can lead to a reduced number of simulations with respect to the finest

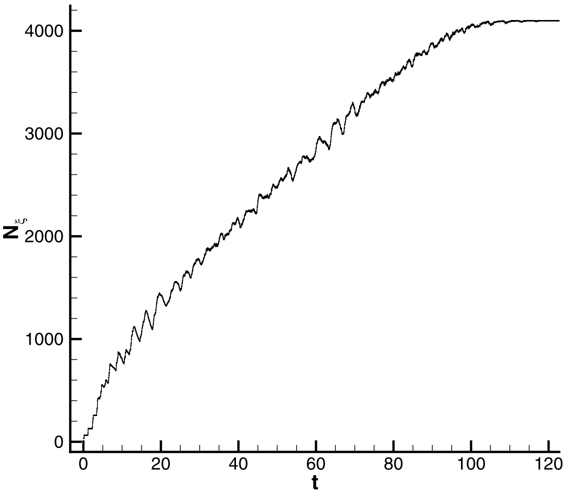


Fig. 27. Number of points in the stochastic space N_ξ employed by the TE algorithm with $m_{\max} = 12$.

Table 3
Time elapsed for the Kraichnan–Orszag problem (with final time $T = 500$) with the performances of the TE algorithm with different thresholds ε.

$m_{\max}$	N_{eval}	t_ε [s]	τ	$\Delta t\%$
$\varepsilon = 10^{-1}$				
6.0000E+00	6.3788E+05	4.0400E+00	1.0190E+00	4.3866E−01
8.0000E+00	2.3891E+06	1.1664E+01	1.0757E+00	5.6885E+00
1.0000E+01	8.3288E+06	3.7696E+01	1.2307E+00	1.9874E+01
1.2000E+01	2.2681E+07	1.0149E+02	1.8064E+00	4.5052E+01
$\varepsilon = 10^{-2}$				
6.0000E+00	6.4681E+05	4.0569E+00	1.0049E+00	2.2180E−02
8.0000E+00	2.5113E+06	1.2081E+01	1.0234E+00	2.3110E+00
1.0000E+01	9.5236E+06	4.4494E+01	1.0763E+00	5.4243E+00
1.2000E+01	3.3905E+07	1.5353E+02	1.2084E+00	2.2288E+01
$\varepsilon = 10^{-3}$				
6.0000E+00	6.4848E+05	4.0533E+00	1.0023E+00	1.1016E−01
8.0000E+00	2.5484E+06	1.2087E+01	1.0085E+00	2.2672E+00
1.0000E+01	9.9572E+06	4.6031E+01	1.0294E+00	2.1558E+00
1.2000E+01	3.7604E+07	1.7134E+02	1.0895E+00	7.2291E+00

resolution level in the initial stage of the computation (for t less than about 120). While, for increasing t, a higher number of simulation is required with no possibility to reduce, even locally, the number of evaluations due to the high non-linearity of the solution (see Fig. 26 in which all the points consist in evaluations). However, the reduction of the number of evaluations in the first part of the computation is sufficient to ensure a good reduction of the overall number of computations.

6.4.1. *Some additional remarks on the computational cost*

In this case, we report explicitly the computational cost associated to TE algorithm in terms of time. In Table 3, the number of points evaluated (N_{eval}), the evaluation compression ratio (see Eq. (33)) and a measure of the saved time $\Delta t\%$ by using the TE algorithm are reported. If we indicate with T_{FULL} the time employed without compression for a certain level $m_{\max}$ and t_ε the time needed by the TE algorithm for a certain threshold ε (obviously with the same maximum level), then $\Delta t\%$ is computed as

$$\Delta t\% = \frac{T_{FULL} - t_\varepsilon}{T_{FULL}} \times 100. \tag{47}$$

All the data are relative to the simulations performed on a personal laptop embedded with `Intel(R) Core(TM)2 Duo CPU P9700 @ 2.80 GHz` and 4 GB of RAM. The evolution of the saved time using the TE algorithm is reported, as function of the maximum level $m_{\max}$ for different threshold in Fig. 28.

6.5. *Kraichnan–Orszag 1D problem with a discontinuous unsteady pdf*

In this section, the 1D Kraichnan–Orszag problem presented in the previous section is solved employing a discontinuous unsteady probability distribution. In particular, the following pdf is retained

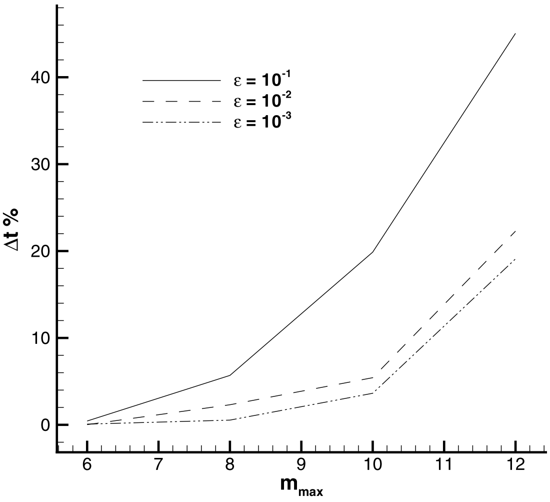


Fig. 28. Time saved using the TE algorithm on the Kraichnan–Orszag problem with different threshold and different maximal levels. The coarsest level is equal to $m_L = 4$ for all the computations.

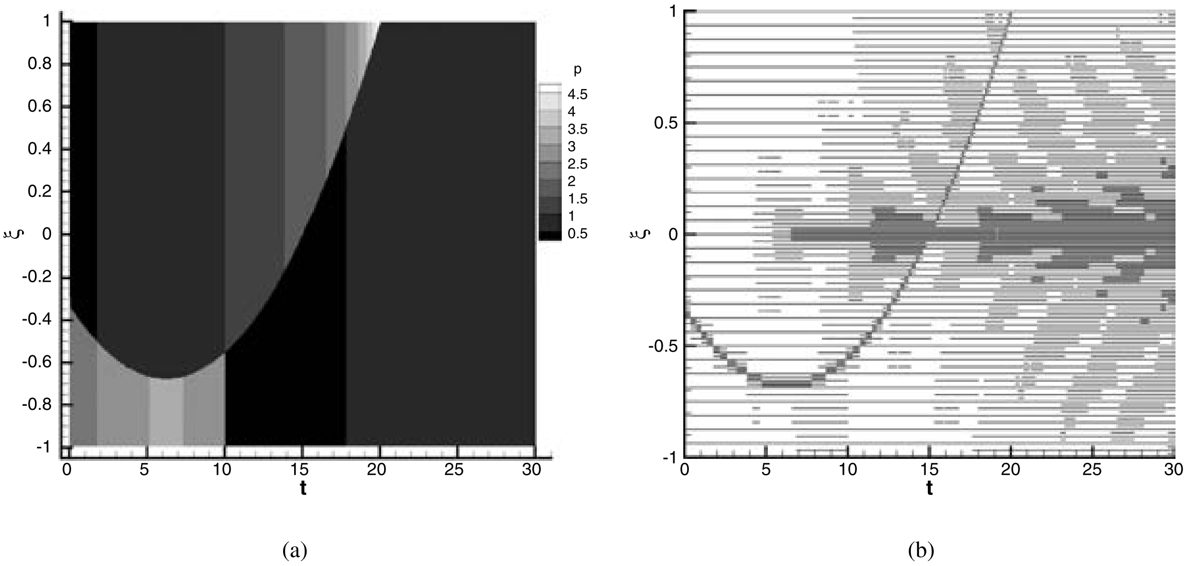


Fig. 29. Probability density function (a) of Eq. (48) and point distribution with TE ($\varepsilon = 10^{-1}$, $m_L = 4$ and $m_{\max} = 8$) (b).

$$p(\omega, t) = \begin{cases} p_I & \omega \leqslant \omega_d, \\ p_{II} & \omega > \omega_d, \end{cases} \tag{48}$$

where

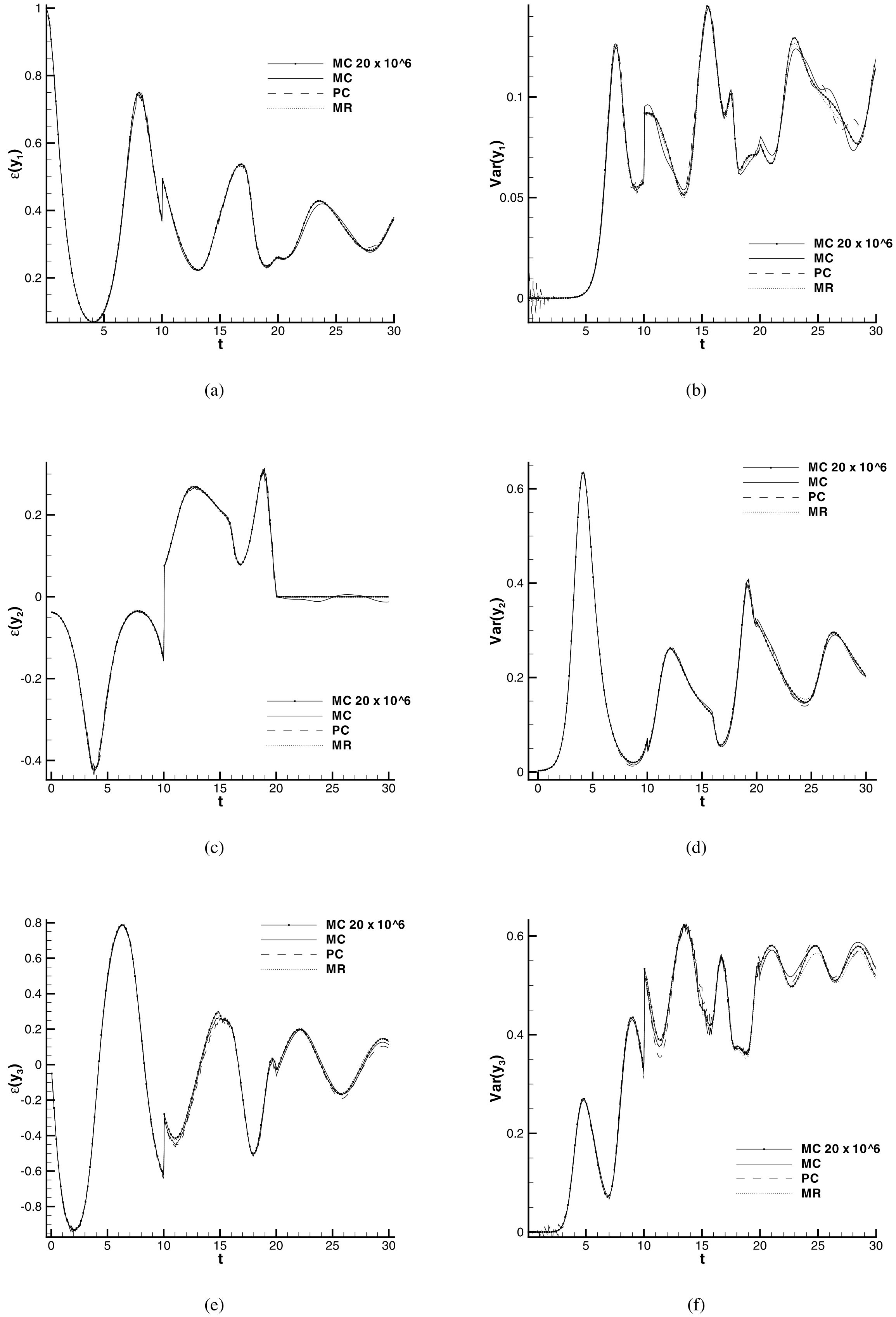
$$p_I = \begin{cases} N p_{II} & 0 \leqslant t \leqslant \bar{t}, \\ p_{II}/N & \bar{t} \leqslant t \leqslant T \end{cases} \tag{49}$$

in which $N = 5$, $\bar{t} = 10$, $T = 30$ and $\omega_d(t) = \frac{11}{2500}t^2 - \frac{11}{200}t + \frac{1}{3}$. Obviously, the normalization condition $\int_{\Xi} p(\xi)\,d\xi = 1$ is satisfied at each time step.

In Fig. 29(a), a contour of the function $p = p(\omega, t)$ is reported. The discontinuity has a parabolic shape and it disappears at nearly $t = 21$. At $t = 10$, the inversion between the left and right part of the discontinuity occurs creating a discontinuity also in time. After the disappearance of the discontinuity, the function reduces to a uniform classical distribution.

The solutions for the statistics of the three variables y_1, y_2 and y_3 are reported in Fig. 30. In Fig. 30, the exact solution obtained with a MC with $N_\xi = 20 \times 10^6$ is reported as reference. The results obtained with the TE algorithm with $\varepsilon = 10^{-1}$, $m_L = 4$ and $m_{\max} = 8$, corresponding to a number of points equal to $N_\xi = 88$, are compared to MC and PC results, computed with the same number of N_ξ.

In this case of unsteady pdf, the point distribution is affected by the presence of a moving discontinuity as well as the high gradients generated by the system responses. This creates a different distribution of points that becomes equal to the previous one (the uniform case) in the last part of the computation. The evolution of points distribution in time is reported in Fig. 29(b).



(a)

(b)

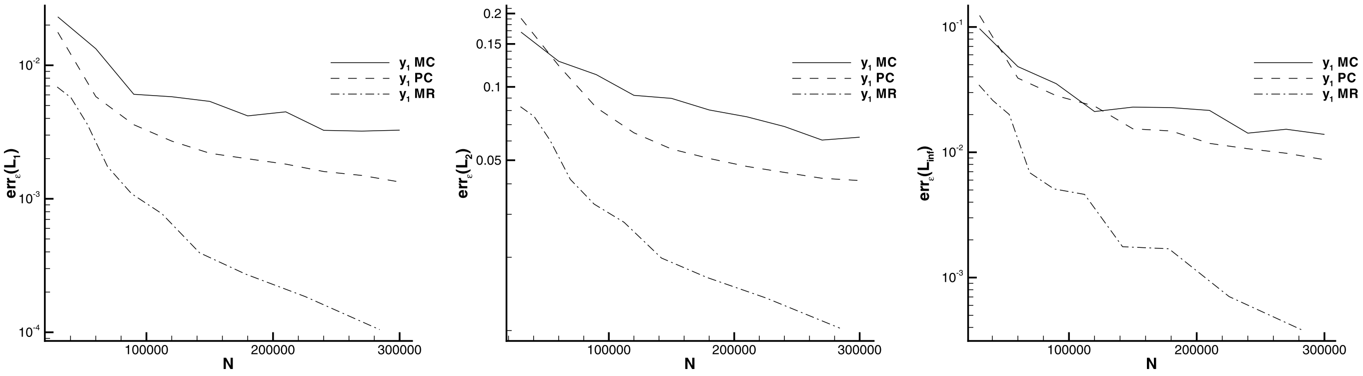
(c)

(d)

(e)

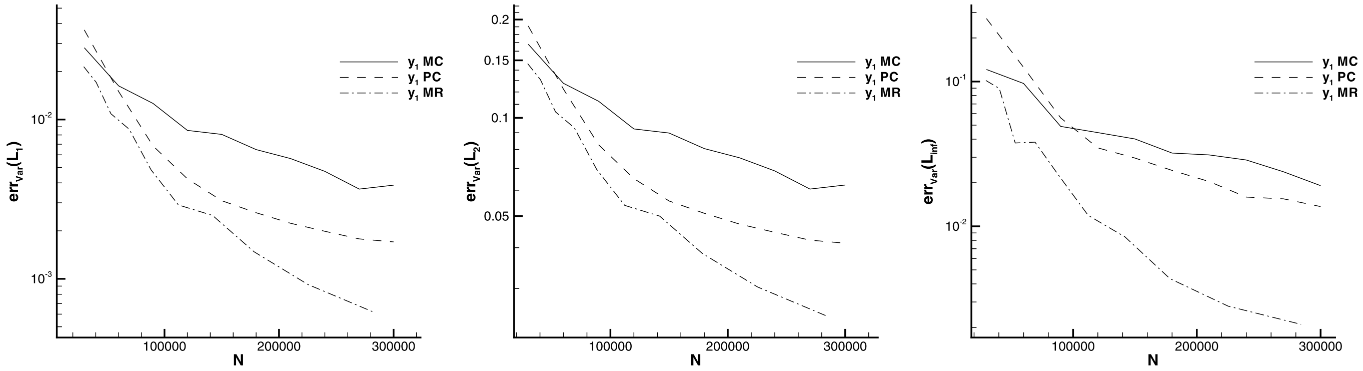
(f)

Fig. 30. Time evolutions of the mean (left) and variance (right) of the three variables y_1 (up), y_2 (middle) and y_3 (bottom) for the 1D Kraichnan–Orszag problem with the unsteady pdf of Eq. (48).



(a) (b) (c)

Fig. 31. Error norms of the mean of the variable y_1 for the 1D Kraichnan–Orszag with unsteady pdf problem in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Fig. 32. Error norms of the variance of the variable y_1 for the 1D Kraichnan–Orszag problem with unsteady pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

Even in the case of discontinuous pdf, the variance for all the variables starts from zero and the mean of the variables y_2 and y_3 could assume some zero-values. For this reason, all the normalized errors are computed for time intervals equal to $[8, 30]$, as done previously for uniform distributions.

In Fig. 31, the errors of the mean of the variable y_1 are reported measured in the L_1, L_2 and L_∞ spaces.

In Fig. 32, the error norms of the variance (computed in the time interval $[8, 30]$, in the L_1, L_2 and L_∞ spaces) are reported for y_1. For all the simulations, the TE algorithm shows better results with respect both MC and PC. In particular a lower error is reached for all the norms for each variable. For all the variable, in the case of L_1 and L_∞ norms the error reached by the TE algorithm is up to an order of magnitude inferior to the correspondent errors, at the same number of points, by the MC and PC methods.

The evolution of points in time is reported in Fig. 33. Remark the increase of the number of points with respect to the points of the uniform distributed case, reported in Fig. 22(b). Remark also the different local behavior in the two cases.

6.6. A two degree-of-freedom subsonic airfoil model for aeroelasticity

In this section, we present the results obtained with the TE algorithm to the motion of a typical wing section with two degree-of-freedom in a subsonic flight condition (see Fig. 34 for a sketch of the problem and the convention adopted).

This simplified model can be employed to perform preliminary aeroelastic computations. The wing section is capable to reproduce the deflection and the torsional motion of a wing by two springs of stiffness k_h and k_t (see Fig. 34). The equations of motion can be derived by an energetic method as the Lagrangian approach. If a lift L and an aerodynamic moment M_a are applied to the wing section in the aerodynamic center AC supposed to be located at the first quarter of the cord (in theory this assumption is valid only for the subsonic motion of a thin plate), the equations of motion, describing the traversal h (in meters) and torsional α (in radiants) motion for a wing in flight at speed U in air with density ρ, are as follows

$$\begin{bmatrix} m & -S \\ -S & I \end{bmatrix} \begin{Bmatrix} \ddot{h} \\ \ddot{\alpha} \end{Bmatrix} + \begin{bmatrix} k_h & 0 \\ 0 & k_t \end{bmatrix} \begin{Bmatrix} h \\ \alpha \end{Bmatrix} = \begin{Bmatrix} L \\ eL + M_a \end{Bmatrix}, \tag{50}$$

in which m is the total mass of the wing section, S is the static inertia moment of the section computed as $S = m(d - e)$, q is the dynamic pressure $q = 1/2\rho U^2$ and the polar inertia moment I is obtained summing the moment referred to the center

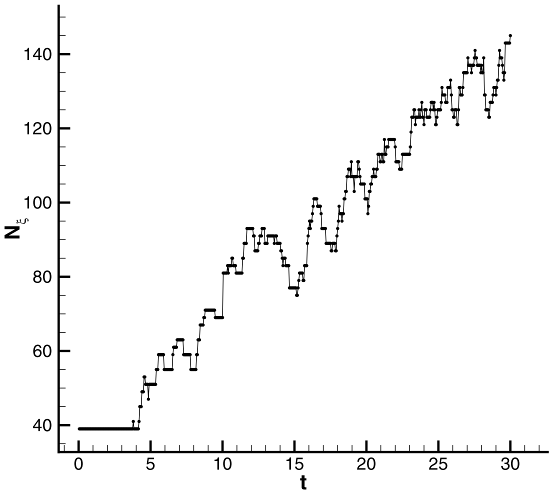


Fig. 33. Evolution of the number of points in time for the TE algorithm with $\varepsilon = 10^{-1}$, $m_L = 4$ and $m_{\max} = 8$, 1D Kraichnan–Orszag problem with discontinuous unsteady pdf.

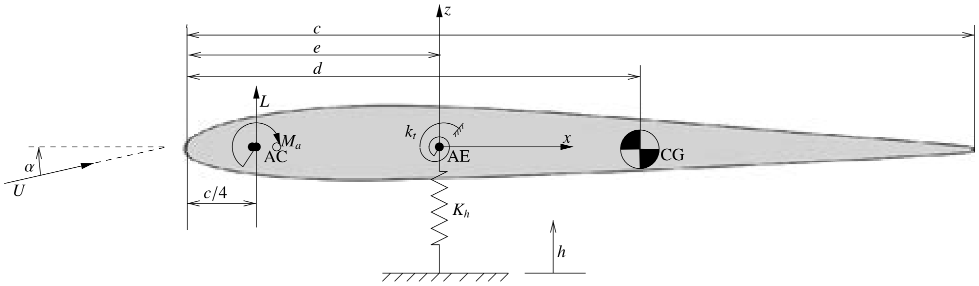


Fig. 34. Sketch of the wing section aeroelastic problem with the forces and the convention adopted.

of gravity I_{CG} and the transport contribution $m(d-e)^2$. The aerodynamics load are computed by means of a simplified quasi-steady strip theory in which the influence of the traversal velocity directly affects the angle-of-attack of the wing section but the memory effect is neglected, i.e. the aerodynamic loads are not dependent from the history of the fluid motion but only from the actual condition. Assuming a lift coefficient C_{L0}, measured with respect to an incidence equal to zero, and a moment coefficient C_M, assumed to be independent from the angle-of-attack, the aerodynamics loads can be written as

$$\begin{aligned} &L = qcC_L = qc(C_{L0} + C_{L/\alpha}\alpha_e), \\ &M_a = qc^2 C_M \end{aligned} \tag{51}$$

in which the slope of the lift coefficient is indicated as $C_{L/\alpha}$ and the effective angle-of-attack α_e is relative to the local stream velocity

$$\alpha_e = \alpha - \frac{\dot{h}}{U}. \tag{52}$$

Obviously, in Eq. (50), the contribution to the lift due to α_e can be recasted to constitute the dumping matrix and can contribute to the stiffness matrix in the following final form

$$\begin{bmatrix} m & -S \\ -S & I \end{bmatrix} \begin{Bmatrix} \ddot{h} \\ \ddot{\alpha} \end{Bmatrix} + \begin{bmatrix} \frac{2\pi qc}{U} & 0 \\ \frac{2\pi qce}{U} & 0 \end{bmatrix} \begin{Bmatrix} \dot{h} \\ \dot{\alpha} \end{Bmatrix} + \begin{bmatrix} k_h & -2\pi qc \\ 0 & k_t - 2\pi qce \end{bmatrix} \begin{Bmatrix} h \\ \alpha \end{Bmatrix} = \begin{Bmatrix} qcC_{L0} \\ qceC_{L0} + qc^2 C_M \end{Bmatrix}. \tag{53}$$

It was assumed that the slope of the C_L–α curve is equal to 2π as demonstrated in the classical thin airfoil theory. This assumption is valid for small angle-of-attack, i.e. roughly speaking between −10° and 10°. This condition is satisfied in the numerical tests reported below.

In Table 4, all the numerical values employed to perform the numerical tests are reported. All the forces are expressed for unit span.

Table 4
Physical characteristics and flight conditions for the wing section.

Geometrical properties	
c [m]	1
e [m]	0.2
d [m]	0.35
Structural properties	
k_h [N/m^2]	1480
k_α [N/rad]	1000
Aerodynamics properties	
C_L [–]	0.1
C_M [–]	−0.001
Flight conditions	
U [m/s]	30
ρ [kg/m^3]	1.2

A parameterization of the mass properties of the section are chosen as follows:

$$\begin{aligned} &m = \xi, \\ &S = m(d-e), \\ &I_{CG} = \xi^2 c^2 + d(d-e)^2 \end{aligned} \tag{54}$$

where the random parameter $\xi = 200\omega + 200$ [kg/m] in which ω is distributed in $[0, 1]$ with the following probability density function

$$p = \begin{cases} 6/5 & \text{for } \omega < \frac{1}{3}, \\ 3/2 & \text{for } \dfrac{1}{3} \geqslant \omega \leqslant \frac{2}{3}, \\ 3/10 & \text{for } \omega > \frac{2}{3}. \end{cases} \tag{55}$$

This kind of parameterization influences the mass matrix and then modifies the oscillation frequency of the system; the parameters (see Table 4) are chosen in order to have a stable system, i.e. the system after the impulsive start oscillates entering in a transitory state and after some cycles reaches another equilibrium state. Of course, for lower parameters of the parameter ξ, the oscillation frequency is greater and the *vice versa* occurs for greater value. In fact, the oscillatory frequency f [Hz] is proportional to root mean square of the ratio between the stiffness of the system and its mass.

Eq. (53) can be written as

$$M\ddot{\mathbf{x}} + C\dot{\mathbf{x}} + K\mathbf{x} = \mathbf{Q}, \tag{56}$$

where M, C and K are the mass, dumping and stiffness matrix, respectively. The unknown vector is $\mathbf{x} = \{h, \alpha\}^{\mathrm{T}}$ while the external forces are collected in the vector $\mathbf{Q} = \{qcC_{L0}, qc(eC_{L0} + cC_M)\}^{\mathrm{T}}$.

An explicit time discretization is applied, with N_t constant time steps $t_n \in [0, T]$ where $t_n = n\Delta t$

$$M\frac{\mathbf{x}_n - 2\mathbf{x}_{n-1} + \mathbf{x}_{n-2}}{\Delta t^2} + C\frac{\mathbf{x}_n - \mathbf{x}_{n-1}}{\Delta t} + K\mathbf{x}_n = \mathbf{Q}_n. \tag{57}$$

This formulation leads to

$$\left(\frac{1}{\Delta t^2}M + \frac{1}{\Delta t}C + K\right)\mathbf{x}_n - \left(\frac{2}{\Delta t^2}M + \frac{1}{\Delta t}C\right)\mathbf{x}_{n-1} + \frac{1}{\Delta t^2}M\mathbf{x}_{n-2} = \mathbf{Q}_n \tag{58}$$

from which the actual value of the unknown vector $\mathbf{x}(t_n) = \mathbf{x}_n$ can be computed explicitly from $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$ as

$$\mathbf{x}_n = \left(\frac{1}{\Delta t^2}M + \frac{1}{\Delta t}C + K\right)^{-1}\left(\left(\frac{2}{\Delta t^2}M + \frac{1}{\Delta t}C\right)\mathbf{x}_{n-1} - \frac{1}{\Delta t^2}M\mathbf{x}_{n-2} + \mathbf{Q}_n\right). \tag{59}$$

Even in this case the TE algorithm is very straightforward to apply. However, attention must be devoted to the interpolation of both the term $\mathbf{x}_{n-1}$ and $\mathbf{x}_{n-2}$ in order to compute the actual value of $\mathbf{x}_n$, as already done in the case of the examples reported above, but only for one value at the previous time step. The two values of $x_{n-1}(\xi_j^k)$ and $x_{n-2}(\xi_j^k)$ are approximated by interpolation (here linear) obtaining

$$\begin{aligned}
&\mathbf{x}_{n-1}\big(\xi_j^k\big) \approx \mathcal{I}\big(\xi_j^k;\mathbf{x}^k(t_{n-1})\big),\\
&\mathbf{x}_{n-2}\big(\xi_j^k\big) \approx \mathcal{I}\big(\xi_j^k;\mathbf{x}^k(t_{n-2})\big).
\end{aligned}\tag{60}$$

As shown in Section 4, the left and right values, with respect to the maximal level reached at each time step, should be used. So, in this case, the left value $\xi_{j_L}^{\bar{k}(t_{n-1})}$ and the right one $\xi_{j_R}^{\bar{k}(t_{n-1})}$ are considered for the time t_{n-1} with respect the point $\xi_j^k(t_n)$ on the maximum level $\bar{k}(t_{n-1)}$ reached at the time t_{n-1}. As a consequence, the same is done at the time t_{n-2}. We remark here that in the general case $\bar{k}(t_{n-1}) \neq \bar{k}(t_{n-2})$. The following steps are performed in order to interpolate the value of $\mathbf{x}(\xi_j^k)$ at the two time levels t_{n-1} and t_{n-2}.

- For both time levels, the left and right values are determined as:

$$\begin{aligned}
&\xi_{j_L}^{\bar{k}(t_{n-1})}(t_{n-1}) = \max\big\{\xi_j^{\bar{k}(t_{n-1})}\big|\xi_j^{\bar{k}(t_{n-1})} < \xi_j^k\big\},\\
&\xi_{j_L}^{\bar{k}(t_{n-2})}(t_{n-2}) = \max\big\{\xi_j^{\bar{k}(t_{n-2})}\big|\xi_j^{\bar{k}(t_{n-2})} < \xi_j^k\big\}.
\end{aligned}\tag{61}$$

- For both the time levels the right values are determined as

$$\begin{aligned}
&\xi_{j_R}^{\bar{k}(t_{n-1})}(t_{n-1}) = \xi_{j_L}^{\bar{k}(t_{n-1})}(t_{n-1}) + h_{\bar{k}(t_{n-1})},\\
&\xi_{j_R}^{\bar{k}(t_{n-2})}(t_{n-2}) = \xi_{j_L}^{\bar{k}(t_{n-2})}(t_{n-2}) + h_{\bar{k}(t_{n-2})}.
\end{aligned}\tag{62}$$

- The values of $\mathbf{x}(\xi_j^k)(t_{n-1})$ and $\mathbf{x}(\xi_j^k)(t_{n-2})$ are interpolated

$$\begin{aligned}
\mathbf{x}\big(\xi_j^k, t_{n-1}\big) \approx \mathcal{I}\big(\xi_j^k;\mathbf{x}^k(t_{n-1})\big) &= \frac{\mathbf{x}(\xi_{j_R}^{\bar{k}(t_{n-1})}, t_{n-1}) - \mathbf{x}(\xi_{j_L}^{\bar{k}(t_{n-1})}, t_{n-1})}{h_{\bar{k}(t_{n-1})}}\xi_j^k + \mathbf{x}\big(\xi_{j_L}^{\bar{k}(t_{n-1})}, t_{n-1}\big)\\
&\quad - j_L(t_{n-1})\big(\mathbf{x}\big(\xi_{j_R}^{\bar{k}(t_{n-1})}, t_{n-1}\big) - \mathbf{x}\big(\xi_{j_L}^{\bar{k}(t_{n-1})}, t_{n-1}\big)\big),\\
\mathbf{x}\big(\xi_j^k, t_{n-2}\big) \approx \mathcal{I}\big(\xi_j^k;\mathbf{x}^k(t_{n-2})\big) &= \frac{\mathbf{x}(\xi_{j_R}^{\bar{k}(t_{n-2})}, t_{n-2}) - \mathbf{x}(\xi_{j_L}^{\bar{k}(t_{n-2})}, t_{n-2})}{h_{\bar{k}(t_{n-2})}}\xi_j^k + \mathbf{x}\big(\xi_{j_L}^{\bar{k}(t_{n-2})}, t_{n-2}\big)\\
&\quad - j_L(t_{n-2})\big(\mathbf{x}\big(\xi_{j_R}^{\bar{k}(t_{n-2})}, t_{n-2}\big) - \mathbf{x}\big(\xi_{j_L}^{\bar{k}(t_{n-2})}, t_{n-2}\big)\big),
\end{aligned}\tag{63}$$

where the indexes $j_L(t_{n-1}) = \xi_{j_L}^{\bar{k}(t_{n-1})}/h_{\bar{k}(t_{n-1})}$ and $j_L(t_{n-2}) = \xi_{j_L}^{\bar{k}(t_{n-2})}/h_{\bar{k}(t_{n-2})}$. The expression obtained in Eq. (63) inserted in Eq. (59) allows recovering the final form of the numerical scheme.

In this numerical test, the section is assumed to be at rest, i.e. zero deflection and torsion with no transversal and angular velocity, and subject to impulsive start, i.e. $\mathbf{x}_{n-2}(\xi) = \mathbf{x}_{n-1}(\xi) = 0$ for all $\xi \in \Xi$. Of course, this condition is equivalent to an impulsive start of the system that has no physical counterpart, but it represents, mathematically, a well-posed problem. The numerical tests are performed with a final time $T = 18$ s and a time step equal to $\Delta t = 0.001$ chosen after a convergence study not reported here for brevity.

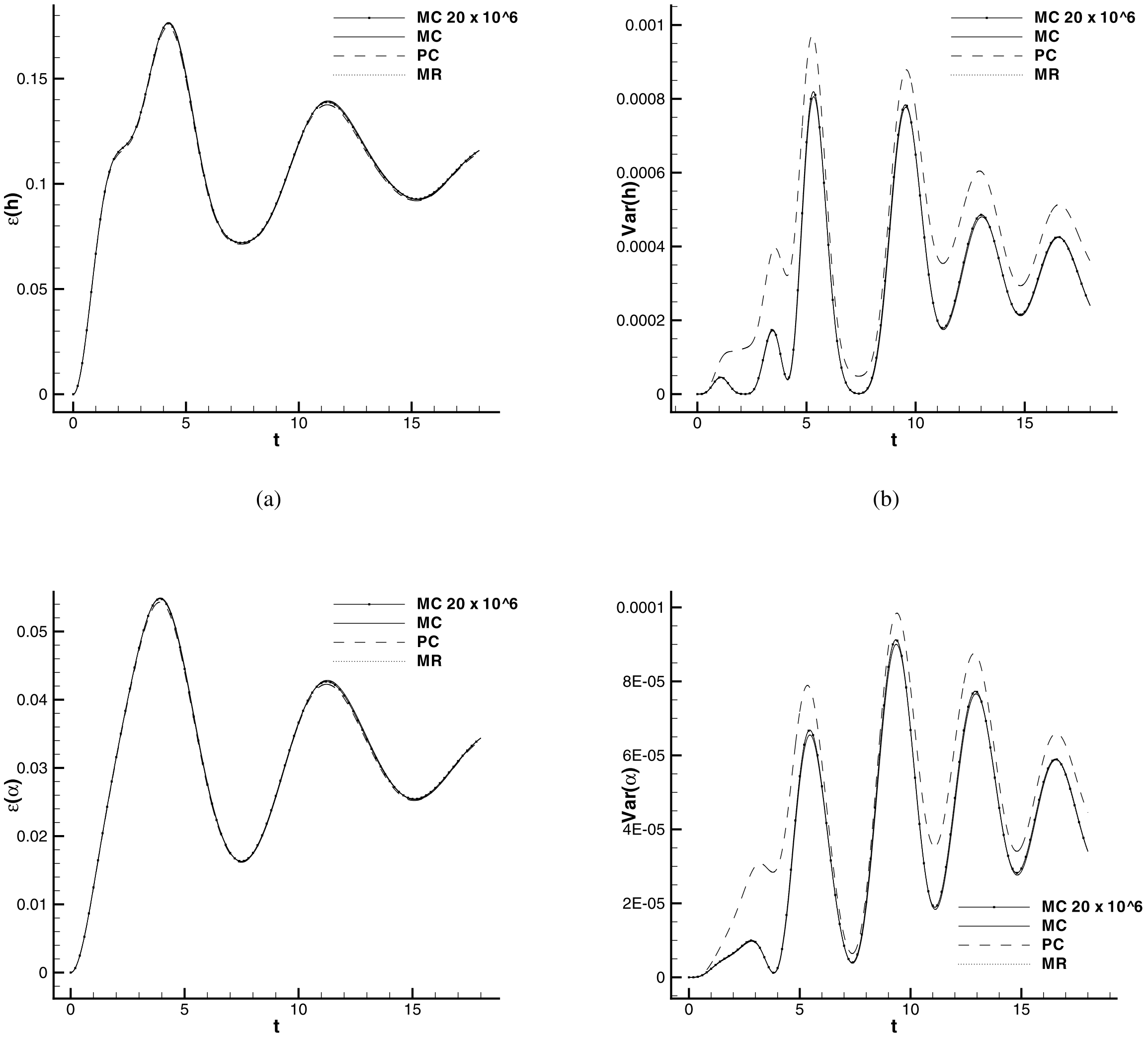
The time evolution of the variables h and α is reported in Fig. 35. In Fig. 35, the solution for MC, PC and the TE algorithm are reported with a number of point N_ξ, in the stochastic space, equal to 86 (equivalent to a TE algorithm of maximal level $m_{\max} = 16$ with the coarsest level $m_L = 3$ and a threshold equal to $\varepsilon = 10^{-1}$). The reference solution obtained with a full converged Monte Carlo computation of $N_\xi = 20 \times 10^6$ points in the stochastic space is also reported.

The mean value of h and α computed by means of the three methods nearly coincide, but, for the variance, a stronger difference appears concerning the PC computations. Even if the shape of the evolution is well solved, the actual values are larger than the exact ones.

As already reported for the other numerical tests, the error of the mean and variance, measured in the L_1, L_2 and L_∞ (see Eqs. (44) and (45)), are computed for the three methods. In this case, the norms are computed for the time interval $[5, 18]$, in order to avoid the normalization with values too close to zero, with an $N_t = 18\,000$.

In Figs. 36 and 37, the errors for the mean and variance in norms L_1, L_2 and L_∞ are reported for the variable h and α, respectively. In the figures, the total number of points in the grid t–ω is reported $N = N_\xi \times N_t$. The value ranges of N_ξ ranges from 50 to 190 with a step of 20 for the MC, from degree 49 to 169 with step 20 for the PC. In order to compute different solutions employing the TE algorithm, the coarsest level is fixed as $m_L = 3$ and the threshold equal to $\varepsilon = 10^{-1}$, while the maximum allowed level, i.e. the finest level, is varied between 12 and 20.

All the numerical results display a superiority of the TE approach with respect to both the MC and PC in terms of level of error and convergence. In particular, all the error curves exhibit a smoother convergence with an error level lower than about an order of magnitude with respect to MC. Provided that the error on the mean remains of the same order of



(a)

(b)

(c)

(d)

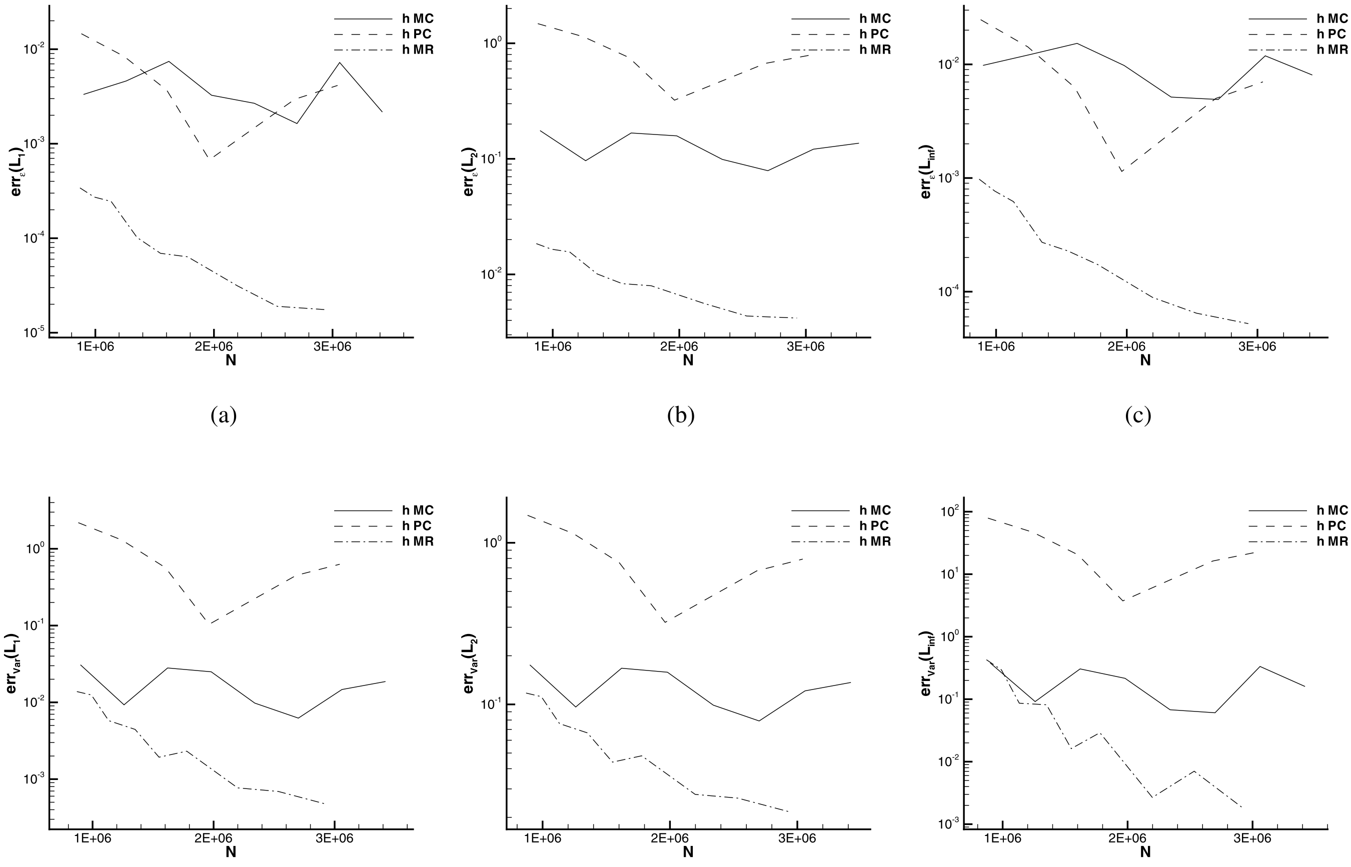
Fig. 35. Time evolution of the mean (left) and variance (right) of the three variables h (up), α (bottom) for the aeroelastic motion described by Eq. (53).

magnitude than MC, the PC shows larger errors for all the norms of the variance for both the variables, as already observed in Fig. 35. Finally, in the range of points considered, only the TE algorithm displays a good convergence behavior while both MC and PC exhibit a very poor rate of convergence and a too oscillatory pattern of the error.

Also in this case, the application of the TE algorithm exhibits good refinement/derefinement properties. For instance, the pattern of the points (Fig. 38(a)) and their number in time (Fig. 38(b)) are reported. Despite to a maximum number of points allowed equal to $N_\xi = 2^{16} + 1 = 65\,536$, the number of points remains always lower than 100 obtaining good compressing results (in term of evaluation compression).

7. Conclusions

In this work an innovative adaptive strategy for stochastic problem, the TE algorithm, inspired to the classical Harten's framework, is presented. A representation of the solution on a finest grid is computed starting from a coarsest grid, with a low number of evaluation of the function. Then, only a reduced set of point values, on the finest grid, is evaluated, while the remaining set is obtained by interpolation (from the previous levels). This procedure moves recursively, with a combination of interpolation and evaluation, from the coarsest level to the finest and from each time step to the successive one. At each time step, the scheme allows to recover the solution on the finest level with a one-time scheme that embeds the *encoding* and the *truncation* procedures of the classical Harten framework. First, this basic algorithm is extended in order to solve a vectorial problem, i.e. computing the stochastic response of a system that has many outputs. Then, slight modifications are suggested for unsteady problems. The TE algorithm must be, in the unsteady case, applied at each time step starting from the coarsest level to the finest one. According to the time-integration technique, the choice of a stencil in time is needed in order to evaluate a solution in the point of the time-stochastic space. The proposed time-advancement strategy can be easily used to build numerical scheme with automatic refinement/derefinement. The proposed formulations



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 36. Error norms of the mean (top) and variance (bottom) of the variable h of the aeroelastic problem in the L_1 ((a) and (d)), L_2 ((b) and (e)) and L_∞ ((c) and (f)) spaces.

permits to recover the same results concerning the interpolation theory of the classical multiresolution approach, but with an extension to uncertainty quantification problems.

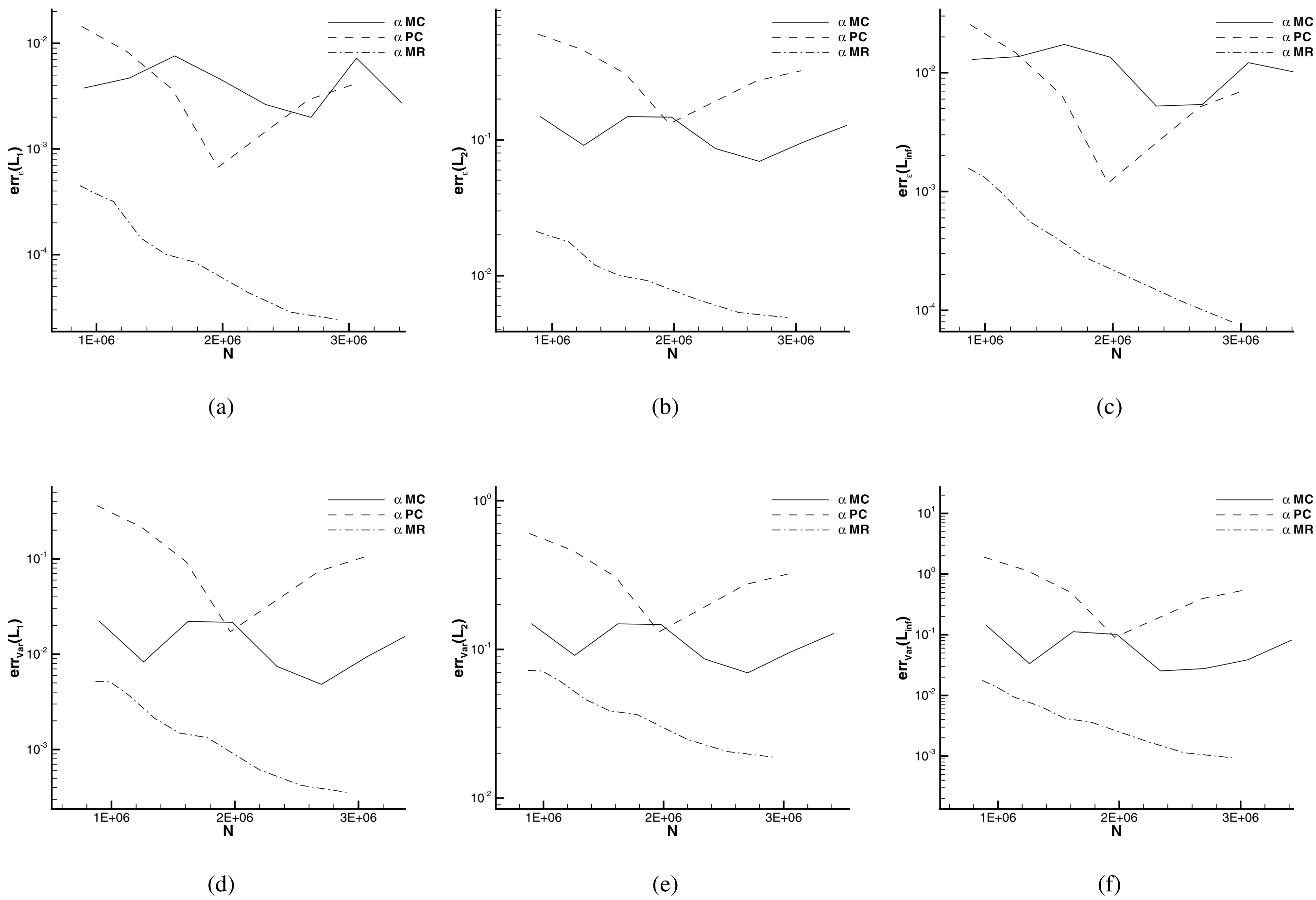
The interest of the present strategy is shown by performing several numerical problems where different forms of uncertainty distributions are taken into account, such as discontinuous and unsteady custom-defined probability density functions.

The TE algorithm is applied to several numerical test-cases in order to check its accuracy and convergence rate with respect to some classical stochastic methods, such as quasi-Monte Carlo and Polynomial Chaos. First, some steady algebraic problems are considered where analytical discontinuous functions are evaluated in terms of their expectancy and variance; for this case the performances in terms of compression and evaluation ratios are also evaluated. The TE algorithm displays a monotone decrease of the errors with respect to the number of evaluations. The capability of the TE algorithm to preserve accuracy for time-evolving solutions is displayed in two test-cases. The first-one deals with an ordinary differential equation (ODE), i.e. taken from [6] but with some modifications in order to obtain a more stiff ξ–t pattern of the function.

In this case, the TE algorithm allows reaching a lower level of error at the same number of points. In the case of L_∞ norm for the variance, at the same number of points it corresponds an order of magnitude of the error inferior up to three orders of magnitude. The convergences exhibit the smoothest behaviors with respect to MC and PC despite the presence of two discontinuities in the inputs. The observed behavior displays the refinement/derefinement capability of the algorithm to move and collocate points in the stochastic space with respect to the time evolution.

The second test-case for checking the convergence properties in long-time integration problem is the so-called Kraichnan–Orszag 1D ODE, a well-known problem in literature for testing UQ methods properties. Different kinds of pdf, i.e. uniform and discontinuous, are considered for both test-cases. For uniform distribution, TE displays better results in terms of accuracy and convergence of the solution with respect to MC and PC. In the case of long-time integration, the advantages of the application of the TE algorithm appear very poor with respect to the short-time case. Anyway, despite to the very high complexity of the solution, all the norms of the errors computed for each variable are smaller in the case of the application of the TE algorithm.

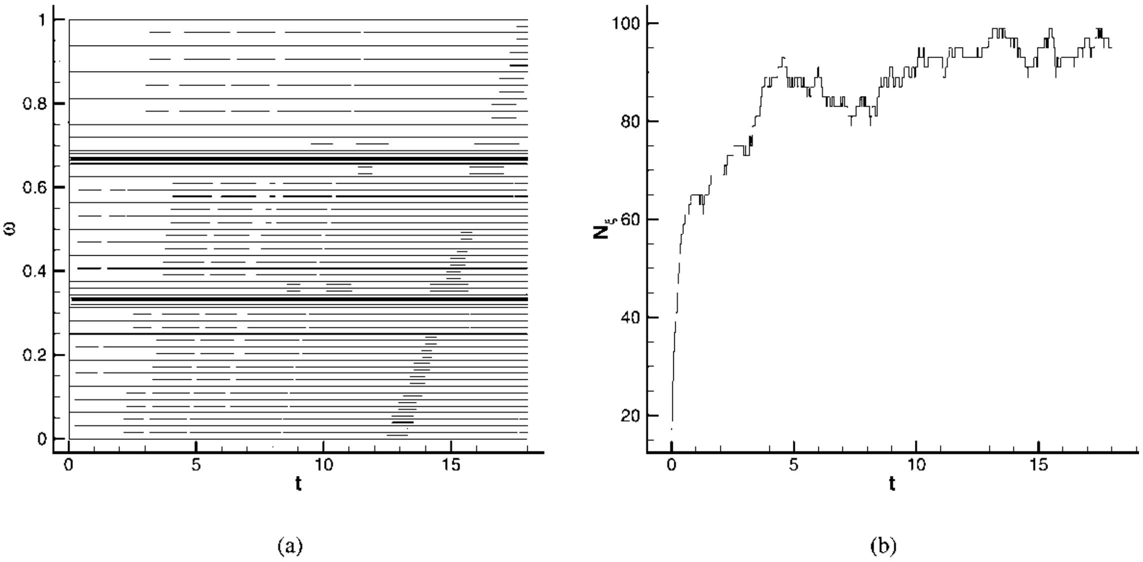
In the case of unsteady discontinuous pdf, the point distribution is affected by the presence of a moving discontinuity as well as the high gradients generated by the system responses. This creates a different distribution of points that reduces the global computational cost when using TE algorithm. As a consequence, performances of TE are very much better with respect to MC and PC solutions in terms of convergence.



(a) (b) (c)

(d) (e) (f)

Fig. 37. Error norms of the mean (top) and variance (bottom) of the variable α of the aeroelastic problem in the L_1 ((a) and (d)), L_2 ((b) and (e)) and L_∞ ((c) and (f)) spaces.



(a) (b)

Fig. 38. Patterns of the evaluations in the space t–ω (a) and the number of points in the stochastic space N_ξ employed by the TE algorithm with $m_{\max} = 16$ (b).

Finally, a simplified model for aeroelastic study, a two degree-of-freedom typical wing section coupled with a quasi-steady strip theory model for aerodynamics, is used to compute the statistics of the motion considering uncertainties on mass properties for discontinuous probability distribution.

All the numerical results display a superiority of the TE approach with respect to both the MC and PC in terms of level of error and convergence. In particular, all the error curves exhibit a smoother convergence with an error level lower than about an order of magnitude with respect to MC. Provided that the error on the mean remains of the same order of magnitude than MC, the PC shows larger errors for all the norms of the variance for both the variables. Finally, in the range of points considered, only the TE algorithm displays a good convergence while both MC and PC exhibit a very poor rate of convergence and a too oscillatory pattern of the error.

Acknowledgements

Remi Abgrall has been partially supported by the ERC Advanced Grant ADDECCO No. 226316, while Gianluca Geraci has been fully supported by the ERC Advanced Grant ADDECCO No. 226316.

The authors are very grateful to anonymous referees for the time spent reading and analyzing the manuscript. Many insightful remarks helped the authors to improve the quality and readability of the paper.

References

[1] I. Graham, F. Kuo, D. Nuyens, R. Scheichl, I. Sloan, Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications, J. Comput. Phys. 230 (2011) 3668–3694.
[2] N. Wiener, The homogeneous chaos, Am. J. Math. 60 (1938) 897–936.
[3] D. Xiu, G.E. Karniadakis, Modeling uncertainty in flow simulations via generalized polynomial chaos, J. Comput. Phys. 187 (2003) 137–167.
[4] X. Wan, G.E. Karniadakis, Beyond Wiener–Askey expansions: handling arbitrary PDFs, J. Sci. Comput. 27 (2005) 455–464.
[5] C. Soize, R.G. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM J. Sci. Comput. 26 (2004) 395–410.
[6] O. Le Maître, O. Knio, Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics, Springer-Verlag, 2010.
[7] R.G. Ghanem, P.D. Spanos, Stochastic Finite Elements. A Spectral Approach, Springer-Verlag, 1991.
[8] X. Wan, G.E. Karniadakis, Long-term behavior of polynomial chaos in stochastic flow simulations, Comput. Methods Appl. Mech. Eng. 195 (2006) 5582–5596.
[9] M. Gerritsma, J.-B. van der Steen, P. Vos, G.E. Karniadakis, Time-dependent generalized polynomial chaos, J. Comput. Phys. 229 (2010) 8333–8363.
[10] I. Babuška, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Rev. 52 (2010) 317.
[11] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Sov. Math. Dokl. (1963) 240–243.
[12] R.E. Bellman, B. Richard, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.
[13] B. Ganapathysubramanian, N. Zabaras, Sparse grid collocation schemes for stochastic natural convection problems, J. Comput. Phys. 225 (2007) 652–685.
[14] M. Griebel, Sparse grids and related approximation schemes for higher dimensional problems, in: L.P. Todd, A. Pinkus, E. Suli, M.J. (Eds.), Foundations of Computational Mathematics (FoCM05), Santander, Cambridge University Press, 2006, pp. 106–161.
[15] J. Foo, G.E. Karniadakis, Multi-element probabilistic collocation method in high dimensions, J. Comput. Phys. 229 (2010) 1536–1557.
[16] O. Le Maître, Uncertainty propagation using Wiener–Haar expansions, J. Comput. Phys. 197 (2004) 28–57.
[17] O. Le Maître, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, J. Comput. Phys. 197 (2004) 502–531.
[18] J.a.S. Witteveen, G. Iaccarino, Simplex elements stochastic collocation in higher-dimensional probability spaces, in: 51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 12–15 April, Orlando, Florida, 2010.
[19] N. Agarwal, N. Aluru, A domain adaptive stochastic collocation approach for analysis of MEMS under uncertainties, J. Comput. Phys. 228 (2009) 7662–7688.
[20] X. Ma, N. Zabaras, An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, J. Comput. Phys. 228 (2009) 3084–3113.
[21] R. Abgrall, P.M. Congedo, A semi-intrusive deterministic approach to uncertainty quantification in non-linear fluid flow problems, J. Comput. Phys. 235 (2013) 828–845.
[22] R. Abgrall, P.M. Congedo, C. Corre, S. Galéra, A simple semi-intrusive method for uncertainty quantification of shocked flows, comparison with a non-intrusive polynomial chaos method, in: J.C.F. Pereira, A. Sequeira (Eds.), V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010, 14–17 June, Lisbon, Portugal, 2010, pp. 1–17.
[23] R. Abgrall, P.M. Congedo, S. Galéra, G. Geraci, Semi-intrusive and non-intrusive stochastic methods for aerospace applications, in: 4th European Conference for Aerospace Sciences, vol. 1, Saint Petersburg, Russia, July 4th–8th, 2011, pp. 1–8.
[24] A. Harten, Multiresolution representation of data: a general framework, SIAM J. Numer. Anal. 33 (1996) 1205–1256.
[25] A. Harten, Multiresolution algorithms for the numerical solution of hyperbolic conservation laws, Commun. Pure Appl. Math. 48 (1995) 1305–1342.
[26] R. Abgrall, A. Harten, Multiresolution representation in unstructured meshes, SIAM J. Numer. Anal. 35 (1998) 2128–2146.
[27] B.L. Bihari, A. Harten, Multiresolution schemes for the numerical solution of 2-D conservation laws I, SIAM J. Sci. Comput. 18 (1997) 315.
[28] S. Amat, F. Aràndiga, A. Cohen, R. Donat, Tensor product multiresolution analysis with error control for compact image representation, Signal Process. 82 (2002) 587–608.
[29] A. Harten, Adaptive multiresolution schemes for shock computations, J. Comput. Phys. 135 (1994) 260–278.
[30] A. Quarteroni, R. Sacco, F. Saleri, Numerical Mathematics, Springer-Verlag, 2000.
[31] S. Müller, Adaptive multiscale schemes for conservation laws, Lect. Notes Comput. Sci. Eng. 27 (2002).
[32] A. Cohen, S.M. Kaber, S. Müller, M. Postel, Fully adaptive multiresolution finite volume schemes for conservation laws, Math. Comput. 72 (2003) 2003.
[33] S. Orszag, Dynamical properties of truncated Wiener–Hermite expansions, Phys. Fluids 10 (1967) 2603–2613.

Paper *P2*

Elsevier Editorial System(tm) for Mathematics and Computers in Simulation
Manuscript Draft

Manuscript Number:

Title: Toward a unified multiresolution scheme in the combined physical/stochastic space for stochastic differential equations

Article Type: Special Issue: MASCOT 2011

Keywords: Multiresolution; Ordinary Differential Equations; Partial Differential Equation; Uncertainty Quantification; Heat Equation

Corresponding Author: Mr Gianluca Geraci,

Corresponding Author's Institution: INRIA Bordeaux--Sud Ouest

First Author: Rémi Abgrall

Order of Authors: Rémi Abgrall; Pietro Marco Congedo; Gianluca Geraci

Abstract: In the present work, an innovative method for solving stochastic partial differential equations is presented. A multiresolution method permitting to compute statistics of the quantity of interest for a whatever form of the probability density function is extended to permit an adaptation in both physical and stochastic spaces. The efficiency of this strategy, in terms of refinement/derefinement capabilities, is displayed for stochastic algebraic and differential equations with respect to other more classical techniques, like Monte Carlo (MC) and Polynomial Chaos (PC). Finally, the proposed strategy is applied to the heat equation, displaying very promising results in terms of accuracy, convergence and regularity.

Toward a unified multiresolution scheme in the combined physical/stochastic space for stochastic differential equations

R. Abgrall, P.M. Congedo, G. Geraci*

INRIA Bordeaux–Sud-Ouest, Equipes Bacchus, 351, Cours de la Liberation, 33405 Talence, FRANCE

Abstract

In the present work, an innovative method for solving stochastic partial differential equations is presented. A multiresolution method permitting to compute statistics of the quantity of interest for a whatever form of the probability density function is extended to permit an adaptation in both physical and stochastic spaces. The efficiency of this strategy, in terms of refinement/derefinement capabilities, is displayed for stochastic algebraic and differential equations with respect to other more classical techniques, like Monte Carlo (MC) and Polynomial Chaos (PC). Finally, the proposed strategy is applied to the heat equation, displaying very promising results in terms of accuracy, convergence and regularity.

Keywords: Multiresolution, Ordinary Differential Equations, Partial Differential Equation, Uncertainty Quantification, Heat Equation

1. Introduction

In the last fifty years, a strong effort has been devoted to develop efficient numerical methods for solving partial differential equations. Estimating the predictivity of a numerical simulation remains very challenging. One of the most important issues is that the physical model and/or the initial/boundary conditions are strongly affected by uncertainties. A general agreement is reached on the necessity to take into account experimental and modeling uncertainties in the numerical simulation. The so-called Uncertainty Quantification (UQ) is a branch of the numerical analysis that has been developed more recently to quantify the uncertainty and to estimate the confidence interval of a certain quantity of interest.

The first and most known UQ method is the Monte Carlo method. The Polynomial Chaos (PC) techniques has acquired great popularity in last years. In the original work of Wiener [18], the solution is expanded in a polynomial Hermite basis, the so-called homogeneous chaos expansion, while in recent years, Xiu and Karniadakis [19] demonstrated that the optimal convergence can be achieved if orthogonal basis are chosen following the so-called Wiener-Askey scheme. This leads to the well-known generalized Polynomial Chaos (gPC) approach. However, problems with discontinuities in the random space can lead to slow convergence. Similarly, long-time integration problems could be encountered [17], where this behavior is due to the modification in time of the statistic properties of the solution inducing an efficiency loss of the polynomial basis in time. Recently, Gerritsma [7] proposed a time-dependent generalized Polynomial Chaos scheme based on the research of a time varying optimal polynomial basis. The majors drawbacks related to the application of the PC to real-like cases is related to the presence of discontinuities, in both physical and stochastic spaces, to long-time integration problems and to the use of a custom-defined form of probability density function (for example discontinuous and unsteady). Actually, handling a discontinuity in both physical and stochastic spaces remains a very challenging issue. In the context of gPC schemes, Wan and Karniadakis introduced an adaptive class of methods for solving discontinuities by using local basis functions, the multi-element generalized Polynomial Chaos (ME-gPC) [6]. This strategy deals with an adaptive decomposition of the domain on which local

*Corresponding author
Email address: gianluca.geraci@inria.fr (G. Geraci)

basis are employed. In order to treat discontinuous response surfaces, Le Maître et al. applied a multiresolution analysis to Galerkin projection schemes [12, 11, 16]. This class of schemes relies on the projection of the uncertain data on a multi-wavelets basis consisting of piecewise polynomial (smooth) functions. This approach is shown to be very CPU demanding. Consequently, two cures are then explored in the context of adaptive methods: automatically refine the multi-wavelets basis or adaptively partitioning the domain.

More recently, Abgrall et al. [1, 2, 3] introduces a new class of finite volume schemes capable to deal with discontinuous problems for shock-dominated flows. The so called semi-intrusive scheme (SI) exhibits promising results in term of accuracy and efficiency compared to more classical Monte Carlo and PC methods. A step-forward for reducing the computational cost and preserving accuracy is made by the authors with a new technique inspired to the Multiresolution framework of Harten [8, 9, 10]. Preliminary results in this direction [4], for problems with custom-defined probability density functions, displays promising results with respect to classical techniques like MC and PC.

In this work, this method is extended to solve not only ordinary differential equations, as made in [4] using the Truncate and Encode (TE) technique, but also partial differential equations. A new stochastic technique, called spatial-TE (sTE), is presented with refinement/derefinement capabilities in time for both the physical and stochastic spaces. The main advantage is the overall reduction of the total number of points needed to reach a certain level of accuracy for the complete stochastic solution.

The approach proposed in the present work is based on a multiresolution concept, as already made in Le Maître et al. [16]. Anyway, the approach differs completely since here no spectral projection is employed, as it will be explained in the next section. Moreover, the possibility to reject a wavelets (equal to an interpolation error as in the original Harten framework) is based only on local tests, then is different from Galerkin projection approach where 1D energy estimators along stochastic dimensions are used. For details on the multiresolution approach applied to Galerkin projection schemes, the reader can refer to the extremely exhaustive reference [13].

This paper is organized as follows. In Section 2, the mathematical problem is defined. The new strategy, *i.e.* the sTE, is illustrated in Section 3. Then, the application to the stochastic heat equation is presented in Section 4. Section 5 presents several numerical results for different test-cases. Finally, some conclusions and perspectives are drawn in Section 6.

2. Mathematical setting

Consider the following problem for an output of interest $u(\mathbf{x}, t, \boldsymbol{\xi}(\omega))$[1]:

$$\mathcal{L}(\mathbf{x}, t, \boldsymbol{\xi}(\omega); u(\mathbf{x}, t, \boldsymbol{\xi}(\omega))) = \mathcal{S}(\mathbf{x}, t, \boldsymbol{\xi}(\omega)), \tag{1}$$

where the operator $\mathcal{L}$ can be either an algebraic or a differential operator (in this case we need appropriate initial and boundary conditions). The operator $\mathcal{L}$ and the source term $\mathcal{S}$ are defined on the domain $D \times T \times \Xi$, where $\mathbf{x} \in D \subset \mathbb{R}^{n_d}$, with $n_d \in \{1, 2, 3\}$, and $t \in T$ are the spatial and temporal dimensions. Randomness is introduced in (1) and its initial and boundary conditions in term of d second order random parameters $\boldsymbol{\xi}(\omega) = \{\xi_1(\omega_1), \ldots, \xi_d(\omega_d)\} \in \Xi$ with parameter space $\Xi \subset \mathbb{R}^d$. The symbol $\omega = \{\omega_1, \ldots, \omega_d\} \in \Omega \subset \mathbb{R}$ denotes realizations in a complete probability space $(\Omega, \mathcal{F}, P)$. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events and $P : \mathcal{F} \to [0, 1]$ is a probability measure. In our case the random variables ω are by definition standard uniformly $\mathcal{U}(0, 1)$ distributed. Random parameters $\boldsymbol{\xi}(\omega)$ can have any arbitrary probability density function $p(\boldsymbol{\xi}(\omega))$, in this way $p(\boldsymbol{\xi}(\omega)) > 0$ for all $\boldsymbol{\xi}(\omega) \in \Xi$ and $p(\boldsymbol{\xi}(\omega)) = 0$ for all $\boldsymbol{\xi}(\omega) \notin \Xi$; we can now drop the argument ω for brevity. The probability density function $p(\boldsymbol{\xi}(\omega))$ is defined as a joint probability density function from the independent probability function of each variable: $p(\boldsymbol{\xi}(\omega)) = \prod_{i=1}^d p_i(\xi_i)$. This assumption allows an independent polynomial representation for each direction in the probabilistic space with the possibility to recover the multidimensional representation by tensorization. The aim of UQ is to find the statistical moments of the solution $u(\boldsymbol{\xi})$.

[1] In the following the exposition is made for a scalar output variable (u) for brevity, but the extension to the multidimensional output case is straightforward.

3. The spatial-TE (sTE) strategy

The aim of the present work is to propose an efficient strategy to solve efficiently stochastic partial differential equation.

In [4], we presented a technique inspired to the classical multiresolution framework of Harten [9, 10], but adapted to the computation of statistics in the stochastic space Ξ in the case of time dependent and eventually discontinuous probability density functions. This so-called truncate and encode strategy (TE) can be employed to obtain non-intrusive solutions in the case of problems defined only in the stochastic space. This basic algorithm is briefly presented in Section 3.1 and 3.2, where the strategy describing the evolution in time is shown.

In this paper, we show how this algorithm could be extended in order to solve stochastic partial differential equation. A detailed description of this new algorithm is then illustrated in Section 3.3.

3.1. The Truncate and Encode strategy

Here, for simplicity, only the 1D case with uniform distribution of points is considered, even if the same conclusions hold for higher dimensional meshes of non structured type (see [5]). In the following, we indicate the generic mesh level k of N_k equally spaced intervals of length h_k as

$$\mathcal{G}^k = \left\{\xi_j^k\right\}_{j=0}^{N_k}, \quad \xi_j^k = jh_k, \quad h_k = 2^k h_0, \quad N_k = N_0/2^k.$$

A representation of the solution on a finest grid is computed starting from a coarsest grid, with a lower number of evaluation of the function (in the space Ξ). The remaining points can be obtained by interpolation under the hypothesis to make an error that can be driven by a threshold parameter ε. The Harten framework consists of three different steps:

- Encoding: the solution represented on the finest mesh $\mathcal{G}^0$ is employed to obtain a hierarchical representation on a nested sequence of levels $k = 1, \ldots, L$ where $\mathcal{G}^k$ are obtained directly from $\mathcal{G}^{k-1}$ without considering the odd points. For each *missing point* $\xi_j^k \in \mathcal{G}^{k+1} - \mathcal{G}^k$, a *detail* or *wavelet* is computed as $d_j^k = u_{2j-1}^{k-1} - \tilde{u}_{2j-1}^{k-1}$, where $\tilde{u}_{2j-1}^{k-1}$ is an approximation of the value employing a whatever interpolation operator $\mathcal{I}(\xi; u^k)$ that interpolates the function u on the level k in the point ξ. In the present work, we have chosen, in order to simplify the exposure, the simplest example, namely a linear interpolation operator. However, the extension to more complex and more accurate interpolation would lead to similar algorithms. The final result of the *encoding* procedure is to obtain a multiresolution u_M representation of u: $(u_M)^{\mathrm{T}} = (d^1, d^2, \ldots, d^L, u^L)$ where $d^k = \{d_j^k\}$ and $k = L$ is the coarsest level. For brevity, the procedure can be re-arranged in matrix form: $u_M = Mu^0$.

- Truncation: to obtain a data compression of the solution at the finest level $\hat{u}^0$, a threshold can be introduced to eliminate the non-significant *wavelets*. In particular, a truncated *detail* is defined as follows

$$\hat{d}_j^k = \begin{cases} d_j^k & \text{if} \quad |d_j^k| > \varepsilon_k \\ 0 & \text{if} \quad |d_j^k| \leq \varepsilon_k. \end{cases} \tag{2}$$

  As a consequence, the truncated multiresolution representation consists in $\hat{u}_M = (\hat{d}^1, \hat{d}^2, \ldots, \hat{d}^L, u^L)$.

- Decoding: once the truncation is performed, the solution on the finest level can be obtained directly from the coarsest one $\hat{u}^0 = M^{-1}\hat{u}_M$. The following estimation holds (see [8] for a proof)

$$||u^0 - \hat{u}^0|| \leq C\varepsilon, \tag{3}$$

  if $\varepsilon_k = \varepsilon/2$.

Now, we can introduce our procedure permitting to perform the encoding and truncation procedure at the same time starting from the coarsest level to the finest. This is necessary if the system is affected by unsteady probability density function, so at each time step a new multiresolution representation should be computed without using information from the previous time steps.

Let us consider only the dependence of a scalar function $u = u(\xi)$ from the stochastic space $\Xi = [0, 1]$. The TE strategy is constituted by the following steps (the notation is the same of the Harten's multiresolution framework, i.e $k = 0$ for the finest level and $k = L$ for the coarsest):

- Initialization

  - Fix a threshold ε (the solution is assumed to be solved with this threshold on the finest grid [2]);

  - Fix an index $m_{\max} \in \mathbb{N}$ for the maximum allowed level ($N_{\max} = N_0 = 2^{m_{\max}}$);

  - Fix an index $m_L \in \mathbb{N}$ for the coarsest level ($N_L = 2^{m_L}$);

  - The condition $m_L < m_{\max}$ must be satisfied.

- Evaluation of the function u at each location at the coarsest level $u(\xi_j^L) = u_j^L$ with $j = 0, \ldots, N_L$ where

$$\mathcal{G}^L = \left\{\xi_j^L\right\}_{j=0}^{N_L}, \quad \xi_j^L = jh_L, \quad h_L = 2^L h_0, \quad N_L = N_0/2^L, \tag{4}$$

and $h_0 = 1/N_0$. Each level can be labeled computing the equivalent index k_{eq}

$$k_{eq} = \log_2\left(\frac{N_0}{N_{k_{eq}}}\right).$$

- Evaluation of the subsequent level, with respect to the coarsest

$$\mathcal{G}^{L-1} = \left\{\xi_j^{L-1}\right\}_{j=0}^{N_{L-1}}, \quad \xi_j^{L-1} = jh_{L-1}, \quad h_{L-1} = 2^{L-1}h_0, \quad N_{L-1} = N_0/2^{L-1}. \tag{5}$$

- Starting of the adaptive strategy by means of a recursive procedure

  A - The *wavelets coefficients* are computed for the present level k as

$$d_j^k = u_j^k - \frac{1}{2}\left(u_{\frac{j+1}{2}}^{k+1} + u_{\frac{j-1}{2}}^{k+1}\right) \quad \text{for} \quad 0 \le j \le N_k \quad \text{with} \quad j \text{ odd;} \tag{6}$$

  This is one of the occurrences where the linear interpolation is used. If more accurate interpolants are used, the detail in (6) will still be the difference between the actual value and the value of the interpolant at the interpolation location.

  B - The wavelets coefficients are compared with the threshold $\varepsilon_k = \varepsilon/2^k$. If $|d_j^k| > \varepsilon_k$ then the two nodes ξ_{2j+1}^{k-1} and ξ_{2j-1}^{k-1} will be flagged as active on the next finer mesh $\mathcal{G}^{k-1}$. If $|d_j^k| < \varepsilon_k$ then the *wavelets* is truncated, i.e. its value is posed zero.

  C - The new level $k-1$ is generated if $k > 0$ and only on the activated points the function u is evaluated.

  D - Moving from a level k to the finer adjacent one $k-1$, three different cases are possible:

  * If $\xi_j^k \in \mathcal{G}^k \cap \mathcal{G}^{k+1}$ then $u_j^k = u_{2j}^{k+1}$ (shifting)
  * If $\xi_j^k \notin \mathcal{G}^k \cap \mathcal{G}^{k+1}$ and it is not flagged then interpolate

$$u_j^k = \frac{1}{2}\left(u_{\frac{j+1}{2}}^{k+1} + u_{\frac{j-1}{2}}^{k+1}\right) \tag{7}$$

    The relation (7) is the second, and last, occurrence where the interpolant is used. In case of more accurate interpolant, (7) is replaced by the value of the interpolant on $\mathcal{G}^{k+1}$ at ξ_j^k.
  * If $\xi_j^k \notin \mathcal{G}^k \cap \mathcal{G}^{k+1}$ and it is flagged as active (by the step B of the algorithm) then evaluate, i.e. call the model.

[2] This is the same hypothesis of the classical MR framework.

E - The algorithm stops when the maximum level is reached or when all the *wavelets* coefficients can be truncated (at a certain level $k > 0$).

Some remarks could be done at this point to make things consistent with the application of this strategy to the computation of statistics. For computing a statistics quantity, the following integrals should be computed,

$$\mathcal{E} = \int_{\Xi} u(\xi) p(\xi) \mathrm{d}\xi, \tag{8}$$

where $\mathcal{E}$ is the expectancy of u dependent on the random parameter ξ with pdf $p(\xi)$ in the space Ξ. The TE strategy presented above is applied to the product of $u(\xi)$ and $p(\xi)$. In the general case of unsteady pdf, this procedure must be also applied at each time step and the information between successive time steps must be exchanged by the time advancing technique presented in the next section.

3.2. An accurate preserving time advancing technique

The aim of the TE strategy and the time stepping technique is to minimize the number of points in the space $\Xi \times T$. The unsteady solution should be solved on all the possible trajectories in the space T, then this implicitly involves to know the solution in all the points in $\Xi \times T$. The procedure we propose relies on the application of a multiresolution encoding and truncation of the solution at each time steps. This ensures that the overall error is bounded by (3). Moving from the initial condition toward the ultimate time step can be performed, for each trajectory, by advancing the overall space Ξ time step by time step. This reflects, in the case of an ordinary differential equation, in the computation of the solution $u(\overline{\xi}, \overline{t})$ in a fixed point $\overline{\xi}$ at the time $\overline{t}$ knowing the solution at the previous time steps for all $\xi \in \Xi$ and $t < \overline{t}$. In a rigorous sense, the solution is known only in a limited set of points, *i.e.* the activated points of the TE strategy. However, relying on the result (3), if a point in the portion of the space $\Xi \times T$ with $t < \overline{t}$ is needed, an interpolation can be performed, with the same operator $\mathcal{I}$ employed by the TE strategy, with an error bounded by ε. The final result is to obtain, for each point $\xi \in \Xi$, some trajectories in T where the evaluations could stop and interpolations could start (from the adjacent ones). Eventually these sequences of interpolations and evaluation can continue to invert virtually at each time step.

A schematic view of these sequences of interpolation and evaluations is reported in figure 1. The points shown are related to the activated points at each time step while the lines indicate the advancing in time that can be performed from a known point (continuous line) or from an interpolated value (dashed lines).

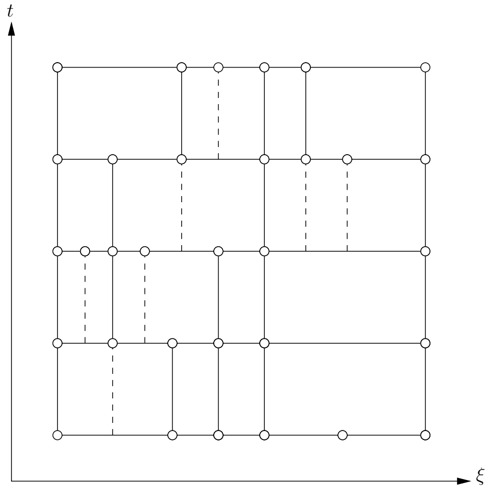


Figure 1: Time advancing. In dashed line the advancing in time from interpolated values and with continuous line the integration from computed value.

3.3. Extension to physical dependent solutions

In this section, we show how the TE strategy, presented in the last two sections, can evolve in the spatial-TE (sTE)strategy, *i.e.* can be extended to partial differential equations. Let us consider partial differential equation defined on 1D physical and stochastic spaces. Obviously, the numerical scheme associated to an adaptive distribution of points in the space $D \times T \times \Xi$, cannot be independent from the specific equation to solve. In this section, the procedure is described in a general way supposing to have a deterministic numerical scheme able to compute the solution $u(\overline{x}, \overline{t}, \overline{\xi})$ knowing all the solutions $u(x, t, \xi)$ for all $x \in D$, $\xi \in \Xi$ and $t < \overline{t}$. In the section §4, an example of the application of the present strategy to the heat equation is illustrated.

The key idea of the algorithm is to fix a finer enough spatial discretization, as well as a time discretization, in order to solve the deterministic problem with the desired accuracy. These requirements are the same of the classical MR approach. In fact, it is clear that the MR scheme cannot produce more accurate solutions than the non compressed finest level solution (remember the estimation (3)). Once the deterministic scheme is provided, the parameters for the TE algorithm, in the stochastic space, must be provided: a maximum level m_{max}, a minimum level m_L and a threshold ε. According to the mathematical setting of the problem, the initial condition must be discretized on the grid $D \times \Xi$ employing the finest resolution level (m_{max}) in the stochastic space and the fixed spatial discretization chosen. Two different cases can arise here: the initial condition is affected by uncertainty or not. However in both cases we can suppose to know analytically the initial condition. After these preliminaries, the spatial-TE (sTE)strategy can be employed as follows. At each time step and for each spatial node, the TE strategy is applied to the associated stochastic space Ξ obtaining the MR representation of the solution $u(\overline{x}, \overline{t}, \xi)$, *i.e.* the representation of the 1D (in this case) stochastic function obtained at a fixed physical space $\overline{x}$ and time $\overline{t}$ location (see figure 2).

Once all the physical points are used by the algorithm, the solution $u(x, \overline{t}, \xi)$ is known. In fact, if a point is evaluated, an *exact* solution is provided for it, otherwise it can be interpolated, employing the operator $\mathcal{I}$ of the TE strategy along the stochastic space with an error bounded by the threshold ε. This procedure continues until the final time step is reached. Obviously, depending on the spatial and time discretization adopted for the problem, different stencils in the physical space could be required. This stencil must be assembled, knowing the solution at the previous time step, eventually by interpolation along the stochastic space. For instance, in the case of a finite element discretization with a fourth order Runge-Kutta scheme the stencil can be identified in an automatic way using only the finite element mass matrix and stiffness matrix. All the details are reported in the section §4. Once the stencil is reconstructed, the value of the function in all the nodes belonging to the stencil must be computed. Two different situation are possible: the point has been already computed or an interpolation must be performed (with the interpolation operator $\mathcal{I}$ along the stochastic space). We remark that the interpolation must be performed always in the stochastic direction while the stencil assembling procedure could require to use different multiresolution representation at different physical locations (see figure 2).

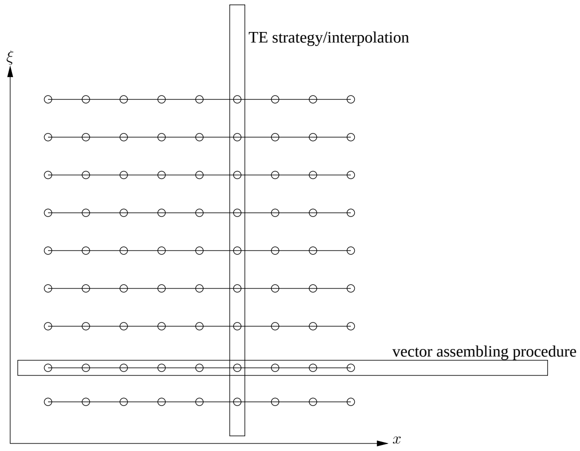


Figure 2: Sketch of the interpolation (same direction of the TE strategy) and stencil assembling procedures.

The entire sTE strategy can be summarized as follows :

- Preliminary
  - Choose a deterministic solver with a spatial fixed discretization and a proper time discretization technique;
  - Fix the parameters for the TE strategy: finest level m_{max}, coarsest level m_L and threshold ε;

- sTE
  - For each time step and for each spatial node, the TE strategy should be applied in order to represent the solution along the stochastic space;
  - The proper stencil must be assembled using different spatial locations (see figure 2); some interpolations along the stochastic space could be necessary at this stage.

In the next section, this algorithm will be adapted to the heat equation discretized by a finite element method in the physical space and a fourth order Runge-Kutta method in time.

4. A finite element deterministic solver

In this section, we present the discretization of the heat equation, described by a parabolic partial differential equation with a random initial condition.

Let us describe the equations for the homogeneous 1D case ($x \in D = [0, 1]$) and temporal domain $t \in T = [0, t_f]$:

$$\begin{cases} \dfrac{\partial u(x,t,\xi)}{\partial t} = \nu \dfrac{\partial^2 u(x,t,\xi)}{\partial x^2}, \quad \xi \in \Xi \\ u(0,t,\xi) = u(1,t,\xi) = 0, \quad \text{for} \quad t \in T \\ u(x,0,\xi) = u_0(x,\xi), \end{cases} \tag{9}$$

where the initial conditions is supposed uncertain.

The problem (9) can be recast in the weak form multiplying both side for a test function $v \in V = H_0^1(0,1)$, *i.e.* the space $H^1(0,1)$ with null elements at the boundary of the domain, and then integrating over the physical space $D = [0,1]$:

$$\int_D \frac{\partial u(x,t,\xi)}{\partial t} v \, \mathrm{d}x + \int_D \nu \frac{\partial u(x,t,\xi)}{\partial x} \frac{\partial v}{\partial x} \, \mathrm{d}x = 0. \tag{10}$$

The Galerkin formulation of the problem can be obtained searching the (approximated) solution in the finite dimensional space: $u_h = \sum_{i=1}^{N_h} u_i(t;\xi)\phi_i(x) \in V_h$. The space V_h is the so-called finite element space of basis $\{\phi_j\}_{j=1}^{N_h}$.

After some manipulations (for more details see 6), the following algebraic problem is obtained :

$$\mathrm{M}\frac{\mathrm{d}\mathbf{U}(t)}{\mathrm{d}t} = -\nu \mathrm{A}\mathbf{U}, \tag{11}$$

where the so-called mass M and stiffness A matrices are of $(N_h \times N_h)$ dimension and the vector $\mathbf{U}$ is employed to collect all the degree of freedom of the problem $\mathbf{U}(\mathbf{t}) = \{u_1(t), u_2(t), \ldots, u_{N_h}(t)\}^{\mathrm{T}}$. We remark here that the present formulation is quite general not depending on the number of physical space dimensions. As reported in 6, the matrices M and K are quite sparse and symmetric. In particular, if the finite element space of linear functions is employed, the matrices are both tridiagonal.

Finally, the initial parabolic partial differential system of equations, is reduced to a system of ordinary differential equations (ODEs). In the next section, the time integration technique is illustrated.

4.1. A recast fourth-order Runge-Kutta

In this section, we aim to use a time integration technique permitting to apply the sTE strategy in order to solve the stochastic partial differential problem (9). The TE technique, described in Section 3, requires the solution of the problem in a specific point of the space $(\overline{x}, \overline{t}, \overline{\xi})$, whenever all the solution at the previous time steps $t < \overline{t}$ are available. This means that the numerical scheme adopted to solve the system of ODEs should be able to compute the solution in a certain node $\overline{i}$ at the time $\overline{t}$ knowing the solution in all the nodes at time $t < \overline{t}$. As it has been described in §3, the TE strategy is employed in the stochastic direction, while, obviously, the deterministic solver produces solutions in the physical space. The coupling between these two spaces will be described more in details in Section §4.3.

For instance, let us suppose to know the solution $\mathbf{U}(\mathbf{t})$ for $t < \bar{t}$ and that the deterministic solver is able to compute the $\bar{i}$-th coefficient of the vector $\mathbf{U}(\bar{\mathbf{t}})$, *i.e.* the $\bar{i}$-th degree of freedom of the finite element expansion of the solution $u_h \in V_h$. In this work, we choose to use an explicit time integration technique, in particular the fourth order Runge-Kutta scheme [15], described as follows for a Cauchy problem :

$$\begin{cases} \dot{y}(t) & = f(t, y(t)) \quad t \in [0, t_f] \\ y(0) & = y_0, \end{cases} \tag{12}$$

where $y \in C(0, t_f)$ can be formulated as [15]

$$y_{n+1} = y_n + \frac{\Delta t}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \tag{13}$$

where $y_n = y(t_n)$ with $t_n = n\Delta t$ and

$$\begin{cases} k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2}k_1\right) \\ k_3 = f\left(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2}k_2\right) \\ k_4 = f\left(t_n + \Delta t, y_n + \Delta t k_3\right) : \end{cases} \tag{14}$$

In this form, the Runge-Kutta (RK4) method is extended to a system of ODEs in a straightforward manner.

In the case of the heat equation, the system of ODEs (11) can be recast to a set of decoupled equations if the so-called *mass lumping* technique is adopted. In particular, as shown in 6, if a trapezoidal integration technique is employed to compute the term of the mass matrix M, a diagonal matrix can be obtained and, in the case of linear element, each term of the mass matrix can be computed as the summation by rows of the elements

$$\hat{m}_{ii} = \sum_{i=1}^{N_h} m_{ij}, \tag{15}$$

where m_{ij} indicates the generic element of the mass matrix M at the i–th row and j–th column.

If we indicate with $\hat{\mathrm{M}}$ the corresponding lumped mass matrix, the system of ODEs can be written as

$$\frac{\mathrm{d}\mathbf{U}(t)}{\mathrm{d}t} = -\nu\hat{\mathrm{M}}^{-1}\mathrm{A}\mathbf{U}(\mathbf{t}) = \mathbf{f}(\mathbf{U}(\mathbf{t})) \tag{16}$$

and the corresponding RK4 scheme is

$$\begin{cases} \mathbf{U}_{n+1} = \mathbf{U}_n + \frac{\Delta t}{6}\left(\mathbf{k_1} + 2\mathbf{k_2} + 2\mathbf{k_3} + \mathbf{k_4}\right) \\ \quad \mathbf{k_1} = -\nu\hat{\mathrm{M}}^{-1}\mathrm{A}\mathbf{U}_n \\ \quad \mathbf{k_2} = -\nu\hat{\mathrm{M}}^{-1}\mathrm{A}\left(\mathbf{U}_n - \nu\hat{\mathrm{M}}^{-1}\mathrm{A}\mathbf{U}_n\right) = \mathbf{k_1} + \nu^2\frac{\Delta t}{2}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^2\mathbf{U}_n \\ \quad \mathbf{k_3} = -\nu\hat{\mathrm{M}}^{-1}\mathrm{A}\left(\mathbf{U}_n + \frac{\Delta t}{2}\mathbf{k_1} + \nu^2\frac{\Delta t^2}{4}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^2\mathbf{U}_n\right) = \mathbf{k_2} - \nu^3\frac{\Delta t^2}{4}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^3\mathbf{U}_n \\ \quad \mathbf{k_4} = -\nu\hat{\mathrm{M}}^{-1}\mathrm{A}\left(\mathbf{U}_n + \Delta t\,\mathbf{k_1} + \nu^2\frac{\Delta t^2}{2}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^2\mathbf{U}_n - \nu^3\frac{\Delta t^3}{4}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^3\mathbf{U}_n\right) \\ \qquad = \mathbf{k_1} + \nu^2\Delta t\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^2\mathbf{U}_n - \nu^3\frac{\Delta t^2}{2}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^3\mathbf{U}_n + \nu^4\frac{\Delta t^3}{4}\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^4\mathbf{U}_n \end{cases} \tag{17}$$

To compute the $\bar{i}$–th term of the vector $\mathbf{U}(t_{n+1})$ is then necessary to compute the corresponding term $\bar{i}$–th term of each vector $\mathbf{k_1}$, $\mathbf{k_2}$, $\mathbf{k_3}$ and $\mathbf{k_4}$. This can be done efficiently if four matrices are stored at the beginning of the computation

$\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)$, $\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^2$, $\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^3$, $\left(\hat{\mathrm{M}}^{-1}\mathrm{A}\right)^4$. We remark that $\hat{\mathrm{M}}^{-1}$ indicates the inversion of a diagonal matrix and then the product $\hat{\mathrm{M}}^{-1}\mathrm{A}$ can be done in a very non expensive way. Once the four matrices are computed, the $\bar{i}$–th term of the four vectors $\mathbf{k_1}$, $\mathbf{k_2}$, $\mathbf{k_3}$ and $\mathbf{k_4}$, can be computed (less than a multiplying factor), as the scalar product between the $\bar{i}$–th row vector (of each of the four $\mathbf{k_1}$, $\mathbf{k_2}$, $\mathbf{k_3}$ and $\mathbf{k_4}$ matrices) and the vector $\mathbf{U}_n$ (already known from the stencil assembling procedure). This procedure allows to select automatically the stencil needed by the time integration technique. For adapting this deterministic scheme to the solution of the stochastic parabolic equation by the sTE strategy, the reconstruction of the vector $\mathbf{U}_n$ from the different multiresolution representation (one for each physical node) of the solution is needed; this will be described in detail in section 4.3.

Obviously, if the deterministic scheme described above is applied for each time step to each node of the physical grid, the time dependent solution of the problem on the whole physical space $x \in D = [0, 1]$ can be computed when a fixed value for the parameter $\xi \in \Xi = [0.2, 0.8]$ is chosen. The sTe strategy, as shown in §3, needs to fix a physical mesh on which the deterministic solution can be represented with the desired accuracy. To identify the proper mesh to employ in the section §5, a spatial convergence study is reported in the next section.

4.2. Space convergence for the deterministic solver

In this section, the space convergence properties of the deterministic scheme is presented. For this reason, a reference solution should be computed. Then, the equivalent modal problem using the method originally proposed by Fourier is solved.

The solution can be searched as a product of two functions depending only from the space and time, respectively. This technique, called separation of variable, with $u(x,t) = f(x)g(t)$, leads to

$$\frac{1}{g(t)}\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \nu\frac{1}{f(x)}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} = -\lambda, \tag{18}$$

where λ must be a constant value not depending neither by x or t. Non trivial solution exist only if $\lambda > 0$ as

$$\begin{cases} f(x) = A\sin(\sqrt{\lambda}x) + B\cos(\sqrt{\lambda}x) \\ g(t) = Ce^{-\nu\lambda t}. \end{cases} \tag{19}$$

The application of the boundary condition to the spatial function $f(x)$ makes possible to compute $B = 0$ and $\lambda = n^2\pi^2$ where the integer n indicates the n–th mode of the function $f(x)$:

$$f(x) = \sum_{n=1}^{\infty} A_n \sin(n\pi x). \tag{20}$$

The solution $u(x,t)$ is the product of the two functions ($f(x)$ and $g(t)$) and it becomes

$$u(x,t) = C\sum_{n=1}^{\infty} A_n \sin(n\pi x)e^{-\nu\lambda t} = \sum_{n=1}^{\infty} H_n \sin(n\pi x)e^{-\nu\lambda t} = \sum_{n=1}^{\infty} H_n \Psi_n(x) \tag{21}$$

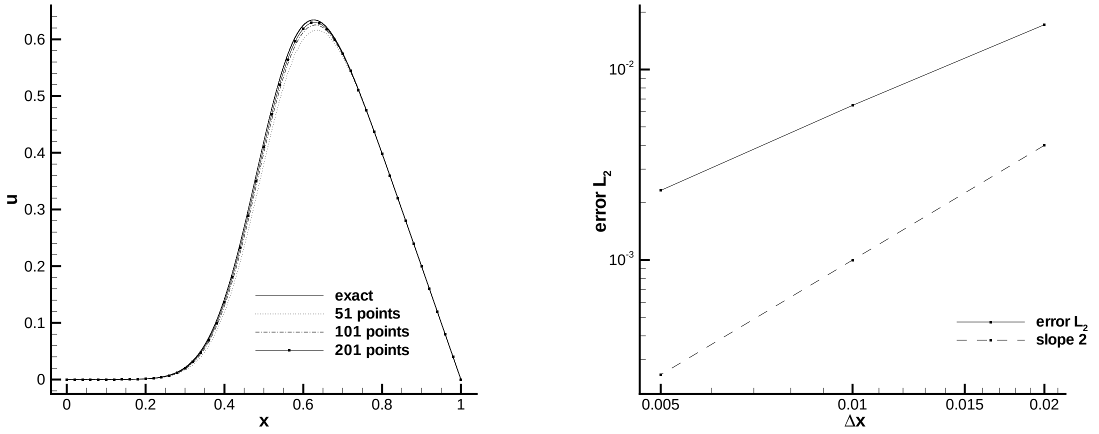
The amplitude H_n for each mode can be obtained by normalization employing the orthogonality between modes, *i.e.* $\int_0^1 \Psi_i\Psi_j \mathrm{d}x = H_n\delta_{ij}$ where δ_{ij} is the Kronecker delta function, and the initial condition $u_0(x)$

$$H_n \int_0^1 \sin(n\pi x)\sin(n\pi x)\mathrm{d}x = \frac{1}{2}H_n = \int_0^1 u_0(x)\sin(n\pi x)\mathrm{d}x \longrightarrow H_n = 2\int_0^1 u_0(x)\sin(n\pi x)\mathrm{d}x. \tag{22}$$

The modal truncated solution with a large number of modes $N_{mod} = 10\,000$ with each term H_n computed by a trapezoidal rule on an equally spaced mesh of 100 000 points has been employed as reference solution. In the figure 3, different solutions computed on uniform meshes of 51, 101 and 201 points are reported in 3(a), while the errors measured in norm L_2 are shown in 3(b) for the same meshes.

As it can be observed from these results, the mesh with 101 points shows to be finer enough to achieve a good spatial accuracy and then it will be adopted for the computations. For stability and accuracy requirements, a time step equal to $\Delta t = 0.001$ (the same employed in the spatial convergence results) is a good trade-off.

The coupling between the deterministic scheme and the TE strategy described in section 3 is described with more detail in the next section.



(a)

(b)

Figure 3: Space convergence for the FE deterministic solver. (a) Solutions with different spatial resolutions and (b) L_2 error norm of the solution with respect the reference modal solution ($\nu = 0.01$ and $a = 1$).

4.3. sTE strategy applied to the 1D heat equation

In this section, the algorithm for the stochastic heat equation is presented. The deterministic solver employs a spatial finite element discretization and a Runge-Kutta method to integrate the solution in time. The scheme is able to compute a specific degree-of-freedom $u_{\bar{i}}(t_{n+1})$ when the vector of all the degree-of-freedom $\mathbf{U}(t_n)$ at the previous time step is provided. This is not dependent on the finite element space, *i.e.* the degree of the basis functions. A sequence of evaluations must be performed for all the activated points (in the TE strategy) at different locations in the physical and stochastic space. However, the vector $\mathbf{U}(t_n)$ could not be available for each parameter ξ (see figure 2). This issue is solved in the sTE strategy performing an interpolation, along the stochastic space, in order to compute the value at a certain physical location by means of the stencil assembling procedure.

The complete algorithm for the sTE strategy applied to the heat equation (9) is as follows :

- Preliminary

  - A fixed uniform spatial resolution is fixed with points: $\{x_i\}_{i=0}^{N+1}$;

  - A uniform time discretization is chosen $t_n = n\Delta t$, where $\Delta t = t_f/N_t$ with N_t number of time steps;

  - The parameters for the TE strategy are fixed: m_L, m_{max} and ε;

  - The mass matrix M is computed, lumped and inversed obtaining $\hat{\mathrm{M}}^{-1}$;

  - The stiffness matrix A is computed;

  - The four matrices $(\hat{\mathrm{M}}^{-1}\mathrm{A})$, $(\hat{\mathrm{M}}^{-1}\mathrm{A})^2$, $(\hat{\mathrm{M}}^{-1}\mathrm{A})^3$, $(\hat{\mathrm{M}}^{-1}\mathrm{A})^4$ are computed and stored.

- sTE strategy

  - For each time step all the (internal) spatial nodes $\{x_i\}_{i=1}^{N}$ are considered;

  - For each spatial node $x_{\bar{i}}$ considered, a MR representation is obtained for $u(x_{\bar{i}}, t_{n+1}, \xi)$ in the stochastic space;

    A- To evaluate the solution in $(x_{\bar{i}}, t_{n+1}, \xi_{\bar{j}})$, the vector $\mathbf{U}(t_n, \xi_j)$ must be assembled;

    B- $k_{1\bar{i}}$, $k_{2\bar{i}}$, $k_{3\bar{i}}$, $k_{4\bar{i}}$ are computed employing the $\bar{i}$–th row of the four matrices $(\hat{\mathrm{M}}^{-1}\mathrm{A})^n$ and $\mathbf{U}(t_n, \xi_j)$;

    C- Evaluation: $u(x_{\bar{i}}, t_{n+1}, \xi_{\bar{j}}) = U_{\bar{i}}(t_n, \xi_{\bar{j}}) + \dfrac{\Delta t}{6}(k_{1\bar{i}} + 2k_{2\bar{i}} + 2k_{3\bar{i}} + k_{4\bar{i}})$.

The vector assembling procedure for the vector $\mathbf{U}(t_n, \xi_{\bar{j}})$, is illustrated in the case of linear interpolation as follows:

- For all the nodes $\{x_i\}_{i=1}^N$, if the point $(x_i, t_n, \xi_{\bar{j}})$ has been already computed, the value is stored in $U_i(t_n, \xi_{\bar{j}})$;
- Otherwise the left and right values are identified for interpolation as follows:
  - Left value: greater value of ξ_j at the $i-$th spatial position less than $\xi_{\bar{j}}$;
  - Right value: point ξ_{j+1} at the $i-$th spatial position;
  - Interpolation: linear interpolation between ξ_j and ξ_{j+1} at the stochastic position $\xi_{\bar{j}}$.

The procedure described in this section is able to reduce the overall number of evaluations in the space $D \times T \times \Xi$ by determining the important point, *i.e.* the points that cannot be interpolated within the prescribed error with the chosen interpolation operator. The final result is an unsteady pattern of the activated points in the space $D \times \Xi$ on which the solution is computed by means of the deterministic solver, while the remaining point are interpolated.

5. Numerical results

In this section, the sTE strategy is applied to some numerical problems: a steady discontinuous function (§5.1); an ordinary differential equation (§5.2) with the application of the time integration strategy reported in §3.2. Finally, the stochastic partial differential equation (9) describing the heat conduction and the evolving temperature u along a 1D rod subjected to an uncertain and discontinuous initial condition, is solved by means of the sTE strategy in (§5.3).

In this section, the expectancy $\mathcal{E}$ and the variance Var for a generic function $f(\xi)$, are computed according to the following definitions

$$\begin{aligned} \mathcal{E}(f(\xi)) &= \int_\Xi f(\xi)p(\xi)\mathrm{d}\xi \\ \mathrm{Var}(f(\xi)) &= \int_\Xi (f(\xi) - \mathcal{E}(f(\xi)))^2 p(\xi)\mathrm{d}\xi, \end{aligned} \tag{23}$$

where the probability distribution $p(\xi)$ is chosen systematically as uniform.

All the results reported in this section are compared to two classical methods in uncertainty quantification, namely Monte Carlo (MC) and Polynomial Chaos (PC). These two methods are employed in a complete non-intrusive way and the reference solution is assumed to be the fully converged Monte Carlo solution.

5.1. Steady problem

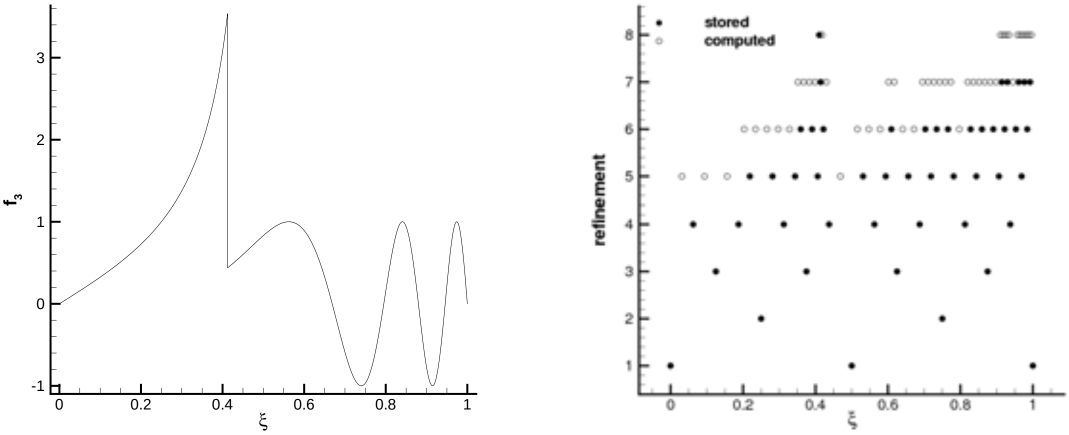
The first example is a function $f(\xi) : \Xi \rightarrow \mathbb{R}$ where ξ is a random parameter having an uniform distribution $\xi \sim \mathcal{U}[0, 1]$. Function f_3 is a piecewise function, composed by a tangent and a wave sine function with decreasing wavelength (see figure 4(a)):

$$f_3(\xi) = \begin{cases} \tan(\xi\pi) & \xi \leq 0.41234 \\ \sin(5\pi\xi^4) & \xi > 0.41234 \end{cases} \tag{24}$$

The coarsest level is assumed to be equal to 2^1 ($m_l = 1$) intervals, while the finest one to 2^8 ($m_{max} = 8$). The threshold is fixed to $\varepsilon = 10^{-1}$ with a variation related to the refinement level (k) equal to $\varepsilon_k = \varepsilon/2^k$. In figure 4(b), the sequence of evaluated points (N_{eval}) is reported. The circles represent the evaluations of the function $f(\xi)$, while a full black dots indicate the activated, i.e. greater than the threshold ε_k, *wavelets* N_w. It is evident that the algorithm is capable to follow the discontinuity and to add some points where needed, *i.e.* in regions with high gradients in the stochastic space.

In the table 1, the compression properties of the sTE strategy are reported when applied to the function f_3. In particular, the compression μ_{cr} and the evaluation τ ratios reported in the table 1 are computed as follows

$$\begin{aligned} \mu_{cr} &= \frac{2^{m_{max}} + 1}{N_w + 2^{m_L} + 1} \\ \tau &= \frac{2^{m_{max}} + 1}{N_{eval}}. \end{aligned} \tag{25}$$



(a)

(b)

Figure 4: Function $f_3(\xi)$ (a) and the pattern of computed and activated (stored) points (b).

They indicate the ratio between the number of points of the non-compressed solution and the number of activated wavelets (these sets include the points of the coarsest level) and the ratio between the number of points at the finest level and the number of evaluations N_{eval} needed by the TE strategy, respectively.

The error norms reported in the table 1 are computed in the L_1 and L_∞ space as

$$\begin{aligned} \mathrm{err}_{L_1} &= \|f^0 - \hat{f}\|_{L_1} = & \frac{1}{N}|f_i^0 - \hat{f}_i| \\ \mathrm{err}_{L_\infty} &= \|f^0 - \hat{f}\|_{L_\infty} = & \max_i |f_i^0 - \hat{f}_i|, \end{aligned} \tag{26}$$

where f^0 is the function at the finest level and $\hat{f}$ is the compressed function, i.e. the function evaluated only in the set of points corresponding to the activated wavelets.

m_{max}	N_{sto}	N_{eval}	μ	τ	err L_1	err L_∞
5	21	29	1.571429	1.137931	0.1341053E-01	0.7356531E-03
6	31	49	2.096774	1.326531	0.1160966E-01	0.8490046E-03
7	39	73	3.307692	1.767123	0.1003391E-01	0.1228993E-02
8	49	95	5.244898	2.705263	0.1291483E-01	0.1506392E-02
9	58	113	8.844828	4.539823	0.8991428E-02	0.1307482E-02

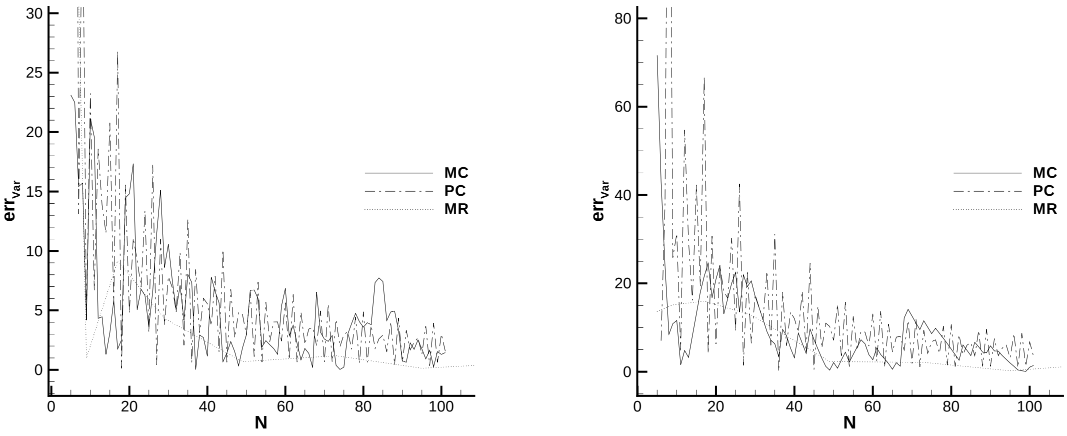
Table 1: Final result for the function f_3 ($\varepsilon = 10^{-1}$)

Thanks to the adaptive distribution of points, the present strategy allows computing the statistical moments very efficiently even with a simple quadrature formula (like the composite trapezoidal rule [15]). This is not the case for MC or PC methods.

The percentage errors with respect the reference MC solution with 2×10^6 deterministic runs is computed as follows

$$\begin{aligned} \mathrm{err}_{\mathcal{E}} &= \frac{|\mathcal{E}-\mathcal{E}_{\mathrm{exact}}|}{\mathcal{E}_{\mathrm{exact}}}100 \\ \mathrm{err}_{Var} &= \frac{|Var-Var_{\mathrm{exact}}|}{Var_{\mathrm{exact}}}100. \end{aligned} \tag{27}$$

They are reported in figure 5 both for mean and variance. The number of points for the PC method are $N = n_0 + 1$, where n_0 is the total degree of the polynomial representation. Concerning the proposed algorithm, several solutions



(a)

(b)

Figure 5: Percentage error with respect the reference MC solution for the mean (a) and the variance (b).

are obtained by varying the maximal level allowed between 2^2 and 2^9 with the coarsest level equal to 2^1 and the threshold $\varepsilon = 10^{-1}$.

The adaptive strategy displays better results both in terms of accuracy and efficiency with respect to the MC and PC methods. For MC and PC, an high non smooth behavior arises when increasing the number of point. This is due to the presence of discontinuities that can prevent the convergence of these quadrature techniques.

5.2. A differential ordinary equation (0D-1D)

In this section, the case of an ordinary differential equation is addressed. In the following, this case is indicated as 0D in the physical space and 1D in the stochastic space, because there is only one uncertainty affecting the solution of the problem. An ordinary differential problem, extracted from [13], has been modified as follows

$$\begin{cases} \frac{\mathrm{d}\rho}{\mathrm{d}t} = \alpha(\overline{\rho} - \rho) - \gamma\rho - \beta(\rho - \overline{\rho})\rho^2 \\ \overline{\rho} = 1 + \frac{1}{2}\sin(5\omega + 8/5) \\ \beta = 20\omega, \end{cases} \tag{28}$$

where $\alpha = 1$, $\gamma = 0.01$ and $\omega \in \mathcal{U}[0,1]$. A discontinuous initial solution in the stochastic space is chosen in order to address a more challenging problem with respect to the one proposed in [13] :

$$\rho(t=0) = \begin{cases} 3/4 \ \text{ if } \ 0.3 < \omega < 0.7 \\ 0 \quad \text{otherwise.} \end{cases} \tag{29}$$

The time integration is performed by means of an explicit Runge-Kutta scheme, the so-called RK4, with a time step $\Delta t = 0.01$. The multiresolution representation at each time step allows advancing the solution in time along patches constituted by true evaluations and interpolations thanks to the accuracy reconstruction embedded in the multiresolution framework. The final results is a refine/derefine capability in the time-stochastic domain that suits very well the efficiency requirement needed in complex and high costly applications. In the figure 6, the pattern in the space $t - \omega$ of the computed, i.e. evaluated points, is reported.

The error of the statistical moments, are reported in figure 7 with respect to a MC reference solution of 2×10^6 points at each time step ($N = 400 \times 10^6$ evaluations in the $\omega - t$ space). Dealing with an unsteady solution, a L_1 norm

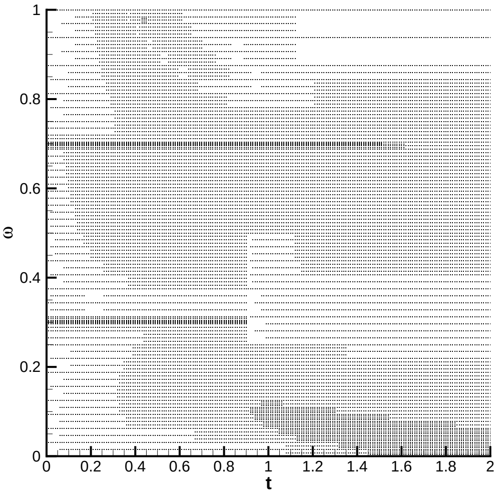


Figure 6: Pattern in the $t - \omega$ space of the computed points of equation 29.

(in time) is employed according to the following definitions

$$
\begin{aligned}
\mathrm{err}_{\mathcal{E}}|_{L_1} &= \|\mathcal{E}(\rho) - \mathcal{E}(\hat{\rho})\|_{L_1} &&= \frac{1}{N_t}\sum_{i=1}^{N_t}\left|\frac{\mathcal{E}_i(\rho) - \mathcal{E}_i(\hat{\rho})}{\mathcal{E}_i(\hat{\rho})}\right|, \\
\mathrm{err}_{\mathrm{Var}}|_{L_1} &= \|\mathrm{Var}(\rho) - \mathrm{Var}(\hat{\rho})\|_{L_1} &&= \frac{1}{N_t}\sum_{i=1}^{N_t}\left|\frac{\mathrm{Var}_i(\rho) - \mathrm{Var}_i(\hat{\rho})}{\mathrm{Var}_i(\hat{\rho})}\right|,
\end{aligned}
$$

where $\mathrm{err}_{\mathcal{E}}$ and $\mathrm{err}_{\mathrm{Var}}$ are the errors for the expectancy and the variance. The solution ρ is compared to the reference solution $\hat{\rho}$ discretized with the same total number of time steps equal to $N_t = \Delta t \times t_f$, where the total time of the simulation t_f is assumed equal to two.

The strategy presented in this work exhibits the fastest convergence and a smoother behavior with respect to Monte Carlo and the Polynomial Chaos both for mean and variance. Similar results are obtained for different norms (L_2, L_∞) not reported here for brevity.

5.3. A partial differential equation (1D-1D)

In this section, the solution of the stochastic parabolic partial differential equation described in (9) is addressed. This unsteady problem is 1D in the physical space and 1D in the stochastic space.

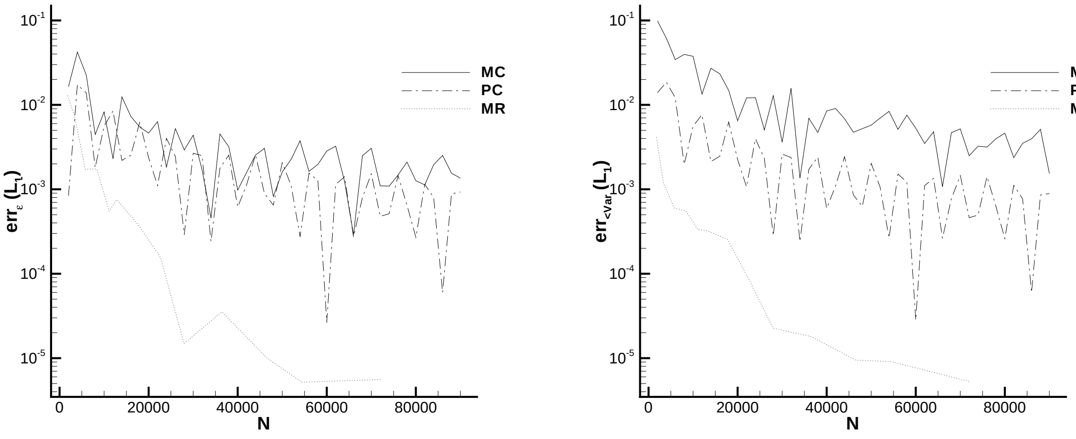
The diffusivity is assumed to be equal to $\nu = 0.01$ and the parameter of amplitude related to the initial condition (30) to $a = 1$.

Let us consider a discontinuous initial condition as follows

$$
u_0(x,\xi) = \begin{cases} 0 & \text{if} \quad x < \xi \\ \dfrac{a}{\xi - 1}(x-1) & \text{if} \quad x \geq \xi, \end{cases} \tag{30}
$$

with the stochastic parameter $\xi \in \Xi = [0.2, 0.8]$ with uniform distribution in Ξ.

For instance, the initial condition for the parameter $\xi = 0.5$ is reported in figure 8 (for $a = 1$). Relying on the convergence study reported in section 4.2, let us consider a physical domain defined in $D = [0, 1]$ discretized by an uniform mesh of $N_x = 101$ nodes and a uniform time step equal to $\Delta t = 0.001$ for a total time of simulation equal to $t_f = 0.5$ ($N_t = 500$ time steps).



(a)

(b)

Figure 7: L_1 norm of the errors for the mean and variance of $\rho(t)$.

Concerning the sTE strategy, the initial condition should be discretized on the finest mesh ($N_x \times (2^{m_{max}} + 1)$) in the space $x - \xi$. The initial condition 9(a) of the problem and the meshes corresponding to the solution at time $t = 0.001$ 9(b), $t = 0.25$ 10(a) and $t = 0.5$ 10(b) are reported for the parameters $m_L = 3$, $m_{max} = 11$ and $\varepsilon = 10^{-1}$.

For performing more accurate comparison, a signal is extracted at fixed space locations. Three different probes at the spatial locations $x = 0.2$ (P1), $x = 0.5$ (P2) and $x = 0.8$ (P3) are considered. For each one of these probes, the mean and variance are stored as functions of the time. The error norms (in time) for the L_1 and L_2 spaces are computed as follows

$$\mathrm{err}_{\mu^m}\Big|_{L_p} = \|\mu^m(u,\, t) - \mu^m(\overline{u},\, t)\|_{L_p} = \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{\mu_i^m(u,\, t) - \mu_i^m(\overline{u},\, t)}{\mu_i^m(\overline{u},\, t)} \right|^p \right)^{1/p}, \tag{31}$$

while for the L_∞ space

$$\mathrm{err}_{\mu^m}\Big|_{L_\infty} = \|\mu^m(u,\, t) - \mu^m(\overline{u},\, t)\|_{L_\infty} = \max_i \left| \frac{\mu_i^m(u,\, t) - \mu_i^m(\overline{u},\, t)}{\mu_i^m(\overline{u},\, t)} \right|, \tag{32}$$

where the reference solution is indicated as $\overline{u}$ and μ^m indicates both mean and variance. The reference solution is obtained in this case with a fully converged MC solution with the same spatial grid ($N_x = 101$), the same time discretization ($\Delta t = 0.001$) but a number of points in the stochastic space equal to $N_\xi = 2.5 \times 10^6$.

The results for mean corresponding to the three probes are reported in the figures 11, 12 and 13, respectively. The sTE strategy is applied with $m_L = 6$, m_{max} between 8 and 16 with a threshold equal to $\varepsilon = 10^{-1}$. MC and PC results obtained on the same physical mesh and with the same time discretization are also reported. In particular, the two methods are employed with a number of points, in the stochastic space, varying between $N_\xi = 100$ and $N_\xi = 300$ for MC and degree between 100 and 300 for PC. In all the presented results, the number of points N represent the overall number of point of the grid in $D \times T \times \Xi$ equal to $N = N_x \times N_t \times N_\xi$.

All the results display systematically very good performances of the presented approach both in term of accuracy and efficiency. The sTE strategy converges smoothly (and in a monotone way) to higher accurate solutions when a large number of points is considered. This is not the case for MC and PC.

The results concerning the variance for the probes P1, P2 and P3 are reported in figures 14, 15 and 16, respectively.

Except for the variance computed at the probe P2, the behavior of the proposed approach is monotone and smoother than both MC and PC. A lower error with a lower global number of points is attained by the sTe strategy

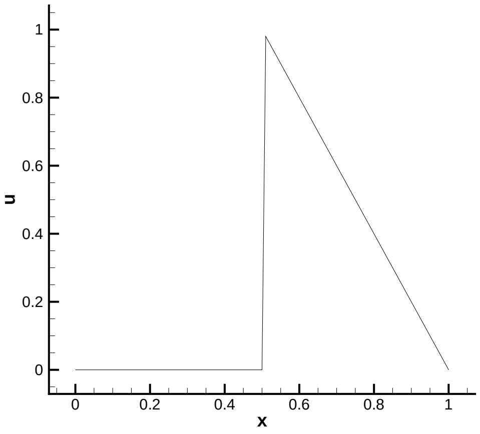
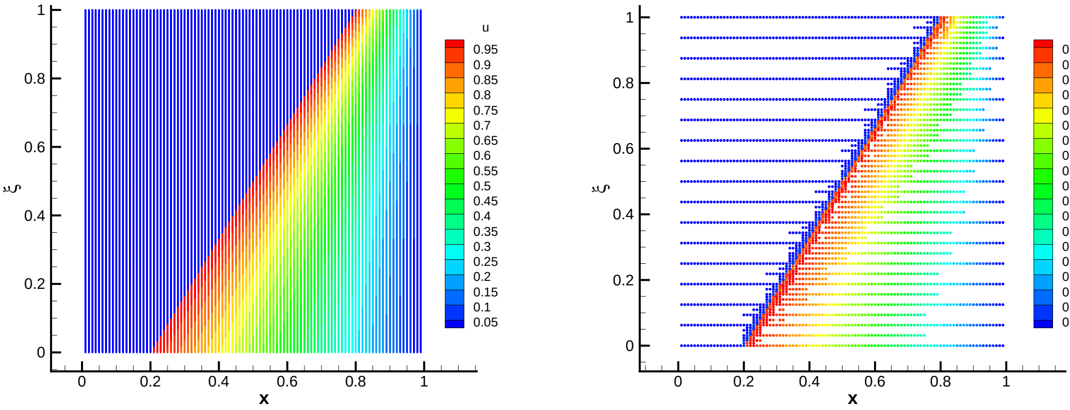


Figure 8: Initial condition for the problem (9) with the stochastic parameter $\xi = 0.5$ on a mesh with 101 equally spaced points.



(a)

(b)

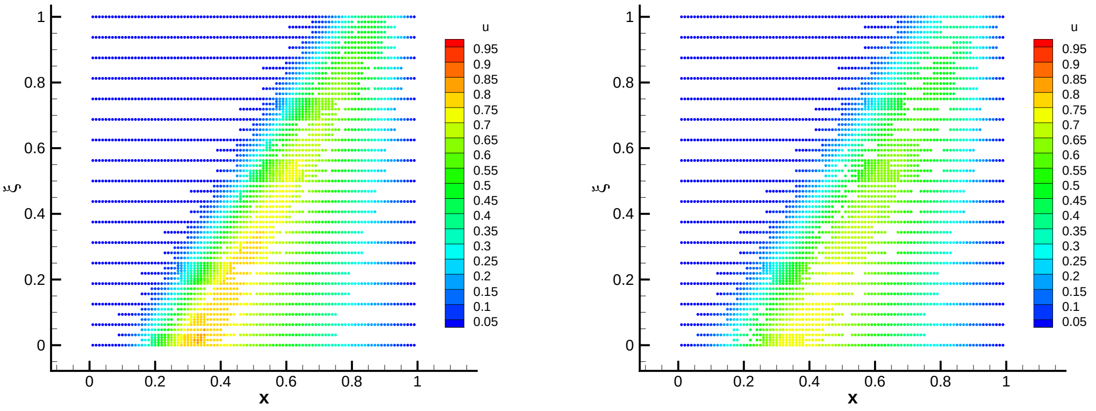
Figure 9: Meshes corresponding to the initial condition (a) of the problem (9) and the derefined mesh after the first time step (b).

with respect to MC and PC. The worse behavior of MC and PC can be justified with the presence of discontinuities in the physical space.

We expect that the sTE strategy will perform much more better, with respect the MC and PC methods, if the solution exhibits a non smooth behavior along the stochastic space. In order to clarify the problem at the probe P2, the solution relative to this probe ($x = 0.5$) is reported as a function of the stochastic space in the figure 17 at the final time step. As clearly shown in figure 17, several discontinuities arise in this case in the stochastic space, even if all the solutions of the heat problem are smooth in the physical space, except for the initial conditions.

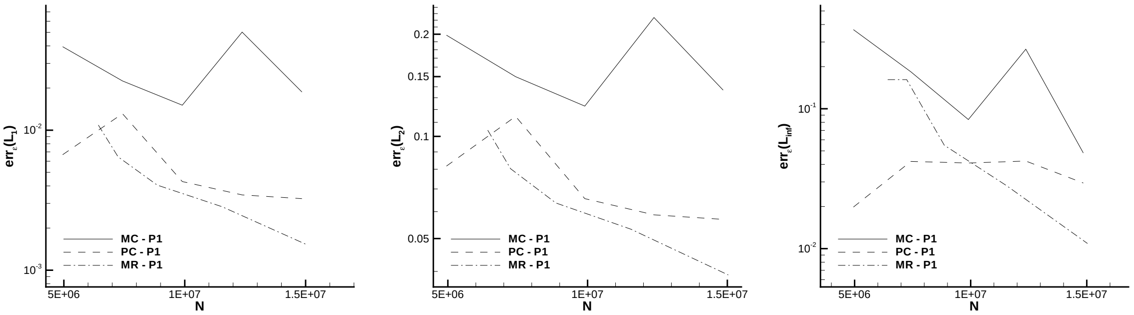
6. Concluding remarks

This paper presents an innovative adaptive strategy for stochastic differential equations, the sTE algorithm, inspired to the classical Harten's framework. A representation of the solution on a finest grid is computed starting



(a) (b)

Figure 10: Meshes corresponding to the half time (t = 0.25) simulation (a) and the final time (t = 0.5) pattern (b).

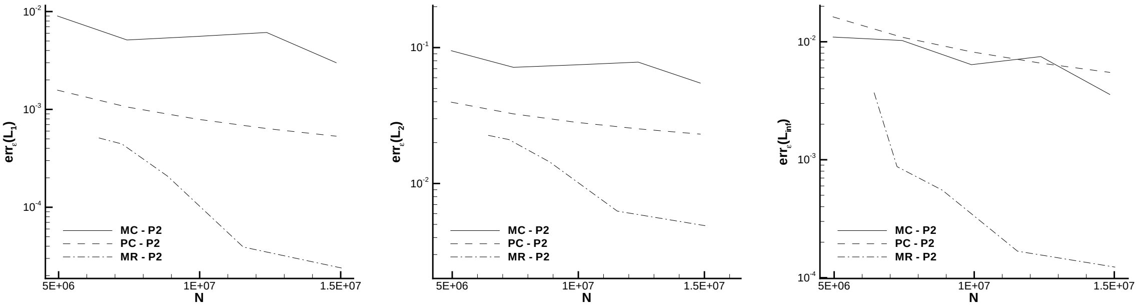


(a) (b) (c)

Figure 11: Error norms of the mean of the variable u, corresponding to the probe P1, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

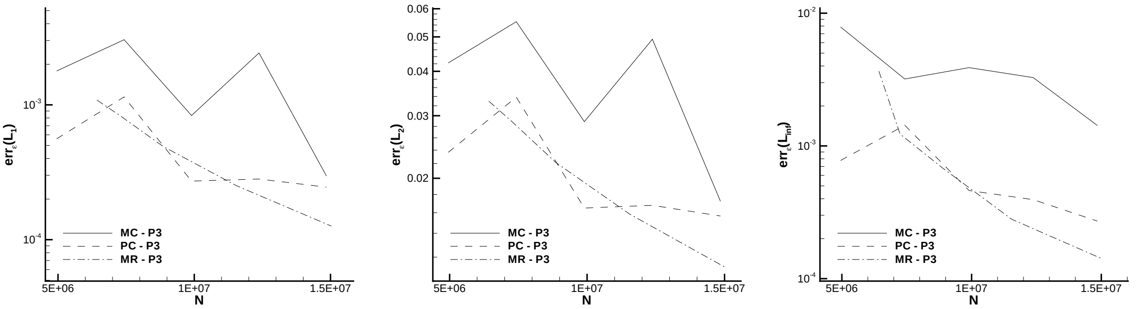
from a coarsest one, with a reduced number of function evaluations. As a consequence, only a reduced set of point values on the finest grid is evaluated, while the remaining set is obtained by interpolation (from the previous levels). This procedure moves recursively, with a combination of interpolation and evaluation, from the coarsest level to the finest and from each time step to the successive one. At each time step, the scheme allows to recover the solution on the finest level with a one time scheme that embeds the encoding and the truncation procedures of the classical Harten framework. Afterwards, this strategy is extended to the partial differential equation. A spatial discretization is chosen, as well as the time discretization, in order to solve the deterministic problem with a desired accuracy. The initial condition is discretized on the grid $D \times \Xi$ employing the finest resolution level in the stochastic space and the chosen spatial discretization. Then, the sTE strategy is applied to the associated stochastic space obtaining the MR representation of the solution at each time step, for each spatial node, *i.e.* the representation of the stochastic function obtained at a fixed physical space and time.

The sTE strategy is applied to some ”simplified” numerical test-cases and compared to classical stochastic methods. Finally, it is applied to the stochastic heat equation discretized by finite elements and integrated in time by means of a fourth order Runge-Kutta method. A discontinuous initial condition is considered. The sTE displays very promis-



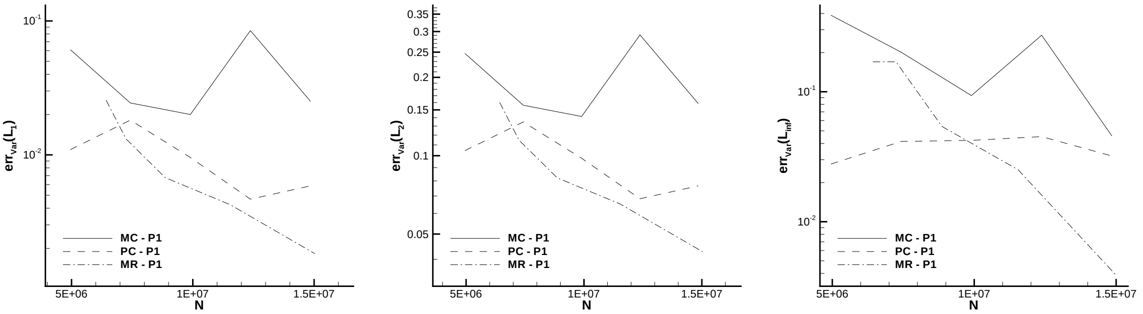
(a) (b) (c)

Figure 12: Error norms of the mean of the variable u, corresponding to the probe P2, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



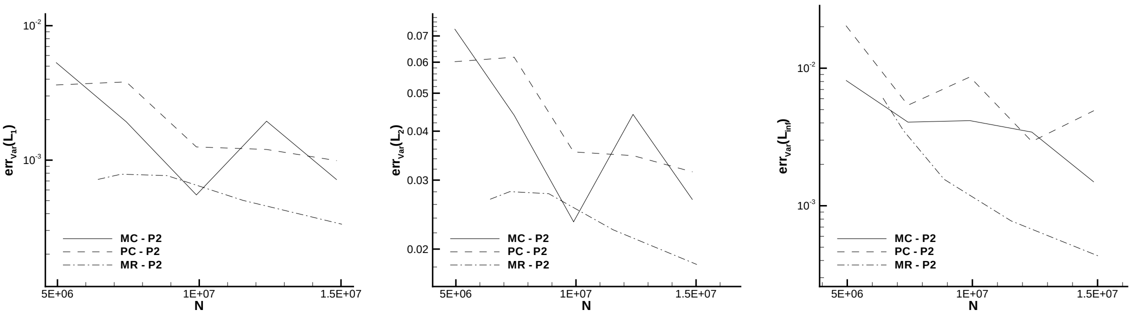
(a) (b) (c)

Figure 13: Error norms of the mean of the variable u, corresponding to the probe P3, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



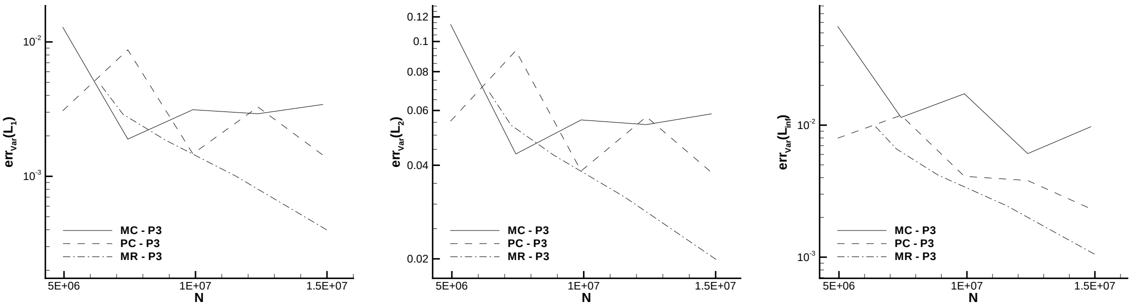
(a) (b) (c)

Figure 14: Error norms of the variance of the variable u, corresponding to the probe P1, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Figure 15: Error norms of the variance of the variable u, corresponding to the probe P2, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.



(a) (b) (c)

Figure 16: Error norms of the variance of the variable u, corresponding to the probe P3, for the 1D heat equation problem with uniform pdf in the L_1 (a), L_2 (b) and L_∞ (c) spaces.

ing results in terms of accuracy, convergence and regularity. Future works will focus on the extension of the present strategy to hyperbolic partial differential equations.

Acknowledgements

Remi Abgrall has been partially supported by the ERC Advanced Grant ADDECCO N. 226316, while Gianluca Geraci has been fully supported by the ERC Advanced Grant ADDECCO N. 226316.

Appendix. Finite element discretization of the 1D heat equation

In this section, we want to provide more details about the finite element scheme employed to solve the heat problem (9). For an exhaustive analysis of the problem, the reader may refer to [14].

Let us consider first, how to reduce the problem (9) from the weak form to the algebraic form (11). The weak formulation can be obtained by multiplying the equation (9), for the test function $v \in V$. As a consequence, the equation (10) is computed by integrating over the physical space

$$\int_D \frac{\partial u(x,t,\xi)}{\partial t} v \, \mathrm{d}x = \int_D \nu \frac{\partial^2 u(x,t,\xi)}{\partial x^2} v \, \mathrm{d}x. \qquad (.1)$$

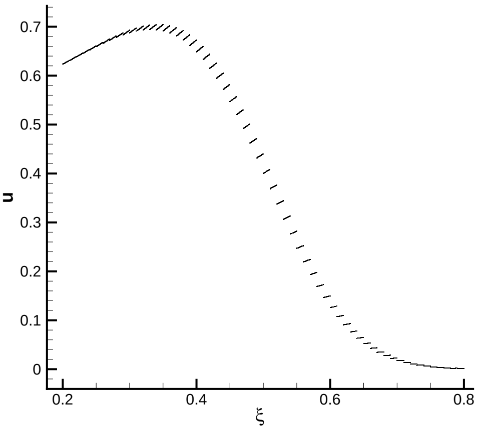


Figure 17: Solution $u(0.5, 0.5, \xi)$ corresponding to the probe P2 at the final time step obtained with 1000 MC samples.

The right-hand side can be decomposed using the integration by parts as follows

$$\begin{aligned}\int_D \nu \frac{\partial^2 u(x,t,\xi)}{\partial x^2} v \, \mathrm{d}x &= \int_D \nu \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x} v\right) \mathrm{d}x - \int \nu \frac{\partial u}{\partial x}\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x \\ &= \nu \left[\frac{\partial u}{\partial x} v\right]_0^1 - \int_D \nu \frac{\partial u}{\partial x}\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x \\ &= -\int_D \nu \frac{\partial u}{\partial x}\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x,\end{aligned} \tag{.2}$$

where the test functions are equal to zero at the boundaries ($v(0) = v(1) = 0$) and a Dirichlet boundary condition is applied. The integrals are well-posed if $v \in H_0^1(0,1)$:

$$H_0^1(0,1) = \left\{v \in H^1(0,1) : v(0) = v(1) = 0\right\}. \tag{.3}$$

The problem reduces to search $u \in V = H_0^1(0,1)$

$$\int_D \frac{\partial u(x,t,\xi)}{\partial t} v \, \mathrm{d}x = -\int_D \nu \frac{\partial u}{\partial x}\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x \quad \forall c \in V. \tag{.4}$$

The Galerkin approximation of the problem can be obtained by searching for an approximate solution $u_h \in V_h \subset V$, where the space V_h has a finite dimension N_h, with the test functions v in the same space V_h. In the latter case, *i.e.* using the same space for the solution and the tests functions, the approximation is the so-called Bubnov-Galerkin approximation.

If the Lagrangian finite element space is chosen for V_h, the classical basis $\{\phi_i\}$ (in the linear case) is equal to

$$\phi_i = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & x_i \le x \le x_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{.5}$$

If the tessellation of the domain is obtained with the nodes $\{x_i\}_{i=0}^{N+1}$, the solution can be expanded as a linear combination of the Lagrangian functions (for all the internal nodes) $u_h(x,t) = \sum_{i=1}^{N} u_i(t)\phi_i(x)$ that, inflated into (.4), becomes

$$\sum_{i=1}^{N}\int_{D_{ij}}\phi_i\phi_j \mathrm{d}x = -\nu\sum_{i=1}^{N}u_i(t)\int_{D_{ij}}\frac{\mathrm{d}\phi_i}{\mathrm{d}x}\frac{\mathrm{d}\phi_j}{\mathrm{d}x}\mathrm{d}x, \quad \forall\phi_j\in V_h \tag{.6}$$

where D_{ij} indicates the non-null support of the product function $\phi_i\phi_j$.

If a mass M and a stiffness A matrices are defined as follows

$$\begin{aligned} \mathrm{M} &= \left[m_{ij}\right], \quad m_{ij} = \int_{D_{ij}}\phi_i\phi_j\,\mathrm{d}x \\ \mathrm{A} &= \left[a_{ij}\right], \quad a_{ij} = \int_{D_{ij}}\frac{\mathrm{d}\phi_i}{\mathrm{d}x}\frac{\mathrm{d}\phi_j}{\mathrm{d}x}\mathrm{d}x, \end{aligned} \tag{.7}$$

the algebraic form (11) can be found as

$$\mathrm{M}\frac{\mathrm{d}\mathbf{U(t)}}{\mathrm{d}t} = -\nu\mathrm{A}\mathbf{U(t)}, \tag{.8}$$

where the vector $\mathbf{U}(t) = \{u_i(t),\dots,u_N(t)\}^{\mathrm{T}}\in\mathbb{R}^N$ is the collection of all the degrees of freedom of the linear expansion for $u_h(t)$. The original (9) parabolic partial differential problem is recast in a set of (coupled) ordinary differential equations.

Thanks to the compact support of the Lagrangian basis both M and A have a regular pattern of sparsity. In particular they are symmetric tridiagonal matrices. For recasting the system of ODEs in a set of decoupled ordinary differential equations, the mass matrix can be approximated by a diagonal matrix $\hat{\mathrm{M}}$, *i.e.* the so-called mass lumping technique. Thanks to the properties of the linear Lagrangian element ($\sum_{j=1}^{N}\phi_j = 1$) the mass matrix can be approximated by

$$\hat{\mathrm{M}} = \mathrm{Diag}(\hat{m}_{ii}) \quad \text{with} \quad \hat{m}_{ii} = \sum_{j=1}^{N}m_{ij} = \int_{D_{ij}}\phi_i\sum_{j=1}^{N}\phi_j\,\mathrm{d}x = \int_{D_{ij}}\phi_i\,\mathrm{d}x. \tag{.9}$$

The lumped matrix becomes $\hat{\mathrm{M}} = \mathrm{Diag}(5/6, 1,\dots,1,5/6)\Delta x$, while for the stiffness matrix we have

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}\frac{1}{\Delta x}, \tag{.10}$$

where we employed a uniform tessellation of the physical space $x_i = i\Delta x$ with $i = 0,\dots,N+1$ and $\Delta x = 1/N$.

The finite element formulation of the problem is obtained here in the physical 1D case only for the sake of simplicity, but both the deterministic scheme and the sTE strategy, described in 4.3, can be easily extended to physical (and even stochastic) multidimensional cases replacing the mass and stiffness matrices with their multidimensional counterparts and providing a nested set of meshes in Ξ.

References

[1] R. Abgrall, P.M. Congedo, C. Corre, S. Galéra, A simple semi-intrusive method for uncertainty quantification of shocked flows, comparison with a non-intrusive polynomial chaos method, in: V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010 J. C. F. Pereira and A. Sequeira (Eds) Lisbon, Portugal,14-17 June 2010, June, pp. 1–17.

[2] R. Abgrall, P.M. Congedo, S. Galéra, A semi-intrusive deterministic approach to uncertainty quantifications in non-linear fluid flow problems, Technical Report, INRIA RR-7820, 2011.

[3] R. Abgrall, P.M. Congedo, S. Galéra, G. Geraci, Semi-intrusive and non-intrusive stochastic methods for aerospace applications, in: 4TH EUROPEAN CONFERENCE FOR AEROSPACE SCIENCES, Saint Petersburg, Russia, July 4th-8th, 2011, 1, pp. 1–8.

[4] R. Abgrall, P.M. Congedo, G. Geraci, A One-Time Truncate and Encode Multiresolution Stochastic Framework, Technical Report, INRIA Bordeaux–Sud-Ouest, 2012.

[5] R. Abgrall, A. Harten, Multiresolution Representation in Unstructured Meshes, SIAM Journal on Numerical Analysis 35 (1998) 2128–2146.

[6] J. Foo, G.E. Karniadakis, Multi-element probabilistic collocation method in high dimensions, Journal of Computational Physics 229 (2010) 1536–1557.

[7] M. Gerritsma, J.B. van der Steen, P. Vos, G.E. Karniadakis, Time-dependent generalized polynomial chaos, Journal of Computational Physics 229 (2010) 8333–8363.

[8] A. Harten, Adaptive multiresolution schemes for shock computations, Journal of Computational Physics 135 (1994) 260–278.

[9] A. Harten, Multiresolution algorithms for the numerical solution of hyperbolic conservation laws, Communications on Pure and Applied Mathematics 48 (1995) 1305–1342.

[10] A. Harten, Multiresolution Representation of Data : A General Framework, SIAM Journal on Numerical Analysis 33 (1996) 1205–1256.

[11] O. Le Maître, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, Journal of Computational Physics 197 (2004) 502–531.

[12] O. Le Maître, Uncertainty propagation using WienerHaar expansions, Journal of Computational Physics 197 (2004) 28–57.

[13] O. Le Maître, O. Knio, Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics, Springer Verlag, 2010.

[14] A. Quarteroni, Modellistica numerica per problemi differenziali, Springer, 2008.

[15] A. Quarteroni, R. Sacco, F. Saleri, Matematica Numerica, Springer, 2008.

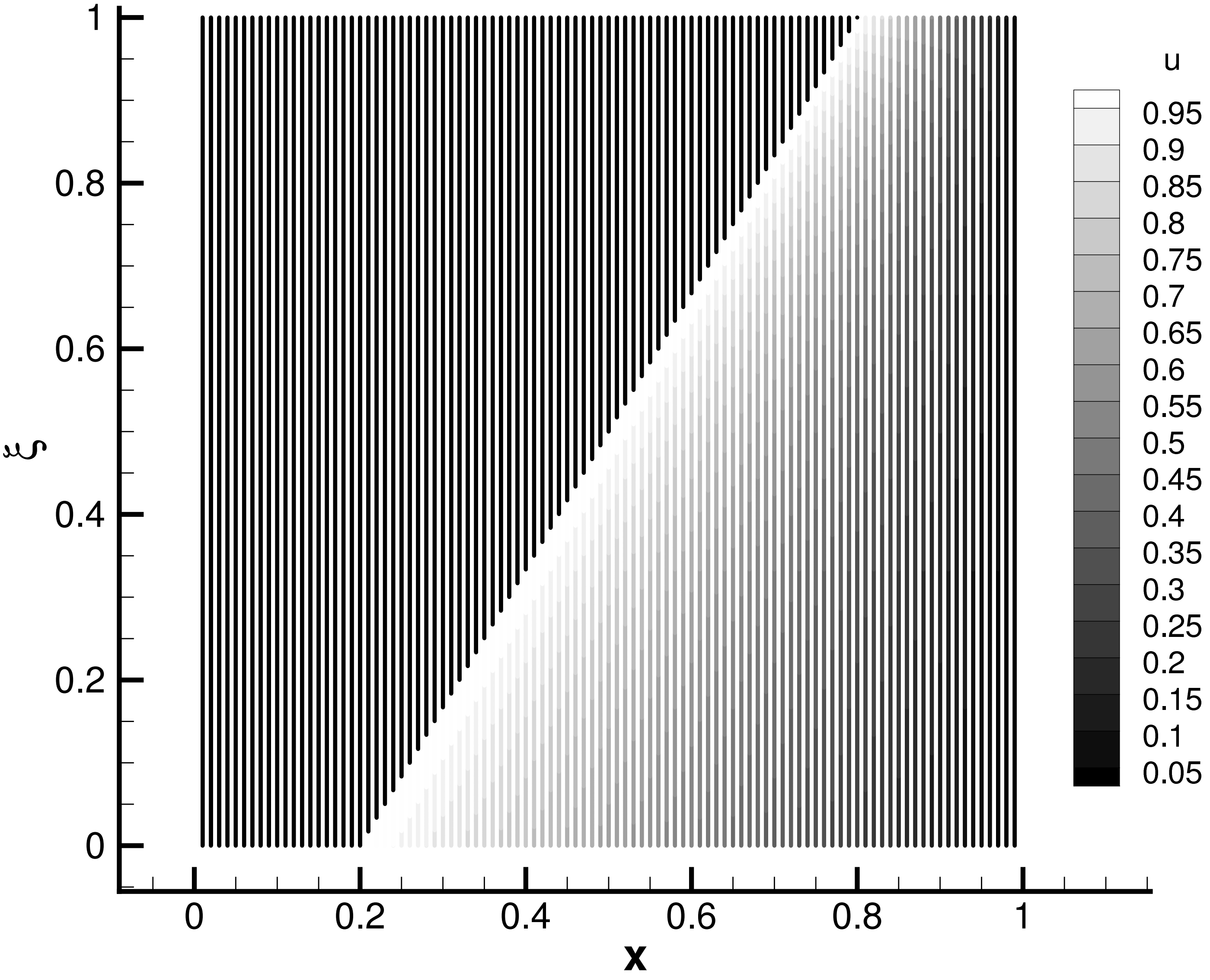
[16] J. Tryoen, O. Le Maître, M. Ndjinga, A. Ern, Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems q, Journal of Computational Physics 229 (2010) 6485–6511.

[17] X. Wan, G.E. Karniadakis, Long-term behavior of polynomial chaos in stochastic flow simulations, Computer Methods in Applied Mechanics and Engineering 195 (2006) 5582–5596.

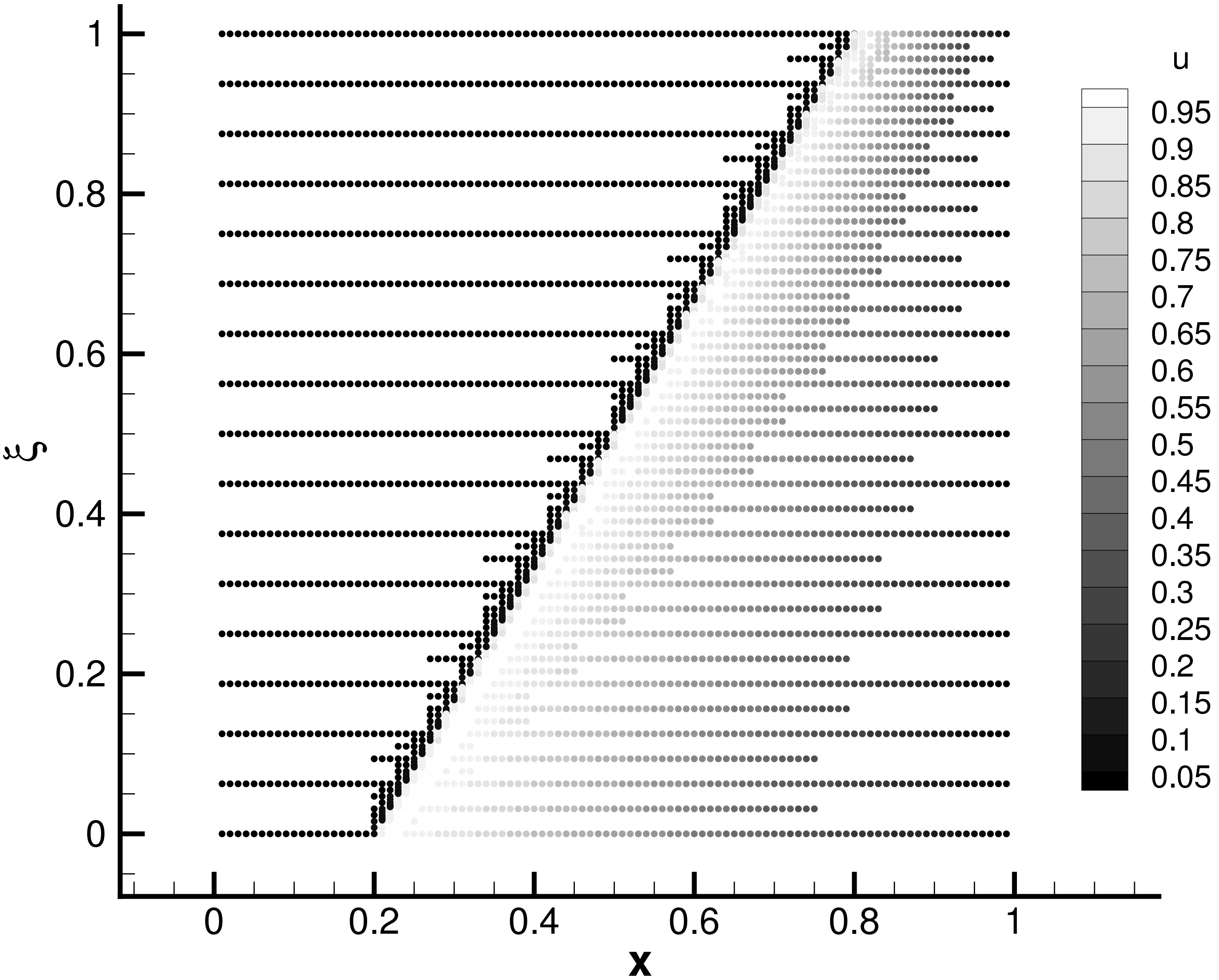
[18] N. Wiener, The Homogeneous Chaos, American Journal of Mathematics 60 (1938) 897–936.

[19] D. Xiu, G.E. Karniadakis, Modeling uncertainty in flow simulations via generalized polynomial chaos, Journal of Computational Physics 187 (2003) 137–167.

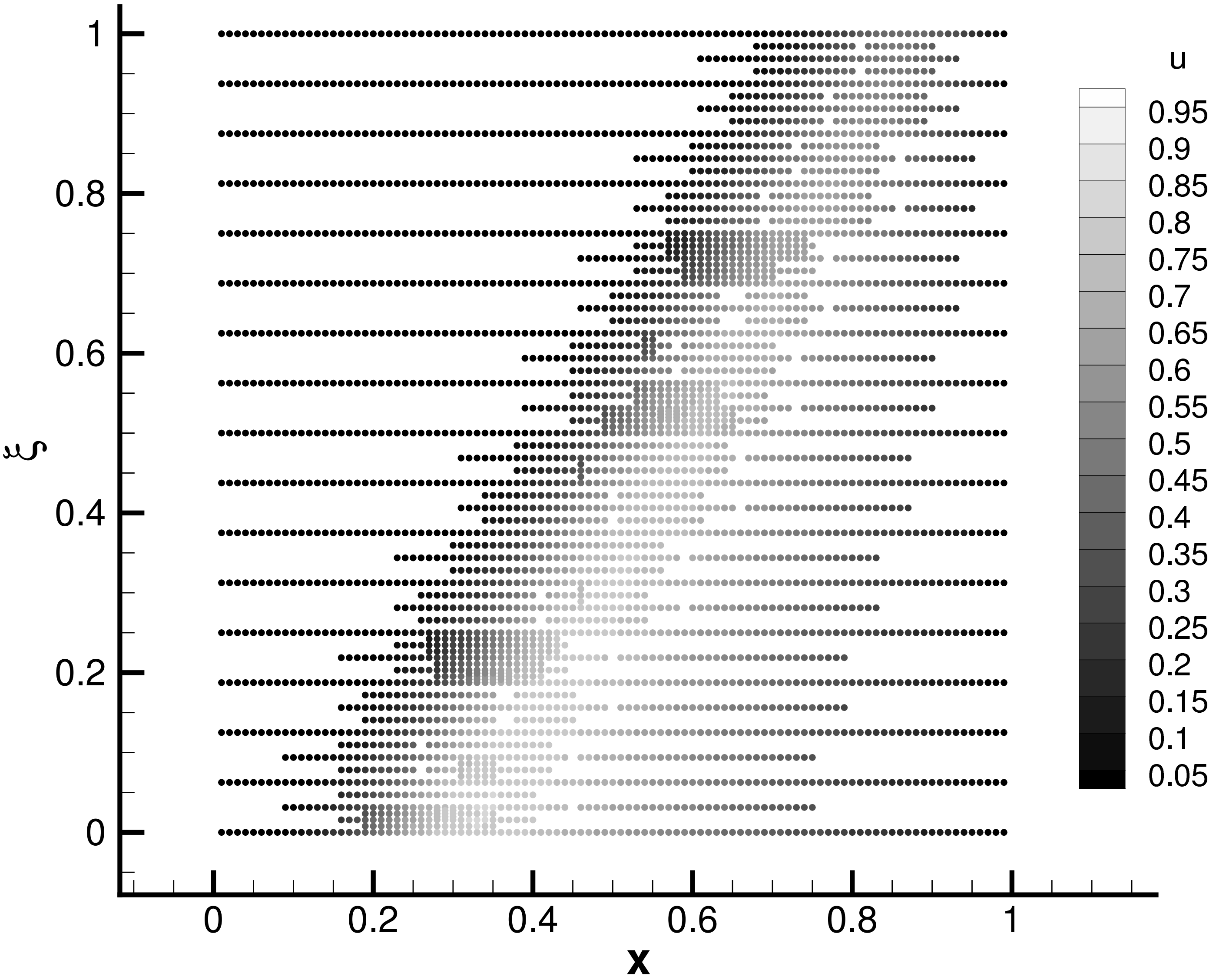
initial_grey.eps (grey version of initial.eps)



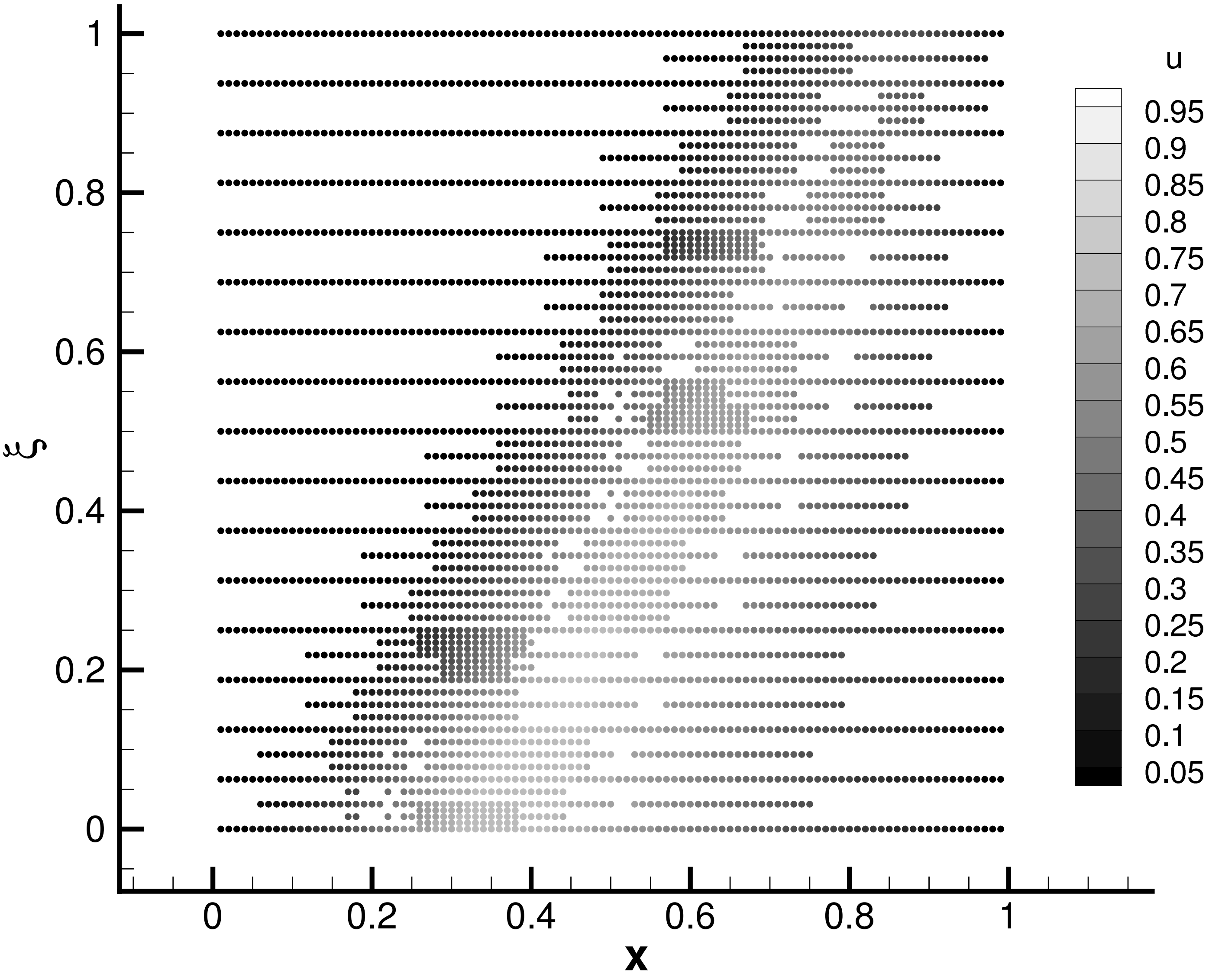
first_step_grey.eps (grey version of first_step.eps)



250_t_grey.eps (grey version of 250_t.eps)



500_t_grey.eps (grey version of 500_t.eps)



Paper *P3*

Journal of Scientific Computing manuscript No.
(will be inserted by the editor)

A novel weakly-intrusive non-linear multiresolution framework for uncertainty quantification in computational fluid dynamics

Pietro Marco Congedo · **Gianluca Geraci** · **Rémi Abgrall** · **Gianluca Iaccarino**

Received: date / Accepted: date

Abstract In this paper, the Harten multi-resolution framework is generalized and extended for taking into account uncertainties into partial stochastic differential equations (sPDE). Innovative ingredients are given by an algorithm permitting to recover the multiresolution representation without requiring the fully resolved solution, the possibility to treat a whatever form of pdf, the use of high-order reconstruction in the stochastic space. Moreover, the sTE method is introduced, that is a weak coupling in spatial/stochastic space for minimizing the computational cost in sPDE and is particularly attractive when treating discontinuities (such as shock in compressible flows). Flexibility of the proposed method is demonstrated by proposing a simple algorithm coupling together high-resolution schemes in the physical and in the stochastic spaces at the same time. Various numerical experiences are performed (algebraic functions and ordinary differential equations), including stochastic partial differential equations. Then, efficiency of the proposed strategy for solving stochastic linear advection and Burgers equations is estimated by comparing with some classical techniques, such as Monte Carlo or Polynomial Chaos method.

Keywords Multiresolution · Uncertainty Quantification (UQ) · Adaptive grid · Hyperbolic conservation laws · Shock tube.

1 Introduction and motivation

Handling uncertain operating conditions, material properties and manufacturing tolerances poses a tremendous challenge to the scientific computing community. In the last year, a great effort has been devoted to the propagation of uncertainties through numerical codes. Two complementary philosophies are actually employed: a non-intrusive propagation approach that consists in using the numerical code as a black box with several and indipendent calls to the code; an intrusive approach that requires the modification of the numerical code to include the propagation of the uncertainties. For the non-intrusive side, different methods are nowadays commonly employed: Monte Carlo family of techniques [13], the collocation family [9] and the non-intrusive Galerkin projection methods. This last family of methods has been introduced for the first time by Ganem and Spanos [12] for the analysis of structural dynamics systems and has been generalized by Xiu and Karniadakis

P.M. Congedo, G. Geraci & R. Abgrall
INRIA Bordeaux–Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talence Cedex, France.
E-mail: pietro.congedo@inria.fr, gianluca.geraci@inria.fr, remi.abgrall@inria.fr

G. Iaccarino
Center for Turbulence Research, Stanford University, Building 500, CA 94305-3035, Stanford, USA
E-mail: jops@stanford.edu

[29] to general probability distributions. Actually, the non-intrusive Galerkin projection represents the state-of-the art of the stochastic analysis for systems with a smooth response surface due to its spectral convergence property.

The Galerkin projection is also the most important technique in order to manage intrusively the uncertainty propagation into a numerical code. In practice, this means that it is possible to obtain an equivalent set of governing equations for the coefficients of a truncated polynomial representation of the quantities of interest [20]. Then, the number of equations is related to the number of coefficients employed in the polynomial expansion, and the numerical code should be deeply modified. In many cases, this leads to complex problems regarding the generality of the approach when *ad hoc* solvers are proposed [27]. More recently, Abgrall and Congedo proposed a novel semi-intrusive approach that extend in a straightforward and natural way the representation of the variables in the physical space also along the stochastic space [1]. This approach leads to a very flexible scheme able to handle whatever form of probability density function even time varying and discontinuous. One of the prominent advantage of this kind of approach is the possibility to extend in an easier way an existing deterministic code to its stochastic counterparts.

Thanks to its intrinsic capability to manage discontinuous responses, the semi-intrusive methods represents a promising alternative to the Galerkin projection techniques for all the applications in which the system is dominated by shocks. A prominent class of application of the semi-intrusive scheme could be, for instance, the computational fluid dynamics for transonic flows. However, actually every class of UQ method suffers from the so-called *curse of dimensionality* [10], *i.e.* the exponential growth of the number of degree of freedom required, associated to the increase of the number of uncertain parameters. In order to propose a cure for the semi-intrusive scheme capable to tackle the curse of dimensionality, in this work we introduce a general multiresolution framework to reduce the number of degrees of freedom necessary for the representation of data along the stochastic space.

Many works focused on multiresolution techniques to reduce the computational cost associated to UQ simulations. In [20], a multiresolution basis is employed to represent the solution of a partial differential equations after fixing the physical coordinate. This representation is very efficient but limited to the case in which the stochastic representation is used at a fixed physical location. To overcome this issue, more recently, Tryoen et al. introduced in [27] a multiresolution wavelets representation in the context of intrusive Galerkin projections. However, the Galerkin approach presented remains very problem-dependent. In fact, using a Roe-type solver requires to know the eigenstructure of the Roe matrix explicitly; this can be very complex. Moreover, *ad hoc* entropy fix should be adopted, thus increasing the numerical cost associated to the representation of discontinuous solution [25]. This original approach has been further improved to obtain a more efficient scheme employing a multiresolution adaptive strategy [26]. However, this approach is limited by the spatial and time discretization accuracy (only first order) that could dominate the overall accuracy.

This work is heavily inspired by the classical Harten's multiresolution framework in the basic principles, but a series of original elements are introduced to make adequate and efficient this technique for UQ purpose. First, there is the introduction of an algorithm capable to recover the multiresolution representation without requiring the fully resolved solution (allowing to consider discontinuous time-varying probability distribution functions). Secondly, the probability distribution function is used as a key element for determining the quality of the representation of the solution. Other main advantages are the following: handling easily a whatever form of pdf, even non classical; treating in a straightforward way correlated uncertainties without solving specific equations for obtaining the wavelet basis as should be required, for example, in a classical wavelet framework as [26]. This last point is even more important when dealing with general geometries in the stochastic space. In fact, the Harten framework allows the extension to general meshes, while the pure wavelets approach via the solution of a dilating equation could be not feasible in a general case. Finally, the general framework presented in this paper allows an easy extension to high-order representations both in physical and time spaces (without any kind of modification of the deterministic scheme), thus proposing a general method for building efficient intrusive scheme in complex applications.

This paper is organized as follows. In section 2, the mathematical setting for the stochastic differential equation is given. Section 3, illustrates the multi-resolution framework of Harten, generalized for the stochastic space. In particular, the following elements are detailed: cell-average and point-values settings, the truncation and stability requirements, the non-linear approach with the introduction of ENO reconstructions. In section 4, the weak coupled spatial/stochastic scheme is presented, extending the basic Truncate and Encode (TE) Algorithm to the spatial-TE (sTE) for treating partial stochastic differential equations. Various results are provided in Section 5. Some algebraic and ordinary differential equation, both scalar and vectorial, are presented and compared with classical Monte Carlo methods and, where possible with gPC non intrusive approach. The challenging monodimensional Kraichnan-Orszag problem is also reported in the classical case of uniform distribution. Then, some partial stochastic differential equations are considered, *i.e.* the advection and the Burgers equations. Some concluding remarks and perspectives have been reported in section 6.

2 Mathematical setting for a stochastic differential problem

Let us consider the following problem for an output of interest $u(\mathbf{x},t,\xi(\omega))^1$:

$$\mathscr{L}(\mathbf{x},t,\xi(\omega);u(\mathbf{x},t,\xi(\omega)))=\mathscr{S}(\mathbf{x},t,\xi(\omega)), \tag{1}$$

where the operator $\mathscr{L}$ can be either an algebraic or a differential operator (in this case we need appropriate initial and boundary conditions). The operator $\mathscr{L}$ and the source term $\mathscr{S}$ are defined on the domain $\Omega\times T\times\Xi$, where $\mathbf{x}\in\Omega\subset\mathbb{R}^{n_d}$, with $n_d\in\{1,2,3\}$, and $t\in T$ are the spatial and temporal dimensions. Let us consider a measurable space (Ξ,Σ,p) where Ξ is the sample space, Σ is its $\sigma-$algebra of events and p a probability measure with the following properties:

- $p(A)\geq 0$ for all $A\in\Sigma$;
- Countable additivity: if $A_i\in\Sigma$ are disjoint sets then $p(\bigcup_i A_i)=\sum_i p(A_i)$;
- as probability measure p is normalized on Ξ: $p(\Xi)=1$.

The $\mathbb{R}^d-$valued random variable ξ specifies a set of events with a corresponding probability. More formally, the random variable ξ is a measurable function that maps the measurable space (Ξ,Σ,p) to another measurable space, *i.e.* the Borel $\mathscr{B}^d$ $\sigma-$algebra of the real space $(\mathbb{R}^d,\mathscr{B}^d,\mathbb{P})$. A set of events ω exists that ξ maps to an output event $A\in\mathscr{B}^d$ with the probability of occurrence of A, where $\mathbb{P}(A)$ is equal to the probability of ω:

$$\mathbb{P}(A)=p(\xi^{-1}(A))=p(\omega:\xi(\omega)\in A). \tag{2}$$

As usual in literature, it is possible to write only $p(\xi\in A)$ and not the induced probability $\mathbb{P}(A)$.

The probability density function $p(\xi)$ is defined as a joint probability density function from the independent probability function of each variable: $p(\xi)=\prod_{i=1}^{d}p_i(\xi_i)$. This assumption allows an independent representation for each direction in the probabilistic space with the possibility to recover the multidimensional representation by tensorization. The aim of UQ is to find the statistical moments of the solution $u(\mathbf{x},t,\xi)$.

For instance, assuming $u(\xi)\in L_2(\Xi,p)$, the following first two central moments can be always defined

$$\begin{aligned}\mathbb{E}(u,\mathbf{x},t)&=\int_\Xi u(\mathbf{x},t,\xi)p(\xi)\mathrm{d}\xi\\ \mathbb{V}\mathrm{ar}(u,\mathbf{x},t)&=\int_\Xi\left(u(\mathbf{x},t,\xi)-\mathbb{E}(u)\right)^2p(\xi)\mathrm{d}\xi.\end{aligned} \tag{3}$$

3 The Generalized Harten MultiResolution approach

In this section, the generalized multiresolution framework of Harten [15–17,4] is briefly recalled. The basic principles are first defined. Then, the whole set of operators are formulated and extended in the UQ framework. They are described in §3.1 and §3.2 in cell-average and point-value, settings.

[1] In the following the exposition is made for a scalar output variable (u) for brevity, but the extension to the multidimensional output case is straightforward

The Harten framework can be considered, as pointed out by Aràndiga and Donat in [7], as a rearrangement of the information in a set of discrete data representing different resolution levels. This approach suits the general scope of a compact representation of the data allowing to neglect information relative to non significative level of resolution.

The building blocks of the framework are the operators of discretization $\mathscr{D}_k$, that allows the transfer of information from the continuous to the discrete space and the reconstruction operator $\mathscr{R}_k$ that performs the inverse operation. Note that these operators need to satisfy certain properties as well as a consistency relation. The multiresolution can be seen as a pyramidal rearrangement of the data making use of a direct transfer of information between different resolution levels. Using both operators of discretization $\mathscr{D}_k$ and reconstruction $\mathscr{R}_k$, the discrete operators of decimation D_k^{k-1} and prediction P_{k-1}^k between consecutive levels of resolutions k (higher resolution) and $k-1$ (lower resolution) can be defined.

More formally, let us start by considering a function $f \in \mathscr{F}$, where $\mathscr{F}$ is a proper space of functions. Let us consider a set of discrete operators of discretization $\{\mathscr{D}_k\}_{k=0}^L$ each of them defined of a vectorial space of finite dimension

$$\mathscr{D}_k : \quad \mathscr{F} \to V_k \quad \text{with} \quad \dim(V_{k+1}) > \dim(V_k) = J_k. \tag{4}$$

The sequence $\{\mathscr{D}_k\}_{k=0}^L$ is nested according to the following properties:

- $\mathscr{D}_k$ is onto
- the null space of each level includes the null space associated to the previous resolution level $\mathscr{N}(\mathscr{D}_k) \subset \mathscr{N}(\mathscr{D}_{k+1})$.

These properties reflect in the following relation between discretization operators

$$\mathscr{D}_{k+1}(f) = 0 \Rightarrow \mathscr{D}_k(f) = 0 \quad \forall f \in \mathscr{F}. \tag{5}$$

Thanks to the onto property of each operator $\mathscr{D}_k$, the reconstruction operator $\mathscr{R}_k$ can be defined as follows

$$\mathscr{R}_k : \quad V_k \to \mathscr{F}. \tag{6}$$

The reconstruction operator is not required to be linear and this makes the Harten's multiresolution more general with respect to the wavelets framework [11].

Both operators $\mathscr{D}_k$ and $\mathscr{R}_k$ need to satisfy the following consistency relationship

$$(\mathscr{D}_k\mathscr{R}_k)(v) = v \quad \forall v \in V_k, \tag{7}$$

thus implying $\mathscr{D}_k\mathscr{R}_k = \mathrm{I}_k$ where I_k is the identity operator on V_k.

In the framework of nested sequences whose elements are defined in (4), the decimation operator D_k^{k-1} can be defined as a linear mapping between V_k onto V_{k-1}:

$$\mathrm{D}_k^{k-1} : \quad V_k \to V_{k-1}, \tag{8}$$

where

$$\mathrm{D}_k^{k-1} v^k = \mathscr{D}_{k-1} f \in V_{k-1} \quad \forall v^k = \mathscr{D}_k f \in V_k. \tag{9}$$

The decimation operator is used to generate recursively the set of discrete data from the highest resolution level ($k = L$) to the lowest ($k = 0$) $\{v^k\}_{k=0}^{L-1}$

$$v^{k-1} = \mathrm{D}_k^{k-1} v^k \quad \forall k = L, L-1, \dots, 1. \tag{10}$$

Inversely, the prediction P_{k-1}^k allows to approximate the set of data v^k from v^{k-1}

$$v^k = \mathscr{D}_k f \approx \mathscr{D}_k(\mathscr{R}_{k-1} v^{k-1}). \tag{11}$$

This leads to the definition of the prediction operator P_{k-1}^k between discrete data on successive resolution levels as

$$\mathrm{P}_{k-1}^k \overset{\text{def}}{=} \mathscr{D}_k\mathscr{R}_{k-1} : \quad V^{k-1} \to V^k. \tag{12}$$

A consistency properties can be found between the discrete operators, *i.e.* $\mathrm{D}_k^{k-1}\mathrm{P}_{k-1}^k = \mathrm{I}_k$, following from

$$v^{k-1} = \mathrm{D}_k^{k-1} v^k = \mathrm{D}_k^{k-1} \mathscr{D}_k f = \mathrm{D}_k^{k-1} \mathscr{D}_k \mathscr{R}_{k-1} v^{k-1} = \mathrm{D}_k^{k-1} \mathrm{P}_{k-1}^k v^{k-1}. \tag{13}$$

Now, an error prediction in the MR framework, e^k, can be defined as

$$e^k \overset{\text{def}}{=} v^k - \mathrm{P}_{k-1}^k v^{k-1} = (\mathrm{I}_k - \mathrm{P}_{k-1}^k \mathrm{D}_k^{k-1}) v^k. \tag{14}$$

The prediction error satisfies (from the consistency property (13))

$$\mathrm{D}_k^{k-1} e^k = \mathrm{D}_k^{k-1}(v^k - \mathrm{P}_{k-1}^k v^{k-1}) = v^{k-1} - v^{k-1} = 0, \tag{15}$$

then it is in the null space of the decimation operator $e^k \in \mathscr{N}(\mathrm{D}_k^{k-1})$. Remembering the definition (8) and applying the rank theorem, it is possible to find that

$$\dim(V_k) = \dim(\mathscr{N}(\mathrm{D}_k^{k-1})) + \dim(V_{k-1}) \to \dim(\mathscr{N}(\mathrm{D}_k^{k-1})) = \dim(V_k) - \dim(V_{k-1}) = J_k - J_{k-1}. \tag{16}$$

The linear independent coordinates of e^k are called wavelets or details d^k. Two operators can be defined to link the prediction error to the details, E^k and G^k, as follows

$$e^k \overset{\text{def}}{=} E^k d^k, \quad d^k \overset{\text{def}}{=} G^k e^k \quad \text{with} \quad E^k G^k : V^k \to \mathscr{N}(\mathrm{D}_k^{k-1}). \tag{17}$$

Using all the operators described in this section, a multiresolution representation of data can be defined.

This is obtained by two procedure: the *encoding* and the *decoding*. The *encoding* moves from the highest resolution level to the lowest one applying recursively (for all $k = L, \ldots, 1$) the decimation operator and computing the details

$$\begin{cases} v^{k-1} = \mathrm{D}_k^{k-1} v^k \\ d^k = G_k(\mathrm{I}_k - \mathrm{P}_{k-1}^k \mathrm{D}_k^{k-1}) v^k. \end{cases} \tag{18}$$

The multiresolution representation v_{MR} refers to the possibility to obtain a one-to-one correspondence between the highest resolution level v^L and the sequence of the details d^k in addition to the lowest resolution level v^0:

$$v_{\mathrm{MR}} \overset{\text{def}}{=} \{v^0, d^1, \ldots, d^L\}. \tag{19}$$

The *decoding* procedure is the dual procedure with respect to the *encoding*: recursively moves from the lowest resolution level v^0 together with the prediction error e^k for all the levels $k = 1, \ldots, L$

$$v^k = \mathrm{P}_{k-1}^k v^{k-1} + E_k d^k. \tag{20}$$

Ideally, *decoding* and *encoding* permit an ideal exchange of information among different resolution levels. In order to be useful, these operations are coupled with an operator of data truncation. This additive operator allows, under a certain tolerance, to eliminate the over abundant information. The compression capability opens several possibilities to the application of the multiresolution framework to compress the data as, for instance, in the signal/image representation schemes [6] or as a fundamental brick in the solution of intrinsically multi scales problems, as demonstrated already in the first seminal works of Harten [16,17]. In this paper, focused on stochastic partial differential equations, the truncation procedure is even more important. In fact, the prediction error is no more an estimation *a posteriori* of the prediction operation, but an indication *a priori* of the possibility to predict the solution. In the following, the multiresolution framework is extended to the partial differential equation solved by conservation methods, in the cell-average and the point-value settings. The approach proposed in the present work remains anyway a weak-coupling approach between the physical and stochastic space. This should be interpreted as follows. In the present paper the multiresolution is employed only for the representation of data in the stochastic space, while the physical space is not affected by it. This approach reflects the situation in which the discretizations between the two spaces remain independent. However Abgrall et al. introduced a novel method for UQ in which the physical and stochastic space are considered as a unique space [2,1]. In a such context the multiresolution of data could be applied not only on the stochastic dimension, but also for the discrete representation of data in the overall physical/stochastic space. The need for a compact representation of data along the stochastic space makes very attractive and natural the choice of the probability density function as weight function for the cell average setting. For these reasons, more general averaging procedures as the hat-based discretizations [11,8] are not investigated in this work (see [11], in particular, for a generalization of the discretization procedure based on averaging).

3.1 Cell average framework

Consider a function $f = f(\xi)$, $f : \Xi \subset \mathbb{R}^d \to \mathbb{R}$ with d the number of dimensions of the stochastic problem, *i.e.* the number of uncertain parameters. The functional space $\mathscr{F}$ is defined as $\mathscr{F} = L^2(\Xi)$ because of the cell-average setting, and the need to have functions with finite variance in the UQ framework. Let us consider the space equipped with a measure

$$\mathrm{d}\mu(\xi) = p(\xi)\mathrm{d}\xi, \tag{21}$$

where $\xi \in \Xi \subset \mathbb{R}^d$ is the vector of uncertain parameters and $p(\xi)$ is the associated probability density function. Let consider a tessellation, of the stochastic space Ξ, satisfying the classical non overlapping requirements

$$\Xi = \bigcup_{j=1}^{N_\xi} \Xi_j, \quad \text{with} \quad \Xi_i \cap \Xi_j = 0 \quad \text{if} \quad i \neq j. \tag{22}$$

A similar approach, reminiscent of a classical finite volume approach in the physical space [1], allow a natural treatment of even unbounded stochastic space Ξ. This is accomplished using the measure $\mathrm{d}\mu$ which is always bounded holding the following relation for the probability function:

$$\int_\Xi p(\xi)\mathrm{d}\xi = \int_\Xi \mathrm{d}\mu(\xi) = \sum_j \mu(\Xi_j) = 1, \tag{23}$$

where the additivity of the integrals is applied considering also the measure of each cell through

$$\mu(\Xi_j) = \int_{\Xi_j} \mathrm{d}\mu(\xi). \tag{24}$$

In this setting, the discretization operator on the k-th level can be defined over the j-th cell Ξ_j^k as

$$(\mathscr{D}_k f)_j \overset{\text{def}}{=} \frac{1}{\mu(\Xi_j^k)} \int_{\Xi_j^k} f(\xi)\mathrm{d}\mu(\xi) = v_j^k. \tag{25}$$

By an agglomeration (splitting) procedure with a generic mesh, even unstructured, it is always possible to obtain a less (higher) resolution level. In a general case, to each cell Ξ_j^k at the lower resolution level corresponds a certain number of cell at the higher resolution level. In order to preserve the nested character between levels, the following properties between meshes must hold:

$$\Xi_j^k = \sum_l^{\bar{l_c}} \Xi_l^{k+1}. \tag{26}$$

Let us consider, for example, a 1D case of equally splitted cells between levels in the case of regular nested meshes (the number of cells of the level $k+1$ impinging at the lower level k is fixed to two ($\bar{l_c}$)).

In this case, the decimation operator could be obtained as follows (see figure 1)

$$\begin{aligned}(\mathrm{D}_k^{k-1} v^k)_j = (\mathrm{D}_k^{k-1}\mathscr{D}_k f)_j = (\mathscr{D}_{k-1} f)_j &= \frac{1}{\mu(\Xi_j^{k-1})} \int_{\Xi_j^{k-1}} f(\xi)\mathrm{d}\mu(\xi) \\ &= \frac{1}{\mu(\Xi_j^{k-1})} \Big(\mu(\Xi_{2j}^k)(\mathscr{D}_k f)_{2j} + \mu(\Xi_{2j-1}^k)(\mathscr{D}_k f)_{2j-1} \Big).\end{aligned} \tag{27}$$

Note that in the general case of an arbitrary PDF $p(\xi)$, even the 1D case sketched in figure 1, the nested sequence produces nested relations with non constant coefficients even for the same level of resolution depending from the measure. To recover the counterparts of the physical space case, the splitting/agglomeration of each cell, based on a Lebesgue measure, should be replaced with a splitting based directly on the probability measure. The nested sequence of the meshes, even in this case, is totally independent from the function and can be generated *a priori* with the only requirement to know the probability distribution $p(\xi)$.

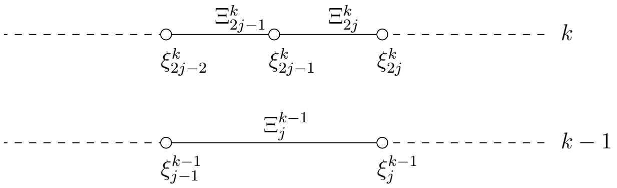


Fig. 1 Example of 1D stochastic nested meshes for the cell-average setting decimation procedure.

However, in this paper uniform PDFs are employed for the numerical test cases, then the two criterion, the splitting/agglomeration based on the Lebesgue or probability measure, are the same. In particular, the corresponding discretization operator can be defined by the convolution of the function f with an Haar scaling function. This approach constitutes a direct link between the Harten framework and the wavelets one.

The reconstruction operator $\mathscr{R}_k$ for the cell average setting has been introduced originally by Harten in the 1D case, using a reconstruction via primitive function. This means that the cell averaged function is replaced by a point valued function that corresponds to its primitive in the nodes of the mesh. This approach, despite its simplicity in the 1D case, results to be difficult to apply in the multidimensional case if the meshes are not structured. To overcome this issue Abgrall and Sonar generalized the reconstruction procedure in [5] even for multidimensional problems on unstructured meshes [4]. Fixed a polynomial degree of reconstruction r, a stencil $\mathscr{S}^k_j$ of cells with cardinality $s = s(r) = \text{card}(\mathscr{S}^k_j)$ can be fixed. On each stencil $\mathscr{S}^k_j$, a polynomial $\mathscr{P}^k_j(\xi; f)$ of degree r can be constructed. The admissibility of this kind of stencil remains subject to a Vandermonde type condition (see [5] for further details). Here, supposing the stencils admissible, the conditions to satisfy for the computation of $\mathscr{P}^k_j$ is the following

$$\mathscr{D}_k(\mathscr{P}^k_j(\xi; f))_l = \mathscr{D}_k(f)_l, \quad \forall l \in \mathscr{S}^k_j. \tag{28}$$

Further details on the choice of a proper polynomial form in the 1D case is reported in section §3.4. The reconstruction operator $\mathscr{R}_k$ is exactly equal to the union of all the polynomials $\mathscr{P}^k_j$ defined on all the cells Ξ^k_j. This makes possible, without introducing confusion, to change $\mathscr{R}_k$ with $\mathscr{P}^k_j$ when the cell Ξ^k_j is of interest.

The prediction operator P^k_{k-1} is obtained following its definition (12), using first the reconstruction procedure (28) for the level $k-1$, and then applying the discretization operator $\mathscr{D}_k(\mathscr{P}^{k-1}_j)$ relative to the level k.

The remaining step is to define a relation between the error e^k and its, linear independent, coordinates d^k (the wavelets). In the general case (26), a number of $\overline{l_c} - 1$ linear dependent relations for the components of e^k must hold. If the case of the dyadic splitting ($\overline{l_c} = 2$) is addressed, then for each cell only one component of d^k is present. Referring directly to the figure 1, on each cell Ξ^{k-1}_j, the value of $(\mathscr{D}_k f)_{2j-1}$ can be obtained via the prediction operator

$$(\mathscr{D}_k f)_{2j} = (\mathrm{P}^k_{k-1} v^{k-1})_{2j} = \frac{1}{\mu(\Xi^k_{2j})} \int_{\Xi^k_{2j}} \mathscr{P}^{k-1}_j \mathrm{d}\mu, \tag{29}$$

where the polynomial $\mathscr{P}^{k-1}_j$ is obtained according to (28). The last relation (29) permits to write

$$\begin{cases} d^k_{2j} = v^k_{2j} - (\mathrm{P}^k_{k-1} v^{k-1})_{2j} \\ v^k_{2j-1} = \dfrac{1}{\mu(\Xi^k_{2j-1})} \left(\mu(\Xi^{k-1}_j) v^{k-1}_j - \mu(\Xi^k_{2j}) v^k_{2j} \right). \end{cases} \tag{30}$$

Relations (30) are useful both in the *encoding* and *decoding* procedures. The cell average framework described in this section permits a direct link between this representation and a numerical scheme for conservations laws. Morevoer, other techniques, such as the sub-cell resolution (SR), introduced by Harten and co-workers [14], for representing a discontinuous response to the cell interfaces allowing to resolving the discontinuity within the cell that contains it, could be easily applied in this context. However, in this paper, the point value setting best suits the scope of weak coupling between the physical space and the stochastic one, since a best exploitation of the information in the only stochastic space is possible. Essentially, in the physical space, it is common to deal with spatially cell averaged value (as in finite volume schemes) while their interpretation in the stochastic space is more commonly a natural point value setting. The approach of a combined finite volume discretization for both the spatial and stochastic space, has been proposed by Abgrall and co-workers in [1,2]. The natural extension of the cell average framework, here described, is its application to the so called *semi intrusive* method of Abgrall, that, however, is left to further works.

3.2 Point-value framework

The point value setting represents since its introduction by Harten in [16] a very flexible setting to build numerical scheme with a multiresolution rearrangement of the information despite its simplicity. Conceptually the point value setting is the most natural dealing with only discretized data.

Let us consider bounded functions $f \in \mathscr{F} = \mathscr{B}(\Xi)$ and a number of stochastic parameters equal to d

$$f : \Xi \subset \mathbb{R}^d \to \mathbb{R} \tag{31}$$

where Ξ must be intended bounded with respect the measure $\mathrm{d}\mu$ previously introduced in (21). On the domain Ξ, let us suppose to generate nested sequences of points $\mathscr{G}^k = \left\{\xi_j^k\right\}$ where $\xi_j^k \in \Xi$. The sequence is nested if the following condition is satisfied

$$\mathscr{G}^{k-1} = \mathscr{G}^k \cap \mathscr{G}^{k-1}, \tag{32}$$

that allows the possibility to increase (decrease) the resolution level only adding (removing) a finite set of points for a fixed level k.

The nested property of the meshes directly turns into the nested character of the discretization operator

$$(\mathscr{D}_k f)_j = f(\xi_j^k) = v_j^k, \tag{33}$$

from which the discretization operator D_k^{k-1} is obtained directly removing from v^k all the components $v_j^k = f(\xi_j^k)$ where $\xi_j^k \in \mathscr{G}^k \setminus \mathscr{G}^{k-1}$.

The reconstruction operator $\mathscr{R}_k$ can be associated to the polynomial interpolation $\mathscr{P}_j^k$ on a fixed stencil $\mathscr{S}_j^k$ relative to the interval $[\xi_{j-1}^k, \xi_j^k]$. More details about the selection of the stencil and the construction of the polynomial $\mathscr{P}_j^k$ are reported in the section §3.4. In this case, the following prediction operator can be used:

$$(\mathrm{P}_{k-1}^k v^{k-1})_{2j-1} = (\mathscr{D}_k \mathscr{P}_j^{k-1})_{2j-1} = \mathscr{P}_j^{k-1}(\xi_{2j-1}^k). \tag{34}$$

In this setting, the error e^k is equal to zero for all the points $\xi_j^k \in \mathscr{G}^{k-1}$, while the number of non-redundant, *i.e.* linear independent, coordinates d^k is equal to $\mathrm{card}(\mathscr{G}^k \setminus \mathscr{G}^{k-1})$, where the wavelets are defined as follows

$$d_j^k = v_{2j-1}^k - (\mathrm{P}_{k-1}^k v^{k-1})_{2j-1} \quad \forall \xi_{2j-1}^k \in \mathscr{G}^k \setminus \mathscr{G}^{k-1}. \tag{35}$$

The point value setting shows a great flexibility in terms of polynomial reconstruction where well established techniques as the essentially non oscillatory (ENO) reconstructions procedure can be introduced [4,5] for virtually any kind of meshes. In the following section, a brief introduction to the truncation procedure is provided.

3.3 Truncation and stability requirements

One of the most important objective of multiresolution framework is the compression capability. The essential idea is to eliminate all the redundant information keeping only the significative one. Note that the compression process always generates an approximation of the initial data. Let us consider to have a function f, discretized on the finest resolution level $v^L = \mathscr{D}_L f$, that is represented in its multiresolution form v_{MR}. The aim of the truncation procedure is to generate an approximated multiresolution representation $\hat{v}_{MR}$. The truncation procedure yields, after the application of the *encoding* algorithm on $\hat{v}_{MR}$, a data set $\hat{v}^L$ which is close, in some norms and under certain accuracy requirements, to the original data v^L. Different truncation procedure can be designed to achieve the correct reproduction of the data obtaining the desired level of compression (see for instance [18]). In this paper, the truncation based on the elimination of the wavelets d^k under a prescribed tolerance is addressed. The problem statement is the following: given a sequence of scale coefficients or wavelets for a fixed level d^k and assigned a level dependent tolerance criterion ε_k, the idea is to generate $\hat{d^k} = \left\{\hat{d^k_j}\right\}_{j=1}^{J_k - J_{k-1}}$ according to

$$\hat{d^k_j} = \mathrm{tr}(d^k_j, \varepsilon_k) = \begin{cases} 0 & |d^k_j| \leq \varepsilon_k \\ d^k_j & \text{otherwise}. \end{cases} \tag{36}$$

Different choices exist in literature for the threshold parameter ε_k: a level independent choice $\varepsilon_k = \varepsilon$ or a dependent criterion $\varepsilon_k = \varepsilon/2^{L-k}$. Since the original work of Harten, the problem of the stability of the MR representation of the data has been analyzed. Harten proposed [15] to modify the *encoding* procedure to preserve the following condition

$$||v^L - \hat{v}^L|| \leq C\varepsilon, \tag{37}$$

with C as a constant with a norm as L^1 or L^∞. The interested reader can found a detailed analysis on the stability questions of the Harten framework with the relative error bounds in [7].

Classically, the effectiveness of a MR approach is measured in term of its compression capability, *i.e.* the number of activated *wavelets* N_w with respect to the dimension of the discrete space at the finest level $\dim(V_L) = J_L$. The compression ratio μ_{cr} is usually defined as follows:

$$\mu_{\mathrm{cr}} = \frac{J_L}{N_w + J_0}. \tag{38}$$

However, in the TE approach presented in this work, the algorithm does not require to know the solution at the finest level. This leads to the definition of an evaluation ratio τ, measuring the number of evaluations N_{eval} of the model with respect to the dimension of the full discrete space J_L

$$\tau = \frac{J_L}{N_{\mathrm{eval}}}. \tag{39}$$

In section §5.1, results on both the compression ratio μ_{cr} and the evaluation ratio τ are provided for different applications of the TE algorithm.

3.4 From the linear approach to the non linear one: introduction of the high-order ENO reconstruction

As previously described, the reconstruction operator $\mathscr{R}_k$ is related to the polynomial interpolant on a fixed stencil. The stencil must be intended as a set of cells or points depending from the setting.

In the following, the details for the selection, even non linear, of the stencil are presented in the case of 1D regular meshes. The reader can find further details on the multidimensional case in [5,4]. The case of the point value setting is explicitly addressed in the following; the cell average setting can be obtained substituting the point notion with the cell one.

The generic stencil $\mathscr{S}$ for a polynomial interpolation of order $r > 0$ is

$$\mathscr{S} = \mathscr{S}(r,s) = \{-s, -s+1, \ldots, -s+r\}, \quad \text{with } r \geq s > 0. \tag{40}$$

On the stencil $\mathscr{S}$ it is possible to define a number of $N_S = \text{card}(\mathscr{S})$ Lagrange polynomials:

$$L_m(y) = \prod_{\substack{l=-s \\ l\neq m}}^{-s+r} \left(\frac{y-l}{m-l}\right) \quad \text{with} \quad L_m(i) = \delta_{i,m} \text{ and } i \in \mathscr{S}. \tag{41}$$

For each $\xi \in [\xi_{j-1}, \xi_j]$ the generic polynomial is defined as

$$q_j(\xi; f, r, s) = \sum_{m=-s}^{-s+r} v_{j+m} L_m\left(\frac{\xi - \xi_j}{h}\right), \tag{42}$$

where $q_j(\xi_l) = v_l = f(\xi_l)$ (this condition must be replaced for the cell average setting imposing a (28)-like condition). To each $q_j(\xi; f, r, s) \in [\xi_{j-1}, \xi_j]$ the topological stencil associated is $\mathscr{S}_j = \left\{\xi_{j-s}, \xi_{j-s+1}, \ldots, \xi_{j-s+r}\right\}$ with $\text{card}(\mathscr{S}_j) = r+1$.

The interpolation error is defined as follows [23]

$$E(\xi) = f(\xi) - q_j(\xi) = f[\mathscr{S}_j, x]\omega_{r+1}(\xi), \tag{43}$$

where $f[\mathscr{S}_j, x]$ is the $r+1$st divided difference of f on the stencil $\mathscr{S}_j$ and ξ:

$$f[\mathscr{S}_j, \xi] = \sum_{\xi_m \in \mathscr{S}_j} \frac{f(\xi_m)}{\omega'_{r+1}(\xi_m)}, \tag{44}$$

where

$$\omega'_{r+1}(\xi_j) = \prod_{\substack{\xi_m \in \mathscr{S}_j \\ \xi_m \neq \xi_j}} (\xi_j - \xi_m). \tag{45}$$

The so-called nodal polynomial $\omega_{r+1}(\xi)$ in (43) is defined as follows

$$\omega_{r+1}(\xi) = \prod_{\xi_m \in \mathscr{S}_j} (\xi - \xi_m). \tag{46}$$

Then, using the notation employed in §3.2, without a lack of generality, let us consider $\Xi = [0,1]$, the polynomial interpolant $\mathscr{P}_j^k(\xi) = q_j(\xi, f, r, s)$ with $\xi \in [\xi_{j-1}^k, \xi_j^k]$ and the reconstruction operator $\mathscr{R}_k$ obtained as the union of each polynomial $\left\{\mathscr{P}_j^k(\xi)\right\}_{j=1}^{J_k}$ on $\mathscr{G}^k = \left\{\xi_j^k\right\}_{j=0}^{J_k}$ with $\xi_j^k = jh_k$ and $J_k = 1/h_k$.

If the stencil is fixed *a priori*, choosing both the degree r and the type of stencil s, the multiresolution framework is said to be linear. Looking at the form of the interpolation error (43), it is evident that the error is minimized by selecting centered stencils (once a divided difference is fixed). However, this general criterion is not sufficient for optimal polynomial interpolations for non smooth functions. The innovative contribution of Harten has been to provide a general non-linear alternative to the wavelet framework. This is accomplished just at level of the selection of the stencil shifting from a data-independent to a data-dependent selection.

The pivotal idea is constituted by the analysis of the interpolation error (43). Classical numerical analysis (see for instance [23]) results allow to link the $r+1$st divided difference $f[\mathscr{S}_j, \xi]$ with the regularity of the function (if it is smooth enough to be differentiable at least $r+1$ times), that is:

$$f[\mathscr{S}_j, \xi] = \frac{f^{(r+1)}(\overline{\xi})}{(r+1)!}, \tag{47}$$

where $\overline{\xi}$ is in the convex hull of the set $\mathscr{S}_j \cup \{\xi\}$.

The equations (47) and (46) inflated in (43) lead to

$$E(\xi) = \frac{f^{(r+1)}(\overline{\xi})}{(r+1)!} \mathscr{O}(h^{r+1}). \tag{48}$$

The case of non smooth function is dramatically different. If the stencil $\mathscr{S}_j$, with cardinality equal to $r+1$, contains a jump discontinuity, *i.e.* the stencil crosses it, the following relation holds

$$\begin{aligned} E(\xi) = f[S_j,\xi]\omega_{r+1}(\xi) = f[\xi_{i-s},\ldots,\xi_{i-s+r},\xi]\omega_{r+1}(\xi) &= \frac{f[\xi_{i-s+1},\ldots,\xi_{i-s+r}] - f[\xi_{i-s},\ldots,\xi_{i-s+r-1}]}{rh}\omega_{r+1}(\xi) \\ &= \cdots = \frac{\mathscr{O}([f])}{h^{r+1}}\omega_{r+1}(\xi) = \mathscr{O}([f]). \end{aligned} \tag{49}$$

The equation (49) is obtained considering $\xi_{i-s+1} < \xi < \xi_{i-s+r}$, where the jump of the function f in a point in the convex hull of $\mathscr{S}_j \cup \xi$ is indicated as $[f]$. If the jump is only at some derivative p, it follows that

$$f[S_j,\xi] = \begin{cases} \mathscr{O}([f^{(p)}])/h^{r-p} & \text{if} \quad r > p \\ \mathscr{O}(||f^{(r)}||) & \text{otherwise.} \end{cases} \tag{50}$$

The presence of discontinuities generates a degradation of the accuracy in each interval $[\xi_{j-1},\xi_j]$ associated to a stencil $\mathscr{S}_j$ which crosses discontinuity. However, a relevant aspect is the locality of the loss of accuracy for the interpolation in presence of non-smooth functions; this is a crucial aspect in the comparison between local approximation technique with respect to global approximation, like the generalized Polynomial Chaos (PC) [20].

In the MR context, a measure of the degradation of the interpolation is contained in the $r+1$ divided difference $f[\mathscr{S}_j,\xi]$. From this observation, Harten et al. introduced, in [19], the so-called Essentially non oscillatory (ENO) interpolation in the context of the numerical methods for conservation laws. The idea is to adapt the stencil, in presence of discontinuity, to avoid crossing the discontinuity; the interpolation is then carried out only using the regions of smoothness. Two different algorithms have been presented in [19]: a hierarchical selection and a non-hierarchical one. The non-hierarchical selection is demonstrated [6] to be able to detect even jump in the derivative of the functions. However, the non-hierarchical selection is, in the same paper, claimed to produce biased stencils away from discontinuity regions. For this reason, aiming to introduce the ENO technique in the MR context to gain in term of compression capabilities, the focus, in the present paper, is on a hierarchical selection of the stencil employing the following algorithm [19]

Algorithm 1: Hierarchical selection of the stencil

```
$s_0 = j$ ;
for $l = 0,\ldots,r-2$ do
    if $\left|f[\xi_{s_l-2},\ldots,\xi_{s_l+l}]\right| < \left|f[\xi_{s_l-1},\ldots,\xi_{s_l+l+1}]\right|$ then
        $s_{l+1} = s_l$ ;
    end
end
$s_j = s_{r-1}$ ;
```

to obtain the stencil $\mathscr{S}_j^{\text{ENO}} = \left\{\xi_{s_j-1},\xi_{s_j},\ldots,\xi_{s_j+r-1}\right\}$.

In the present paper, two different interpolations are used: the linear interpolation ($r=1$) and the cubic interpolation ($r=3$). In the linear case, the stencil is constituted by only two points and no ENO selection is possible. In case of a discontinuity, the interpolation degrades, as all the other techniques, but no influence on the adjacent intervals exists. The situation changes for the cubic interpolation. In this case, the stencil contains four points, than the presence of a discontinuity has the effect to degrade the interpolation in three intervals: the interval that contains the discontinuity and the two neighborings. In practice all the intervals contained in the stencil $\mathscr{S}_j$ are affected by the presence of the discontinuity.

The direct effect is to generate three significative coefficients for each discontinuity with a systematic lost in terms of compression capabilities in presence of non-smooth solutions, thus demanding the use of a ENO selection. In the A, the coefficients for the cubic interpolation case are reported with some further details on the selection of the stencils at the boundaries.

In this section, the generalized Harten framework has been reformulated with respect to the stochastic problem. Now, in the next section, a weak coupled scheme in the spatial/stochastic space with multiscale capabilities is presented.

4 Design of a weak coupled spatial-stochastic scheme

In this section, a MR scheme based on the elements presented in §3 is illustrated. First, the procedure to obtain a multi-scale representation of the solution performing the truncate and encode (TE) algorithm, is presented (§4.1). The time-advancing strategy is presented in section §4.2. Then, the TE strategy is extended in §4.4 to the case when solution depends on both physical space and stochastic space. The introduction of the physical space naturally leads to multi-scale schemes based on different physical discretizations. The general idea can be applied to virtually any kind of spatial discretization. However, in this paper, standard (Godunov first order) and high-resolution (MUSCL) finite volume schemes for fluid-dynamics are considered.

The aim of the present paper is the introduction of efficient schemes to address the solution of stochastic partial differential equations in presence of uncertain parameters. Classically the Harten's MR scheme has been introduced to allow the compact representation of a function in the physical space. In the case of time dependent problems, the evolution of the significant details has been connected to the equation by means of a CFL-based criterion. This criterion is not adapted to the stochastic space. The approach here presented is to obtain, at each time step and for each physical location, a compact discrete representation of the function of interest in the stochastic space. The compact representation is obtained in a way to require a reduced number of true evaluations of the model and, by means of the reconstruction operator $\mathscr{R}_k$, it can be considered, under the prescribed tolerance ε, to be equivalent to the full discrete solution. For each time step and for each physical location the so-called truncate and endode (TE) algorithm (see section §4.1) must be applied to obtain a compact representation of the discrete data in the stochastic space. The basic algorithm is extended to the partial differential equations, the spatial-TE (sTE), in the section §4.4.

4.1 The fundamental brick: a one time Truncate and Encode approach in the stochastic space

In this section, the truncate and encode TE algorithm is presented. The pivotal idea of the algorithm is to identify in the prediction error e^k, trough its linear independent components d^k, of a certain $k-$th level a measure of the quality of the predictor operator P_{k-1}^{k}. The results presented in §3.4 show that the interpolation error diminishes, moving from a coarser level to a finer one, with respect to the local regularity of the function and the local order of the polynomial employed for the interpolation. On the contrary, in presence of discontinuities, the error remains constant and of the order $\mathscr{O}[1]$. This allows to claim that, starting from the exact knowledge of a finer enough level k, and using the discretization operators $\mathscr{D}_k$, the recursive combinations of prediction operations via the operators P_{k-1}^{k} and evaluations of the error e^k are sufficient to determine the region where the the solution is well-known (under a certain criterion) or not.

In the following the algorithm is presented in an abstract way, while, in §4.3, the 1D algorithm is more precisely described. The focus is devoted to the point-value setting, while the cell-average framework and, in particular, its link with the semi-intrusive method [1] is left to further research. Moreover the point value setting is explicitly addressed and employed in the numerical test cases reported in §5. The algorithm starts with the definition of the coarsest level of resolution $k = 1$. On this level, the discretization operator is applied obtaining the discrete data v^1: $v^1 = \mathscr{D}_1 f$. By decimation, it is also possible to obtain the discrete data on the level $k = 0$ knowing only v^1:

$$v^0 = \mathrm{D}_1^0 v^1. \tag{51}$$

An *encoding* step analogous to what is normally done in the MR classic context (see (18)) is then completed computing the linear independent coefficients d^k of e^k for $k = 1$:

$$d^k = G_k(\mathrm{I}_k - \mathrm{P}_{k-1}^{k}\mathrm{D}_k^{k-1})v^k. \tag{52}$$

The truncation is applied on d^1 with respect to the threshold ε defined by the user and to the relation $\varepsilon_k = \varepsilon_k(\varepsilon, k)$:

$$\hat{d}^1 = \mathrm{tr}(d^1, \varepsilon_k). \tag{53}$$

This operation however is based on the knowledge of the resolution $k = L$ of the finest level on which the threshold is always equal to ε (see (36)). The knowledge of the resolution is intended to be exactly the integer $k = L$ assigned to the finest level if the coarsest is marked as $k = 0$ and at each refinement k is increased by one.

The data d^1 are investigated to locate the region of the domain in which the accuracy of the prediction via P^k_{k-1} is not adequate. The non-zero wavelets d^1_j are identified. At each non-zero wavelets corresponds a region in which the knowledge of the solution is insufficient under the criterion used in the truncation (53), thus demanding more information. In particular, after the generation of the mesh on the level $k = 2$, on all the cells/points inside the regions of the domain (at level $k = 0$) used to generate the corresponding wavelets d^1_j (this correspond in the 1D case to the interval $[\xi^0_{j-1}, \xi^0_j]$), the discretization operator $\mathscr{D}_2$ is applied, while in the region marked as well-represented, the *decoding* procedure is performed:

$$v^2 = \mathrm{P}^2_1 v^1 + E_2 d^2 \simeq \mathrm{P}^2_1 v^1. \tag{54}$$

The assumption in the equation (54) is that for every null wavelets at a level $k-1$, the corresponding wavelets at level k are null too. In the case of non null details, the equation (54) is not applied, but substituted by a direct (exact) discretization of the function by the operator $\mathscr{D}_k$ for $k = 2$.

Knowing v^2 and v^1, the *encoding* is performed computing d^2 and their truncated counterpart $\hat{d}^2$ by means of (36). The algorithm is then repeated until reaching the finest level L or a full satisfactory prediction, *i.e.* $d^k_j = 0$ for all $j = 1, \dots, J_k - J_{k-1}$.

The preliminary sequence of operations is the following:

- Generation of a nested set of meshes $\mathscr{G}^k$ for $k = 0, \dots, L$ (0 is the coarsest mesh);
- Definition of the operator $\mathscr{D}_k$, $\mathscr{R}_k$, D^{k-1}_k and P^k_{k-1} according to the setting (cell-average or point value) as described in sections §3.1 and §3.2;
- Setting a proper threshold ε and a proper relation for $\varepsilon_k = \varepsilon_k(\varepsilon, k; L)$
- Discretization of the level $k = 1$: $v^1 = \mathscr{D}_1 f$;
- Decimation of the discrete data v^1 to obtain $v^0 = \mathrm{D}^0_1 v^1$

The Truncate and Encode algorithm can be resumed as follows:

Algorithm 2: General Truncate and Encode algorithm

```
while $2 \le k \le L$ do
    Encoding:   $d^{k-1} = G_{k-1}(\mathrm{I}_{k-1} - \mathrm{P}^{k-1}_{k-2}\mathrm{D}^{k-2}_{k-1})v^{k-1}$;
    Truncation:   $\hat{d}^{k-1} = \mathrm{tr}(d^{k-1}, \varepsilon_{k-1})$ ;
    for $j = 1, \dots, J_k$ do
        if $v^{k+1}_j = (v^{k+1})_j \in V^{k+1} \setminus V^k$ then
            if $d^k_{j^\star} > 0$ then
                $v^{k+1}_j = (\mathscr{D}_{k+1} f)_j$ ;
            else
                $v^{k+1}_j = (\mathrm{P}^{k+1}_k v^k)_j$ ;
            end
        else
            $v^{k+1}_j = (\mathrm{P}^{k+1}_k v^k)_j$ ;
        end
    end
end
```

The index $j^\star$ is used to indicate, in this abstract version of the algorithm, the index of the wavelet at level k to the point at level $k+1$. The explicit example is reported in the 1D case in §4.3. Obviously, the algorithm stops when $\hat{d}^k = 0$, *i.e.* no further discretization are needed to reach the proper accuracy.

Note that the last step of the algorithm is possible only in presence of the consistency between the operators of prediction P_{k-1}^{k} and decimation D_{k}^{k-1} (see (13)); the decimation operator is never used (except in the preliminary stages D_{1}^{0}) to obtain a level $k-1$ from the finer k. Moreover, the consistency property between the discrete operator P_{k-1}^{k} and D_{k}^{k-1} guarantees that is always possible to obtain $k-1$ from k by decimation *a posteriori* and that the wavelets d^k represent the linear independent components of the error vector.

4.2 Time advancing strategy with accuracy preserving properties

In this section, the time advancing procedure applied to the TE algorithm is described. Here, the case of ordinary differential equations is considered to present the main idea, while the most complex case of partial differential equations is presented in §4.4, where the full spatial-TE algorithm (sTE) is presented.

Let us consider a stochastic Cauchy problem, *i.e.* find statistics for $y \in \mathscr{C}(0,T)$

$$\begin{cases} \dfrac{\mathrm{d}y(t,\xi)}{\mathrm{d}t} = f(t,y(t,\xi),\xi) \\ y(0,\xi) = y_0(\xi), \end{cases} \tag{55}$$

where ξ is the vector of d stochastic parameters $\xi \in \Xi \subset \mathbb{R}^d$ with probability distribution function $p(\xi,t) > 0$ for all $t \in [0,T]$.

Let us consider a uniform discretization of N_t intervals of length Δt on the time space $t = [0,T]$: $t_n = n\Delta t$ with $T = N_t \Delta t$. The solution can be discretized in time for fixed position in the stochastic space: $y_n(\xi) = y(t_n,\xi)$ and $f_n(\xi) = f(t_n,y_n,\xi)$. In the general case of linear multistep method of $q+1$ ($q \geq 0$) steps [23], it can be written:

$$y_{n+1}(\xi) = \sum_{r=0}^{q} \alpha_r y_{n-r}(\xi) + \Delta t \sum_{r=0}^{q} \beta_j f_{n-r}(\xi) + \Delta t \beta_{-1} f_{n+1}(\xi), \tag{56}$$

where if $\beta_{-1} = 0$ then the scheme is explicit. Note that the presence of an implicit scheme in time do not pose any problem in this context while it could represent an issue in the case of partial differential equation.

The TE algorithm (Algorithm 2) applied at each time step t_n provides the truncated multiresolution representation of $y(t_n,\xi)$. However, the application of the discretization operator $\mathscr{D}_k$ at a certain level k in a certain point j requires the knowledge of the solution in $q+1$ previous time steps (see (56)). In theory, these values are always known if the *decoding* procedure is applied to obtain the truncated solution at level L, and then by applying a sequence of decimation operators:

$$(y^k)_j = (\mathrm{D}_{k+1}^{k} \cdots \mathrm{D}_{L-1}^{L-2} \mathrm{D}_{L}^{L-1} y^L)_j \quad \text{with} \quad k < L. \tag{57}$$

As described in §4.1, the algorithm could stop at a level $k < L$ if the information related to the wavelets are sufficient to build the multiresolution representation of the solution within a prescribed accuracy. In this case, despite the possibility to *decode* (until L) and decimate (applying (57)) it is still possible to move directly from the maximum level reached $\overline{k}$ to the level needed k ($k > \overline{k}$). The solution in a generic point j on the level k is then obtained applying a sequence of prediction operators

$$(y^k)_j = (\mathrm{P}_{k-1}^{k} \mathrm{P}_{k-2}^{k-1} \cdots \mathrm{P}_{\overline{k}}^{\overline{k}+1} y^{\overline{k}})_j \quad \text{with} \quad k > \overline{k}. \tag{58}$$

The equation (56) can be recasted, considering explicitly the random vector $\xi_j^k \in \mathscr{G}^k$:

$$\begin{aligned} y_{n+1}(\xi_j^k) &= \sum_{r=0}^{q} \alpha_r y_{n-r}^k(\xi_j^k) + \Delta t \sum_{r=0}^{q} \beta_r f_{n-r}^k(\xi_j^k) + \Delta t \beta_{-1} f_{n+1}^k(\xi_j^k) \\ &= \sum_{r=0}^{q} \alpha_r (y_{n-r}^k)_j + \Delta t \sum_{r=0}^{q} \beta_r (f_{n-r}^k)_j + \Delta t \beta_{-1} (f_{n+1}^k)_j, \end{aligned} \tag{59}$$

where each term $(y^k_n)_j$ is obtained via the decimation or prediction sequences reported in the equations (57) and (58) and $(f^k_n)_j = f(t_n,(y^k_n)_j,\xi^k_j)$.

Practical applications of the described algorithm are given in the next two sections, in cell average and point-value settings.

4.3 1D TE example in point-value setting

Let us consider the stochastic ordinary problem (55) with one uncertain parameter $\xi \in \Xi \subset \mathbb{R}$. First, let us focus on the generation of the nested set of meshes. Equally spaced intervals can be defined both with respect to a Lebesgue or a probability measure. In the case of unbounded stochastic spaces the tessellation can be performed only on the basis of the probability measure. On the contrary, in the case of bounded distribution, both the approaches are still possible. When the stochastic space is bounded, it is always possible to map it, with a linear transformation, in the unitary hypercube $[0,1]^d$, so, without loss of generality to make the things clear unitary stochastic space are considered in the following. In both the point-value setting and the cell-average setting, when time dependent pdf are considered, if the measure is used to build the meshes, then the set of nested meshes $\mathscr{G}^k$ is a function of the time:

$$\mathscr{G}^k = \mathscr{G}^k(t). \tag{60}$$

This possibility can be applied in the TE algorithm, using at each time step nested set of meshes and consistent operators. In this paper, only uniform probability distributions are used, then the Lebesgue measure coincides with the probability measure. Some results on bounded time dependent and even discontinuous distributions can be found in [3] where they are afforded with the topological approach.

The preliminary operations for the TE algorithm are the following:

- Generation of a nested set of meshes $\mathscr{G}^k$ for $k = 0,\dots,L$ (0 is the coarsest mesh):

$$\mathscr{G}^k = \left\{\xi^k_j\right\}_{j=0}^{J_k} \quad \text{where} \quad \xi^k_j = j\frac{1}{J_k}. \tag{61}$$

- Definition of the operator $\mathscr{D}_k$, $\mathscr{R}_k$, D^{k-1}_k and P^k_{k-1} according to §3.2:

$$\begin{cases} (\mathscr{D}_k y(t,\xi))_j = y(t,\xi^k_j)p(\xi^k_j,t) \\ \mathscr{R}_k : (\mathscr{R}_k v^k)_l = (\mathscr{D}_k y(t,\xi))_l = y(t,\xi^k_l)p(\xi^k_l,t) \quad \text{with} \quad l \in \mathscr{S}^k_j \\ (\mathrm{P}^k_{k-1}v^{k-1})_j = (\mathscr{R}_{k-1}v^{k-1})_j \end{cases} \tag{62}$$

In the case of point-value setting, as presented in §3.2 the reconstruction operator $\mathscr{R}_k$ is, in each interval $\xi \in [\xi^k_{j-1},\xi^k_j]$, a polynomial of order r with a stencil of cardinality $\#S^k_j = r+1$.

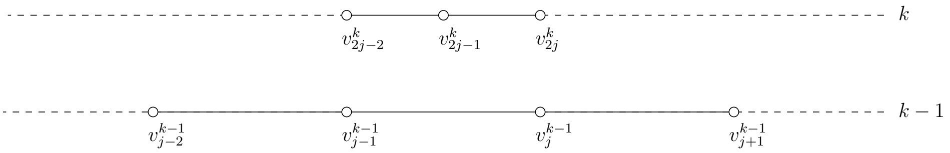


Fig. 2 Example of 1D stochastic nested meshes for the point-value setting decimation procedure.

Let us consider the situation sketched in figure 2, the decimation operator D^{k-1}_k is simply obtained as

$$(\mathrm{D}^{k-1}_k v^k)_j = v^{k-1}_j = y(t,\xi^k_{2j}) \quad \forall j = 0,\dots,J_k \tag{63}$$

- Setting of a proper threshold ε and a proper relation for $\varepsilon_k = \varepsilon_k(\varepsilon,k;L)$

- Discretization of the level $k=1$: $(v^1)_j = (\mathscr{D}_1 y)_j$ for all $j = 0,\dots,J_1$;
- Decimation of the discrete data v^1 to obtain $(v^0)_j = (\mathrm{D}_1^0 v^1)_j$ for all $j = 0,\dots,J_0$.

The TE algorithm in this case is resumed as follows

Algorithm 3: Truncate and Encode algorithm for the point value setting in 1D stochastic space.

```
while $2 \le k \le L$ do
    for $j = 1,\dots,J_{k-2}$ do
        Encoding:   $(d^{k-1})_j = v^{k-1}_{2j-1} - (\mathrm{P}^{k-1}_{k-2} v^{k-2})_{2j-1} = v^{k-1}_{2j-1} - \left(\mathscr{R}_{k-2} v^{k-2}\right)(\xi^{k-1}_{2j-1})$ ;
        Truncation: $\hat{d}^{k-1}_j = \mathrm{tr}(d^{k-1}_j, \varepsilon_{k-1})$ ;
    end
    for $j = 1,\dots,J_k$ do
        if $(v^k)_j \in V^k \setminus V^{k-1}$ then
            if $d^{k-1}_{j^\star} > 0$ then
                Discretization: $v^k_j = (\mathscr{D}_k y(t,\xi))_j = y(t,\xi^k_j) p(\xi^k_j)$ ;
            else
                Reconstruction: $(\mathscr{R}_{k-1}(t) v^{k-1})_l = (\mathscr{D}_{k-1}(t) y(t,\xi))_l$   with   $l \in \mathscr{S}^{k-1}_j$ ;
                Prediction: $v^k_j = (\mathrm{P}^k_{k-1} v^{k-1})_j = (\mathscr{R}_{k-1} v^{k-1})(\xi^k_j)$ ;
            end
        else
            Prediction: $v^k_j = (\mathrm{P}^k_{k-1} v^{k-1})_j = (\mathscr{R}_{k-1} v^{k-1})(\xi^k_j)$ ;
        end
    end
end
```

The index $j^\star$ is relative to the wavelets corresponding to the common interval. The typical situation is sketched in the figure 3 where the two discrete data generated by the generic wavelets $d_{j^\star}$ are explicitly reported. It is remarkable that in the case of $(v^k)_j \notin V^k \setminus V^{k-1}$ the prediction, due to the Lagrangian interpolation, produces exactly the shifting of the value at the previous resolution level ($v^{k+1}_{2j} = v^k_j$).

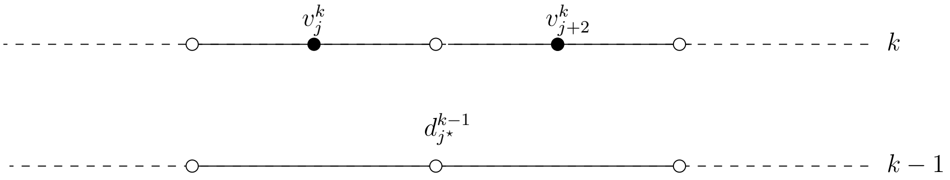


Fig. 3 Correspondence between the $j^\star$ index of the wavelet and the discrete data dependent from it at the successive higher resolution level.

Also in this setting, if the problem is time dependent as in (55), each evaluation depends on a finite number of previous time step at the same stochastic location. It is again possible to employ the time stepping procedure described in 4.2 (see equation (59)).

A cubic ENO-interpolation for the reconstruction operator $\mathscr{R}_k$ is presented in A with all the coefficients and details. The Lagrangian cubic interpolant described in A has been employed to obtain both the linear and non-linear (ENO) multiresolution schemes.

4.4 Extending the TE to partial stochastic differential equations of the solution

In this section, the TE strategy is applied to partial stochastic differential equations. In particular, the TE algorithm is applied to each physical coordinate, at each time step in the stochastic space with a proper technique to exchange data between different time steps with respect to the chosen spatial discretization for the deterministic solver. The overall strategy is named spatial-Truncate and Encode (sTE).

Note that in this work the case of a weak coupling between the physical and stochastic space is addressed; this means that it is possible to obtain a different multiresolution representation of the function of interest in the stochastic space for each physical space point while the MR representation remains explicitely addressed only in the stochastic space. However at each time step, the union of all the multiresolution representation for all the physical coordinates represents a compressed, *i.e.* more compact, representation of the solution in the overall space, both physical and stochastic. Another approach could consist in a strong coupling: represent at each time step the overall solution in the physical/stochastic space with only one multiresolution representation, applying a TE algorithm on the overall space. But, it demands a reformulation of the framework §3, and a much higher complexity in the implementation (definition of a new scheme on the coupled stochastic/physical space, building of a complex mesh). On the contrary, the sTE is very flexible and efficient.

The following stochastic problem, written in a conservative form, is considered:

$$\begin{cases} \dfrac{\partial u(\mathbf{x},t,\xi)}{\partial t} + \dfrac{\partial f(u(\mathbf{x},t,\xi))}{\partial \mathbf{x}} = 0 \\ \qquad\qquad + \text{initial condition} \\ \qquad\qquad + \text{boundary conditions}, \end{cases} \tag{64}$$

where the vector of the physical coordinates is $\mathbf{x} \in \Omega \subset \mathbb{R}^n$, t is the coordinate on the time space $t \in T \subset \mathbb{R}^+$ and the vector of the stochastic parameters $\xi \in \Xi \subset \mathbb{R}^d$. Let us suppose that the flux function could be a non linear function of the solution as this is the case in Burgers' equation.

First, the tessellation is generated for the physical space. This tessellation can be, without limitation, based on a whatever kind of grid, *i.e.* structured, non structured and even conformal or not. The tessellation is anyway a set of points or cells in the physical space depending on the deterministic scheme employed. From hereafter, the generic tessellation is indicated as $\mathscr{T} \subset \Omega$.

An overall threshold, ε, and the minimum $k = 0$ and maximum $k = L$ level of resolutions have to be chosen in terms of the number of elements in the stochastic meshes, *i.e.* choosing J_0 and J_L.

The sTE algorithm consists in the application, for all the time steps $t_n = n\Delta t \geq 0$, of the TE algorithm to each element $\mathbf{x}_i$ in $\mathscr{T}$ representing the function in the only stochastic space: $u(\mathbf{x}_i, t_n, \xi)$.

A sketch of the algorithm for N_t time intervals and N_x elements of the physical tessellation $\mathscr{T}$ is the following:

Algorithm 4: Generic spatial Truncate and Encode algorithm.

```
while $n = 1, \dots, N_t$ do
    for $i = 1, \dots, N_x$ do
        Preliminary operationsfor TE: ;

        – Generation of $\mathscr{G}^k$ for $k = 0, \dots, L$
        – Discretization on $k = 1$: $v^1 = \mathscr{D}_1 u(\mathbf{x}_i, t_n, \xi)$
        – Decimation $v^0 = \mathrm{D}_1^0 v^1$

        while $2 \le k \le L$ do
            Encoding:     $d^{k-1} = v^{k-1} - \mathrm{P}_{k-2}^{k-1} v^{k-2}$ ;
            Truncation:   $\hat{d}^{k-1} = \mathrm{tr}(d^{k-1}, \varepsilon_{k-1})$ ;
            for $j = 1, \dots, J_k$ do
                if $v_j^k \in V^k \setminus V^{k-1}$ then
                    if $d_{j/star}^{k-1} > 0$ then
                        Discretization: $v_j^k = (\mathscr{D}_k u(\mathbf{x}_i, t_n, \xi))_j$ ;
                    else
                        Reconstruction: $(\mathscr{R}_{k-1} v^{k-1})_l = (\mathscr{D}_{k-1} u(\mathbf{x}_i, t_n, \xi))_l$   with   $l \in \mathscr{S}_j^{k-1}$ ;
                        Prediction: $v_j^k = (\mathrm{P}_{k-1}^k v^{k-1})_j$ ;
                    end
                else
                    Prediction: $v_{2j}^k = (\mathrm{P}_{k-1}^k v^{k-1})_{2j}$ ;
                end
            end
        end
    end
end
```

The sTE, as described in the algorithm 4, could be seen as an ordinate sequence of applications of TE algorithm at different spatial locations $\mathbf{x}_i$ and at different time steps t_n. Obviously, being the problem (64) time dependent in general the solution at certain time steps has a dependence from the solution discretized at a previous time step. However in this case, the situation is slightly different from (59). Depending on the spatial discretization chosen, the solution, for a fixed time and space $u(\mathbf{x}_i, t_n, \xi)$, could depend from different spatial locations at the previous time step. This dependence is directly related to the spatial discretization chosen. In theory, this is not a problem, since the solution is represented knowing the multi-scale representation for each spatial location at each previous time step of interest. From an implementation point-of-view, an additional algorithm is demanded, the physical assembling (PhAs) algorithm, able to reconstruct at a time step t_n, when required a general vector containing the values of the function $u(\mathbf{x}, t, \xi_j)$ for the opportune spatial locations at the previous time steps $t < t_n$ for an assigned stochastic location ξ_j. Without any limitations the required vector could be the full discrete solution at the previous time step $v^k(\mathbf{x}, t, \xi_j)$ (for a generic level k and $t < 0$). The procedure assembles the value at different spatial locations that could be, in principle, not all already computed. In this case, the reconstruction operator is used to compute the values of the function where needed. The accuracy of the procedure is guaranteed by the previous application of the TE algorithm to the different spatial locations of interest at the previous time steps. Also in this case, it could be necessary to use the decimation or prediction cascade already presented in (57) and (58).

To make things clearer, let us consider a finite difference scheme with a second order spatial discretization for the fluxes. The equation (64) on an uniform 1D spatial mesh with step Δx, with an explicit Euler scheme in time ($\beta_{-1}=0$ and $q=0$ to reference to (59)), becomes:

$$u(x_i,t_{n+1},\xi)=u(x_i,t_n,\xi)-\frac{\Delta t}{2\Delta x}\Big(f(u(x_{i+1},t_n,\xi))-f(u(x_{i-1},t_n,\xi))\Big). \tag{65}$$

The stencil for each spatial point x_i at time t_{n+1} is then constituted by the points $\{x_{i-1},x_i,x_{i+1}\}$ at time t_n. These three points belong to the three different multi-scale representations associated to three different applications of the TE algorithm. If in some positions in the stochastic space ξ_j at a certain time t_{n+1} the function should be evaluated by the application of the operator $\mathscr{D}_k$, the exact (*i.e.* obtained by the solution of the equation) solution is computed using the values of the solutions $\{u(x_{i-1},t_n,\xi_j),u(x_i,t_n,\xi_j),u(x_{i+1},t_n,\xi_j)\}$. The role of the PhAs algorithm is to enter in the three multi-resolution representations $v(x_{i-1},t_n,\xi)$, $v(x_i,t_n,\xi)$ and $v(x_{i+1},t_n,\xi)$ and obtain by the cascade decimation (57) or prediction (58) the three values at ξ_j. More complex spatial discretization could be employed. In this paper, in the context of high-resolution FV scheme, the MUSCL-Hancock method (MHM), described in B, is used to solve the linear advection equation presented in section §5.

In section §5 an example of the application of the PhAs algorithm for the MHM is also presented.

5 Numerical results

In this section, various numerical results are presented. First, a so-called steady stochastic equation is presented in §5.1. In this case, various aspects of the application of the TE algorithm are presented on a function depending only from the stochastic space $f=f(\xi)$. This case is, in some sense, the equivalent of an image or signal compression case in which the classical MR framework can be applied knowing the full solution at the finest level. Here, the TE algorithm is applied showing the stability properties and the compression capabilities in term of both the compression ratio of evaluation and storage. This simple steady case is optimal to present the effect of the introduction of a non-linear approach, via the ENO interpolation to obtain the reconstruction operator $\mathscr{R}_k$, with respect to the linear one in term of the compression capability for discontinuous functions. The time advancing strategy presented in §4.2 is applied to two different stochastic ordinary differential equations in §5.2.1 for a scalar case and §5.2.2 for a vectorial one. Successively the solution of stochastic partial differential equations is addressed introducing the complete sTE algorithm. Two scalar cases are presented: the linear advection equation presented in §5.3 while the non-linear inviscid Burgers equation is presented in §5.4.

All the meshes are obtained using the agglomeration/splitting based on the Lebesgue measure, *i.e.* the points are equally spaced with respect to the parameter space Ξ and the number of intervals of each mesh $\mathscr{G}^k$ is chosen in order to be equal to

$$N_k=2^{m_k},\quad \text{where}\quad m_{k+1}>m_k \quad \text{and} \quad m_k\in\mathbb{N}. \tag{66}$$

If a unitary stochastic space is considered, each mesh $\mathscr{G}^k$ is then defined as the set of points $\mathscr{G}^k=\left\{\xi_j^k\right\}_{j=0}^{N_k}$ with $\xi_j^k=j\frac{1}{N_k}=jh_k$.

The relation used to generate the threshold level ε_k is

$$\varepsilon_k=\frac{\varepsilon}{2^{L-k}}. \tag{67}$$

All the results reported here, are compared with two standard non-intrusive UQ techniques, namely the Monte Carlo (MC) approach and the Polynomial Chaos method (PC). In this paper, a quasi-MC method based on Sobol sequences is used for all the numerical tests. For an exhaustive theoretical background of the two method the reader can refer to the book [20].

5.1 An introductory example: a steady stochastic function

The first problem is a steady equation in which a function $f = f(\xi)$ with $\xi \in \Xi \subset \mathbb{R}$ describes the stochastic output. The system is affected by an uncertain parameter ξ with distribution $p(\xi)$ here assumed uniform, *i.e.* $\xi \sim \mathscr{U}[0,1]$.

The aim is to compute the two first statistical moments, namely the expectancy $\mathbb{E}$ and the variance $\mathbb{V}\text{ar}$ following the definitions (3).

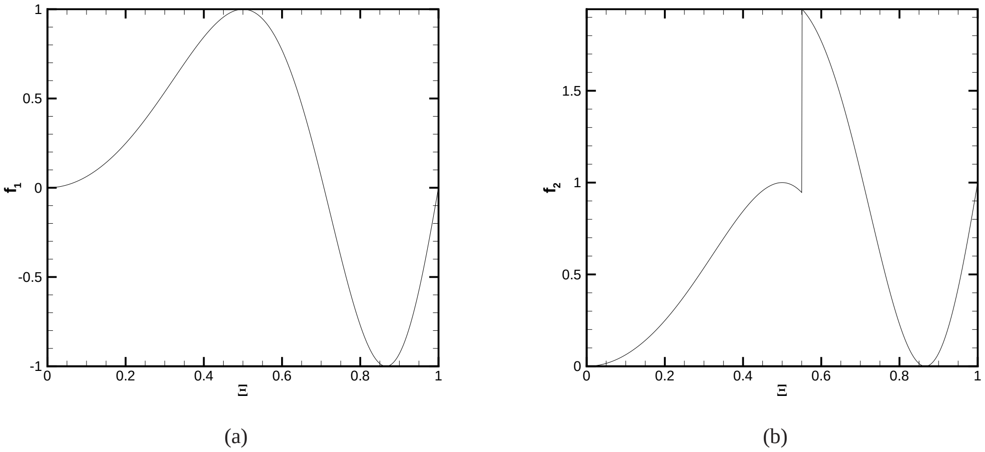
Two steady functions are studied. A continuous one $f_1 = f_1(\xi) = sin(2\xi^2\pi)$ and its discontinuous counterpart $f_2 = f_2(\xi)$ where

$$f_2 = \begin{cases} sin(2\xi^2\pi) & \text{if} \quad \xi \leq 11/20 \\ sin(2\xi^2\pi)+1 & \text{otherwise}. \end{cases} \tag{68}$$

In figure 4, both functions f_1 and f_2 are plotted.

For both cases, the exact solutions for the expectancy and the variance can be computed. These are used as reference values for computing normalized statistical errors, as follows

$$\begin{cases} \text{err}_{\mathbb{E}} = \dfrac{\mathbb{E} - \mathbb{E}_{\text{ex}}}{\mathbb{E}_{\text{ex}}} \\ \text{err}_{\mathbb{V}\text{ar}} = \dfrac{\mathbb{V}\text{ar} - \mathbb{V}\text{ar}_{\text{ex}}}{\mathbb{V}\text{ar}_{\text{ex}}}. \end{cases} \tag{69}$$



(a)

(b)

Fig. 4 Representation of the two steady functions f_1 (a) and f_2 (b).

The TE algorithm is then applied by varying the following parameters: a coarsest level with $m_0 = 3$, a finest level with an increasing $m_L = 6, \ldots, 20$ and $\varepsilon = 10^{-1}$ for the linear case ($r = 1$) and the high-order ($r = 3$), with and without the ENO interpolation to build the reconstruction operator $\mathscr{R}_k$. In table 1, results for the function f_1 are reported. In particular, different information are reported for each maximum level m_L employed: the number of activated wavelets N_w, *i.e.* the total number of details d_j^k which are greater than the threshold ε_k, the number of evaluated points N_{eval}, *i.e.* points in which the value of the function is obtained applying the discretization operator $\mathscr{D}_k$. The two ratio of compression μ and of evaluations τ, defined in (38) and (39), are also reported. Knowing the analytical description of the function, the norms in L_1 and L_∞ can be measured as follows

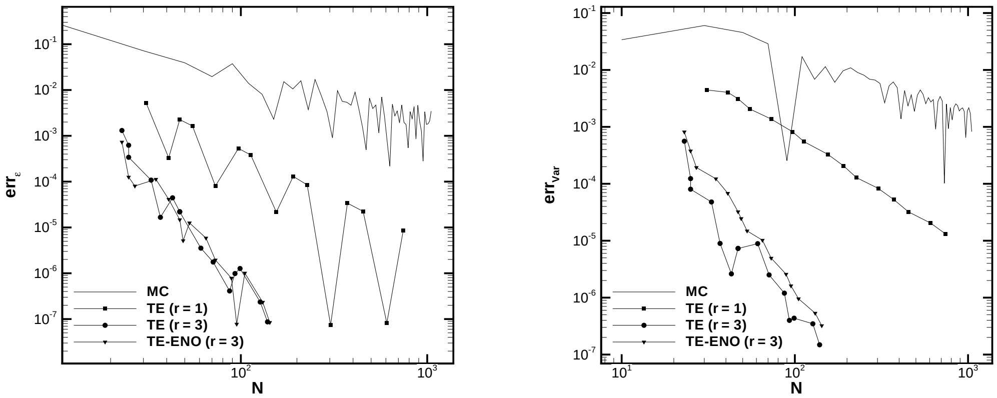
$$\begin{cases} \mathrm{err}_{L_1} = ||v^L - \hat{v}^L||_{L_1} \;=\; \dfrac{1}{N}\displaystyle\sum_{j=0}^{N_L+1} |v_j^L - \hat{v}_j^L| \\ \mathrm{err}_{L_\infty} = ||v^L - \hat{v}^L||_{L_\infty} \;=\; \max_j |v_j^L - \hat{v}_j^L|, \end{cases} \tag{70}$$

in which v^L represents the function discretized on the finest level and $\hat{v}^L$ its counterpart obtained by the application of the TE algorithm.

From the table 1, it is evident the advantage of the application of the TE algorithm employing the high-order interpolation as reconstruction operator $\mathscr{R}_k$. This can be seen by comparing the compression μ and evaluation τ ratios. However, differences between the high-order with and without the ENO selection of the stencil are negligible. In this case, the function is continuous and the ENO selection of the stencil obviously produces a slightly less accurate interpolation with respect to the high-order without ENO. As already described, this is due to the presence of the nodal polynomial in the error estimation (43).

In a UQ perspective, the interest is not only to obtain a compressed representation of the function at a fixed time, but to have the possibility to compute statistics in a more efficient way, *i.e.* with the lower possible number of simulation for a prescribed accuracy. Results in terms of the error for computing the expectancy and variance of the function f_1 with respect to the analytical solution are reported in figure 5 (as a function of the number of applications of the discretization operator $\mathscr{D}_k$). The errors are reported for the TE algorithm with $r = 1$, $r = 3$ (with and without the ENO selection of the stencil) and for the MC method. The MC results in figure 5 are obtained with an increasing number of simulations from 10 to 1050 (with step of 20 simulations). Of course, in this smooth case a direct comparison with the PC is not significative because it is well-known that PC represents the best polynomial approximation, with a spectral convergence.

The TE algorithm achieve the best efficiency in term of reached accuracy with a prescribed number of simulations, *i.e.* with a fixed number of exact evaluations of the model via the discretization operator $\mathscr{D}_k$, with respect to the MC. The introduction of the high-order interpolation ($r = 3$) for the reconstruction operator increases the performances for both the computation of expectancy $\mathbb{E}$ and variance $\mathbb{V}\mathrm{ar}$ with respect to the linear ($r = 1$) scheme. In this particular case no advantage is seen introducing the ENO selection of the stencil, as already discussed, due to the interpolation error (43).



(a)

(b)

Fig. 5 Statistical errors for the expectancy (a) and variance (b) following the definitions (69).

Then, the TE algorithm is applied to the function f_2. In table 2, the results are reported. Note that the parameters employed in this case are the same of the previous one.

m_L	N_w	N_{eval}	μ	τ	err_{L_1}	err_{L_∞}
			$r = 1$			
6	16	31	0.4062500E+01	0.2096774E+01	0.1076455E-01	0.2208432E-02
7	21	41	0.6142857E+01	0.3146342E+01	0.5951887E-02	0.1444001E-02
8	24	47	0.1070833E+02	0.5468085E+01	0.3773556E-02	0.1146526E-02
9	28	55	0.1832143E+02	0.9327272E+01	0.3072174E-02	0.9540855E-03
10	37	73	0.2770270E+02	0.1404110E+02	0.1877892E-02	0.5144779E-03
11	49	97	0.4181633E+02	0.2112371E+02	0.9460814E-03	0.2664874E-03
12	57	113	0.7187719E+02	0.3625664E+02	0.7898986E-03	0.2083328E-03
13	78	155	0.1050385E+03	0.5285806E+02	0.3863797E-03	0.1087813E-03
14	96	191	0.1706771E+03	0.8578534E+02	0.2366288E-03	0.6745963E-04
15	114	227	0.2874474E+03	0.1443568E+03	0.1965388E-03	0.5082159E-04
16	152	303	0.4311645E+03	0.2162937E+03	0.9923035E-04	0.2805389E-04
17	187	373	0.7009251E+03	0.3514021E+03	0.5915678E-04	0.1755482E-04
18	227	453	0.1154824E+04	0.5786865E+03	0.4904217E-04	0.1269737E-04
19	304	607	0.1724635E+04	0.8637380E+03	0.2434267E-04	0.6956671E-05
20	372	743	0.2818755E+04	0.1411275E+04	0.1478987E-04	0.4410474E-05
			$r = 3$			
6	12	23	0.5416667E+01	0.2826087E+01	0.1633177E-02	0.1789895E-03
7	13	25	0.9923077E+01	0.5160000E+01	0.7117425E-03	0.1119303E-03
8	13	25	0.1976923E+02	0.1028000E+02	0.7554024E-03	0.1135765E-03
9	17	33	0.3017647E+02	0.1554545E+02	0.2774332E-03	0.5256011E-04
10	19	37	0.5394737E+02	0.2770270E+02	0.2718788E-03	0.3384478E-04
11	22	43	0.9313636E+02	0.4765116E+02	0.2718788E-03	0.2265925E-04
12	24	47	0.1707083E+03	0.8717021E+02	0.5072283E-04	0.8208742E-05
13	24	47	0.3413750E+03	0.1743192E+03	0.5072283E-04	0.8209936E-05
14	31	61	0.5285484E+03	0.2686066E+03	0.5072513E-04	0.4289376E-05
15	36	71	0.9102500E+03	0.4615352E+03	0.1184295E-04	0.1585562E-05
16	44	87	0.1489477E+04	0.7532988E+03	0.6027842E-05	0.6814684E-06
17	47	93	0.2788787E+04	0.1409387E+04	0.3622172E-05	0.4827021E-06
18	50	99	0.5242900E+04	0.2647929E+04	0.2246242E-05	0.3843166E-06
19	64	127	0.8192016E+04	0.4128260E+04	0.1560115E-05	0.1572141E-06
20	70	139	0.1497967E+05	0.7543719E+04	0.7982875E-06	0.1013178E-06
			$r = 3$ (ENO)			
6	12	23	0.5416667E+01	0.2826087E+01	0.1633177E-02	0.2377694E-03
7	13	25	0.9923077E+01	0.5160000E+01	0.7526094E-03	0.1674474E-03
8	14	27	0.1835714E+02	0.9518518E+01	0.8017963E-03	0.1342711E-03
9	18	35	0.2850000E+02	0.1465714E+02	0.2852286E-03	0.5381801E-04
10	21	41	0.4880952E+02	0.2500000E+02	0.1601275E-03	0.2512965E-04
11	24	47	0.8537500E+02	0.4359575E+02	0.6364479E-04	0.1202350E-04
12	25	49	0.1638800E+03	0.8361224E+02	0.4692711E-04	0.1017512E-04
13	27	53	0.3034445E+03	0.1545849E+03	0.4408086E-04	0.7086252E-05
14	33	65	0.4965151E+03	0.2520769E+03	0.2570142E-04	0.3325651E-05
15	37	73	0.8856486E+03	0.4488904E+03	0.1224823E-04	0.1989242E-05
16	45	89	0.1456378E+04	0.7363708E+03	0.6257585E-05	0.8766845E-06
17	48	95	0.2730688E+04	0.1379716E+04	0.2960107E-05	0.6389484E-06
18	53	105	0.4946132E+04	0.2496619E+04	0.2399054E-05	0.4202189E-06
19	66	131	0.7943773E+04	0.4002206E+04	0.1513161E-05	0.1868357E-06
20	72	143	0.1456357E+05	0.7332706E+04	0.8283872E-06	0.1276827E-06

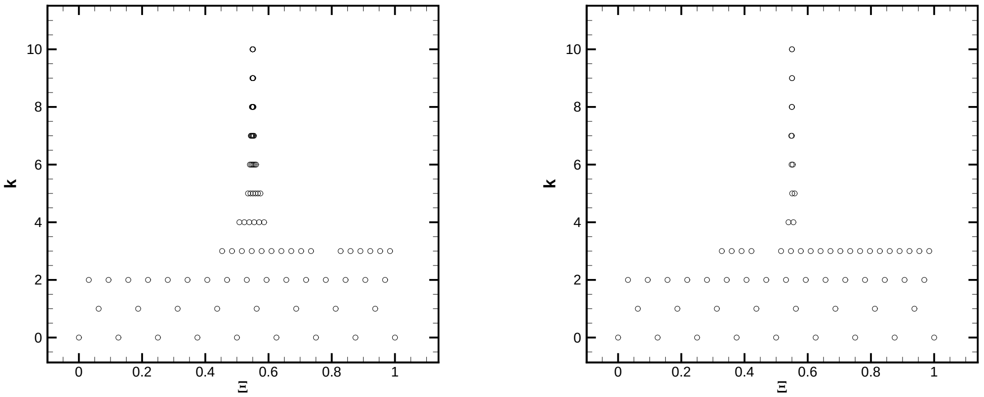
Table 1 Results of the application of the TE algorithm on the steady functions f_1 (see figure 4(a)).

In this case, the function is discontinuous and the best performances are achieved with the ENO selection of the stencil for the reconstruction operator $\mathscr{R}_k$. The effect of the ENO interpolation is then to obtain the same level of error with a lower number of evaluations. This is due to the degradation of the stencil in just one interval of the mesh respectiveness of the cardinality of the stencil associated to the polynomial degree chosen for the reconstruction operator $\mathscr{R}_k$. To make explicit the advantage related to the non-linear MR framework associated to the ENO selection, in the figure 6 the distribution of evaluated points is shown for both the TE algorithm with $r = 3$ with and without the ENO selection (the maximum level is equal to $m_L = 13$). It is evident how the

m_L	N_w	N_{eval}	μ	τ	err_{L_1}	err_{L_∞}
			$r = 1$			
6	18	33	0.3611111E+01	0.1969697E+01	0.1076455E-01	0.2036121E-02
7	24	45	0.5375000E+01	0.2866667E+01	0.5504224E-02	0.1247305E-02
8	27	51	0.9518518E+01	0.5039216E+01	0.3773556E-02	0.1120426E-02
9	32	61	0.1603125E+02	0.8409836E+01	0.3072174E-02	0.9275682E-03
10	42	81	0.2440476E+02	0.1265432E+02	0.1877892E-02	0.4878891E-03
11	54	105	0.3794444E+02	0.1951429E+02	0.9460814E-03	0.2630980E-03
12	63	123	0.6503175E+02	0.3330894E+02	0.7898986E-03	0.2049418E-03
13	85	167	0.9638824E+02	0.4905988E+02	0.3863797E-03	0.1053899E-03
14	103	203	0.1590777E+03	0.8071429E+02	0.2366288E-03	0.6703958E-04
15	122	241	0.2685984E+03	0.1359709E+03	0.1965388E-03	0.5040153E-04
16	161	319	0.4070621E+03	0.2054451E+03	0.9923035E-04	0.2763382E-04
17	196	389	0.6687398E+03	0.3369486E+03	0.5915678E-04	0.1750254E-04
18	237	471	0.1106097E+04	0.5565711E+03	0.4904217E-04	0.1264509E-04
19	315	627	0.1664410E+04	0.8361866E+03	0.2434267E-04	0.6904394E-05
20	383	763	0.2737799E+04	0.1374282E+04	0.1478987E-04	0.4403922E-05
			$r = 3$			
6	18	33	0.3611111E+01	0.1969697E+01	0.7117425E-03	0.8735547E-04
7	21	39	0.6142857E+01	0.3307692E+01	0.7117425E-03	0.9449274E-04
8	24	45	0.1070833E+02	0.5711111E+01	0.7554024E-03	0.9610891E-04
9	31	59	0.1654839E+02	0.8694915E+01	0.2774332E-03	0.3455729E-04
10	36	69	0.2847222E+02	0.1485507E+02	0.1440048E-03	0.1581990E-04
11	41	79	0.4997561E+02	0.2593671E+02	0.5072283E-04	0.7745037E-05
12	44	85	0.9311364E+02	0.4820000E+02	0.5072283E-04	0.7747611E-05
13	47	91	0.1743192E+03	0.9003297E+02	0.5072283E-04	0.7748743E-05
14	57	111	0.2874561E+03	0.1476126E+03	0.5072513E-04	0.3817509E-05
15	64	125	0.5120156E+03	0.2621520E+03	0.1184295E-04	0.1398270E-05
16	75	147	0.8738267E+03	0.4458299E+03	0.4712994E-05	0.4914680E-06
17	79	155	0.1659152E+04	0.8456323E+03	0.3622172E-05	0.4409493E-06
18	84	165	0.3120774E+04	0.1588758E+04	0.2246242E-05	0.3779013E-06
19	101	199	0.5190980E+04	0.2634618E+04	0.1560115E-05	0.1507988E-06
20	110	217	0.9532519E+04	0.4832152E+04	0.7982875E-06	0.9490253E-07
			$r = 3$ (ENO)			
6	14	25	0.4642857E+01	0.2600000E+01	0.3700040E-02	0.3713698E-03
7	16	29	0.8062500E+01	0.4448276E+01	0.7526094E-03	0.1680260E-03
8	18	33	0.1427778E+02	0.7787879E+01	0.8017963E-03	0.1342630E-03
9	23	43	0.2230435E+02	0.1193023E+02	0.2852286E-03	0.5378220E-04
10	28	53	0.3660714E+02	0.1933962E+02	0.1601275E-03	0.1929446E-04
11	31	59	0.6609677E+02	0.3472881E+02	0.4692711E-04	0.1011992E-04
12	32	61	0.1280312E+03	0.6716393E+02	0.4692711E-04	0.1012630E-04
13	35	67	0.2340857E+03	0.1222836E+03	0.4408086E-04	0.7037415E-05
14	42	81	0.3901190E+03	0.2022840E+03	0.2570142E-04	0.3276826E-05
15	47	91	0.6972128E+03	0.3600989E+03	0.1224823E-04	0.1940414E-05
16	57	111	0.1149772E+04	0.5904234E+03	0.4841718E-05	0.6950600E-06
17	59	115	0.2221576E+04	0.1139765E+04	0.2960107E-05	0.6360469E-06
18	65	127	0.4033000E+04	0.2064134E+04	0.2399054E-05	0.4181538E-06
19	79	155	0.6636570E+04	0.3382510E+04	0.1513161E-05	0.1847706E-06
20	86	169	0.1219276E+05	0.6204598E+04	0.8283872E-06	0.1256176E-06

Table 2 Results of the application of the TE algorithm on the steady functions f_2 (see figure 4(b)).

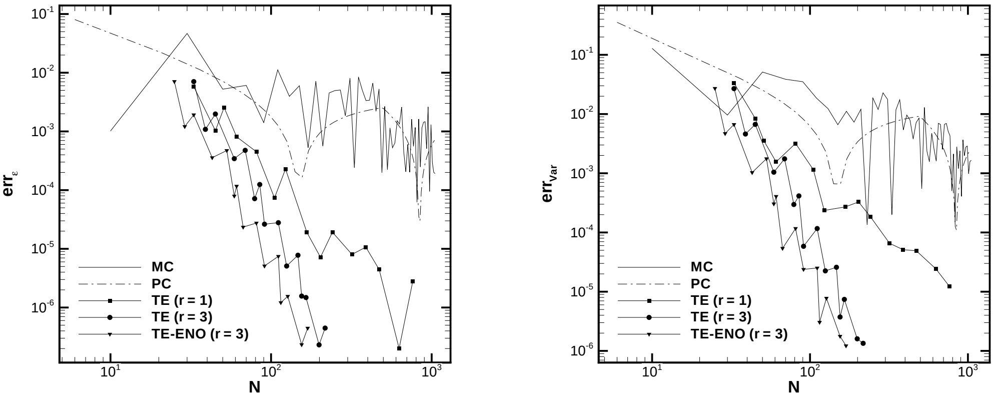
ENO selection of the stencil is associated to the locality of the degradation of the interpolation that traduces in a less polluted stencil, *i.e.* only the two wavelets generated at a level $k+1$th in the interval containing a discontinuity at level kth are affected by the degradation of the interpolation. Outside the interval containing the discontinuity, the interpolation is performed without degradation limiting the polluted region [6]. For the TE with a centered stencil $r = 3$, at each stencil containing a jump discontinuity corresponds three interval in which the interpolation degrades; of course in each degraded interval two active wavelets are generated. This means six active wavelets for each jump discontinuity against the two in presence of the ENO selection.



(a) (b)

Fig. 6 Distribution of the evaluated points for the function f_2 with $m_L = 13$, $m_0 = 3$ and $\varepsilon = 10^{-1}$ without (a) and with (b) the ENO selection of the stencil.

The results in terms of statistics for the function f_2 are reported in figure 7 where the PC with a number of simulation between 6 and 1041 (with steps of 15 simulations) and the MC with simulation between 10 and 1050 (with steps of 20 simulations) are reported. Also in this case the error are computed according to (69) using the exact solution for the statistics. It is evident how the TE algorithm works better than the MC and PC. Moreover, the high-order TE improves the performances of the linear TE. An increasing in terms of efficiency, *i.e.* lower error with the lower possible number of evaluations, is achieved with the introduction of ENO interpolation for $\mathscr{R}_k$. The direct effect is to increase, at the same level of error, the compression capability of the scheme (see also the table 2).



(a) (b)

Fig. 7 Statistical errors for the expectancy (a) and variance (b) following the definitions (69).

In the next section, the TE algorithm is employed in conjunction with the time integration procedure, presented in §4.2, to solve stochastic ordinary differential equations.

5.2 Scalar and vectorial ordinary differential equations

In this section, the TE algorithm is employed to solve ordinary differential equations. Two cases are analyzed: a scalar case §5.2.1 and a vectorial one §5.2.2.

The first case is inspired by the time evolution of the coverage surface presented in [20], while the vectorial case is the well-known Kraichnan-Orszag system of differential equations introduced in the 1967 by the authors as an inviscid turbulence model.

In both cases, the aim is to compute statistics of the solution in presence of an uncertain parameter (here supposed to be uniform distributed). The statistics are time dependent:

$$\begin{cases} \mathbb{E}(t) = \int_{\Xi} f(\xi,t)p(\xi)\mathrm{d}\xi \\ \mathbb{V}\mathrm{ar}(t) = \int_{\Xi} (f(\xi,t) - \mathbb{E}(t))^2 p(\xi)\mathrm{d}\xi. \end{cases} \tag{71}$$

In both cases the analytical solution is not known, then a fully converged MC solution is employed as reference solution in order to compute the error of the statistics at each time step. To measure the global error, three different norms, *i.e.* namely the L_1, L_2 and L_∞, are used according to the following definitions

$$\begin{cases} \mathrm{err}_{\mu^m}\big|_{L_p} = ||\mu^m(t) - \mu^m_{\mathrm{ref}}(t)||_{L_p} = \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{\mu_i^m(t) - \mu^m_{\mathrm{ref},i}(t)}{\mu^m_{\mathrm{ref},i}(t)} \right|^p \right)^{1/p} \\ \mathrm{err}_{\mu^m}\big|_{L_\infty} = ||\mu^m(t) - \mu^m_{\mathrm{ref}}(t)||_{L_\infty} = \max_i \left| \frac{\mu_i^m(t) - \mu^m_{\mathrm{ref},i}(t)}{\mu^m_{\mathrm{ref},i}(t)} \right|, \end{cases} \tag{72}$$

where the integer p can be equal to one or two and μ^m indicates the generic statistical moment.

5.2.1 A volume coverage evolution equation

The scalar case here analyzed is extracted from [20]

$$\begin{cases} \frac{\mathrm{d}\rho}{\mathrm{d}t} = \alpha(1-\rho) - \gamma\rho - \beta(\rho-1)\rho^2 \\ \rho(t=0) = \rho_0, \end{cases} \tag{73}$$

where $\rho \in [0,1]$ represents the surface coverage for a species. The evolution is governed by the surface absorption rate α, the desorption rate γ and the recombination rate β. This model is modified in this paper to obtain a non constant final state

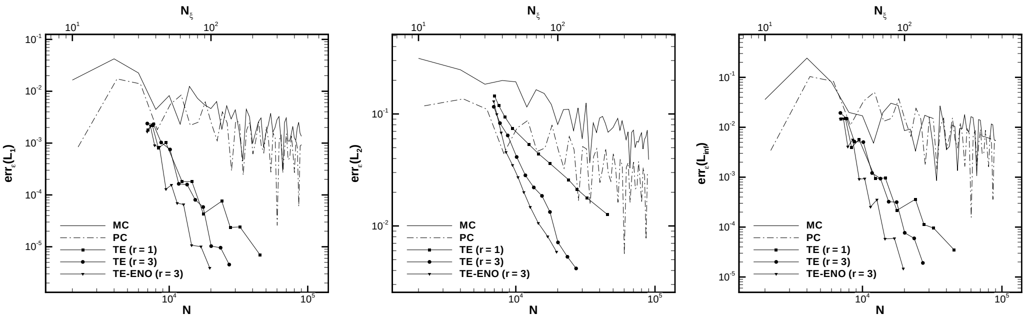
$$\begin{cases} \frac{\mathrm{d}\rho}{\mathrm{d}t} = \alpha(\bar{\rho}-\rho) - \gamma\rho - \beta(\rho-\bar{\rho})\rho^2 \\ \bar{\rho} = 1 + \frac{1}{2}\sin(5\xi + 8/5) \\ \beta = 20\xi, \end{cases} \tag{74}$$

where $\alpha = 1$, $\gamma = 0.01$ and $\xi \sim \mathscr{U}[0,1]$. A discontinuous initial solution in the stochastic space is chosen in order to obtain a discontinuous response

$$\rho(t=0) = \begin{cases} 3/4 \ \text{ if } \ 0.3 < \xi < 0.7 \\ 0 \quad \text{otherwise.} \end{cases} \tag{75}$$

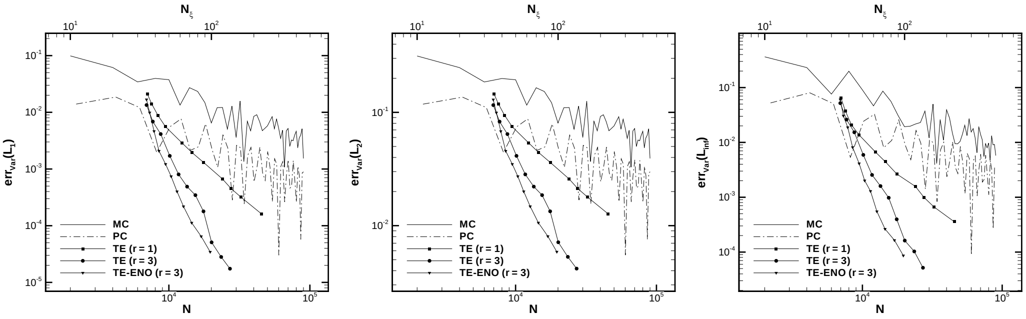
In this test case, the time space $T = [0,20]$ is discretized by a constant step equal to $\Delta t = 10^{-2}$ and an explicit Runge-Kutta with four increments, *i.e.* the classical RK4 method, is used. In the following, the results are presented in terms of total number of exact evaluations of the model, *i.e.* number of applications of the discretization operator $\mathscr{D}_k$, indicates with N. However, the total number of evaluations corresponds to a number of points in the stochastic space N_ξ that is constant in the case of the MC and PC and vary in time for the TE algorithm, *i.e.* $N_\xi = N_\xi(t)$. To make the comparisons clear both the number of points are reported in the following.

The results of the application of the TE algorithm for $m_0 = 4$ and a maximum level between 6 and 16, with $\varepsilon = 10^{-1}$, are reported in figures 8 and 9 for the mean and variance, respectively. The quasi-MC Sobol and PC are also reported. In all the cases, the TE algorithm displays better performances than both the MC and PC. The high-order algorithms ($r = 3$) performs also better than the linear one ($r = 1$). The non linear version of the scheme, the TE-ENO, shows the best performances for both the statistics.



(a) (b) (c)

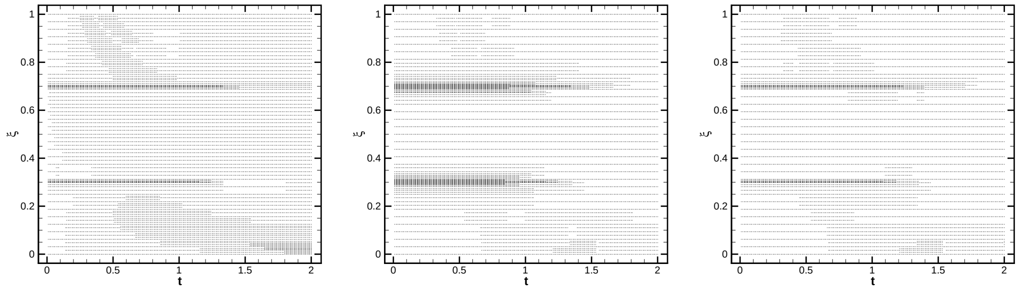
Fig. 8 Errors for the expectancy $\mathbb{E}$ of the surface coverage (74), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.



(a) (b) (c)

Fig. 9 Errors for the variance of the surface coverage (74), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

In figure 10, the distribution of the discretization operations are represented. For each time, each application of the operator $\mathscr{D}_k$ is marked by a point. Once fixed the minimum level $m_0 = 4$, the maximum $m_L = 12$ and the threshold $\varepsilon = 10^{-1}$, it is possible to compare the distribution of points for the three different TE schemes, namely linear $r = 1$, high-order $r = 3$ and high-order $r = 3$ with ENO selection of the stencil. It is evident how the introduction of the higher-order reconstruction for $\mathscr{R}_k$ reduces the number of points with the lower possible number of evaluations in the smooth regions. However, the presence of the two discontinuities causes the formation of a polluted region, *i.e.* in which the interpolation degrades, then no high compression in the discontinuous region 10(b) can be obtained. The introduction of the ENO selection of the stencil for the reconstruction operator $\mathscr{R}_k$ recover the narrow polluted region of the linear ($r = 1$) scheme 10(c).



(a) (b) (c)

Fig. 10 Distribution, in the combined time-stochastic space $T - \varXi$, of the application of the discretization operator $\mathscr{D}_k$ for the TE algorithm with $r = 1$ (a), $r = 3$ (b) and with $r = 3$ and the ENO selection of the stencil (c).

5.2.2 An inviscid turbulence model: the Kraichnan-Orszag problem

The present section deals with the solution of vectorial stochastic differential equations. The extension of the scalar TE strategy is straightforward. In a general case, it is sufficient to compute a wavelets for each component of the vectorial function and to choose, for each cell/point, the maximum of the wavelets. If a vectorial function $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is of interest, the TE algorithm is applied to each component y_i and the wavelets d_j^k are computed as the maximum of the wavelets $d_{j,i}^k$ relative to each ith component, *i.e.*

$$d_j^k = \max(d_{j,1}^k, \ldots, d_{j,n}^k). \tag{76}$$

The text case proposed in this section is the Kraichnan-Orszag problem proposed in 1967 [22] by Orszag. It is a three mode problem modeling an inviscid turbulence system given by a set of ordinary differential equations. The original system [22], rotated by $\pi/4$ around the axis y_3 can be written following [28] as

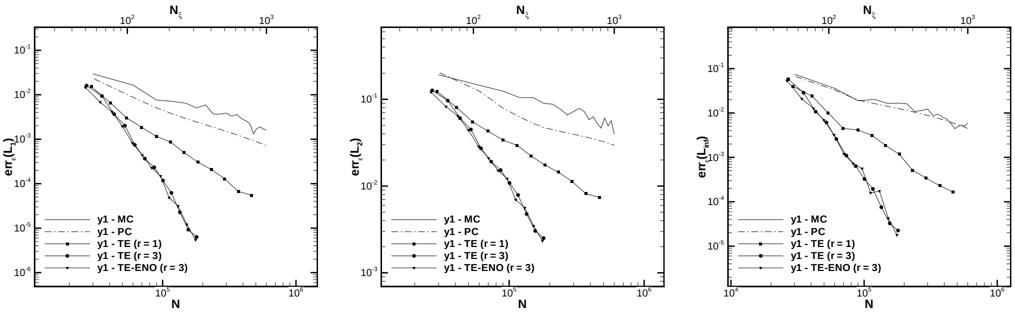
$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}t} = y_1 y_3 \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}t} = -y_2 y_3 \\ \dfrac{\mathrm{d}y_3}{\mathrm{d}t} = -y_1^2 + y_2^2. \end{cases} \tag{77}$$

The equation (77) is correlated to the following (uncertain) initial condition $\mathbf{y}(t = 0) = (1, 0.1\xi, 0)^{\mathrm{T}}$, where the parameter ξ is uniformly distributed between -1 and 1, $\xi = 2\omega - 1$, and $\omega \sim \mathscr{U}[0,1]$. A classical RK4 formula is used for the time integration. The time $T = [0, 30]$ is divided in $N_t = 600$ equal steps of length $\Delta t = 0.05$.

The TE algorithm is applied with the following parameters: $m_0 = 4$, m_L between 6 and 17 and a threshold equal to $\varepsilon = 10^{-1}$. As in the previous section, three reconstruction operators R_k are employed: linear $r = 1$, centered cubic $r = 3$ and cubic with ENO selection of the stencil. The MC and PC are also used with a number of simulation N_ξ, for each time step between 50 and 1000 for the MC and a degree between 50 and 1000 for the PC. The reference solution is a fully converged solution with $N_\xi = 20 \times 10^6$ samples generated with a quasi-MC Sobol sequences.

In this section, the equations (72) are employed to compute errors norms for the expectancy and variance for all the variables. However, due to the presence of null values for the expectancy of the second y_2 and third y_3 variable, the norms are computed on the time interval $[8,30]$, that correspond to 440 time steps.

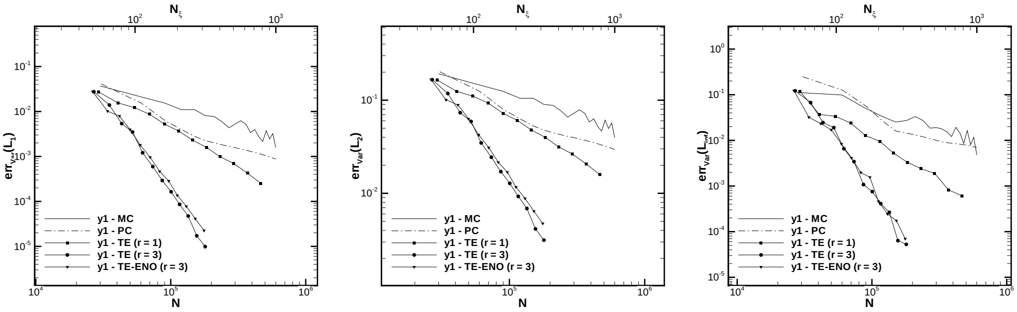
In figure 11, the error norms for the expectancy of y_1 are reported in the three norms L_1 11(a), L_2 11(b) and L_∞ 11(c). The TE algorithm shows better convergence rate with respect to both MC and PC method and for all the norms. Moreover, the introduction of the cubic reconstruction improves the rate of convergence of the method, instead the ENO selection of the stencil does not introduce an improvement with respect to the centered cubic reconstruction. It is possible to note that the improvement of the ENO selection of the stencil in the region near $\xi = 0$ (where a jump discontinuity is present) does not compensate the slightly degradation of the interpolation due to the employment of non-symmetric stencils.



(a) (b) (c)

Fig. 11 Errors for the expectancy $\mathbb{E}$ of the variable y_1 of the problem (77), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

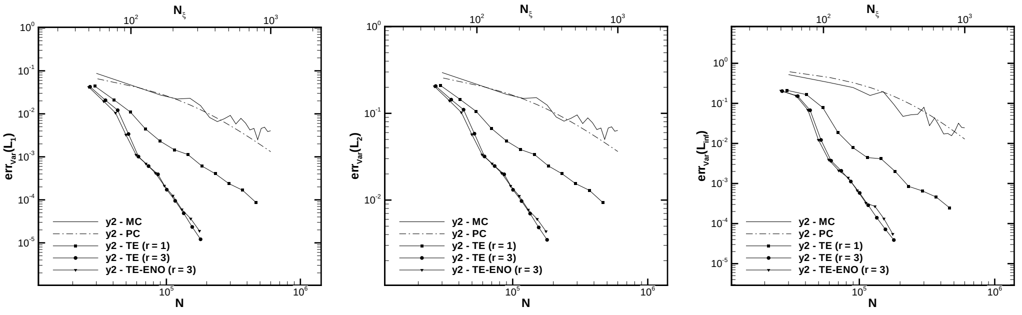
The error norms for the variance of y_1 are then reported in figure 12 where the same qualitative results of the norms for the mean hold.



(a) (b) (c)

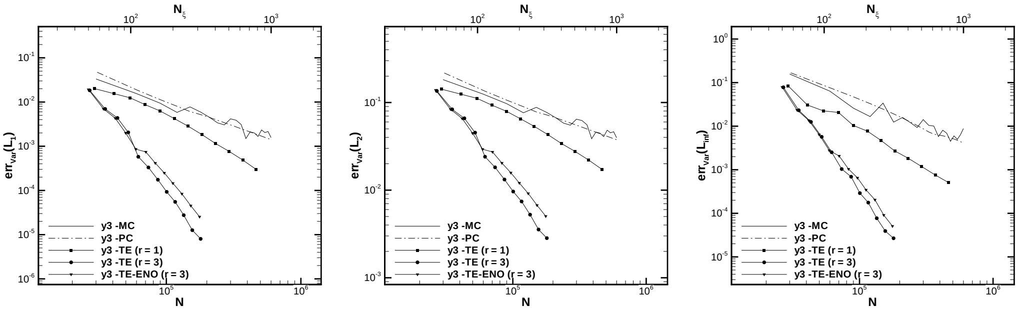
Fig. 12 Errors for the variance $\mathbb{V}\mathrm{ar}$ of the variable y_1 of the problem (77), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

For the variables y_2 and y_3, only the norms for the variance are reported in figures 13 and 14 respectively. The TE algorithm again shows the better convergence properties with respect to both MC and PC methods. The application of higher reconstruction operators makes possible to achieve better convergence properties and again the ENO selection of the stencil does not introduce any improvements. For y_3, the ENO selection of the stencil causes a slightly increase of the error with respect to the cubic centered reconstruction. A cure could be to employ an adaptive shifting between centered and ENO selection of the stencils based on some regularity criterion of the function.



(a) (b) (c)

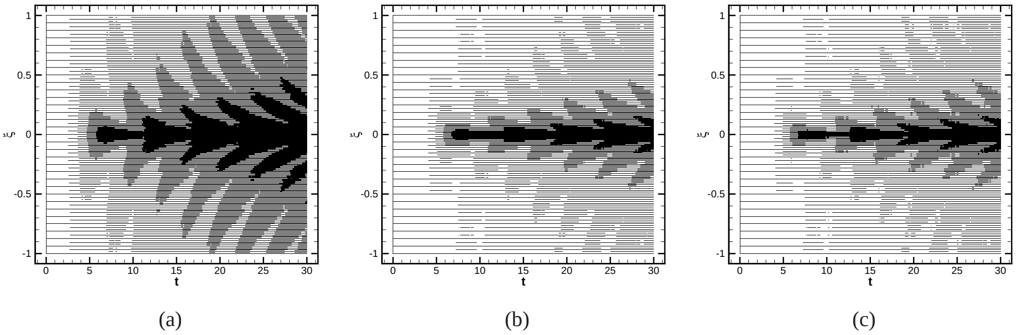
Fig. 13 Errors for the variance $\mathbb{V}\mathrm{ar}$ of the variable y_2 of the problem (77), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.



(a) (b) (c)

Fig. 14 Errors for the variance $\mathbb{V}\mathrm{ar}$ of the variable y_3 of the problem (77), with norms L_1 (a), L_2 (b) and L_∞ (c) following (72). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

In order to compare the compression capabilities of the different TE schemes, in the figure 15 the distribution of the discretized values in the plane $T - \Xi$ are reported. The pattern of points corresponds to the application of the TE algorithm with the following values for the parameters: $m_0 = 4$, $m_L = 12$ and $\varepsilon = 10^{-1}$. The overall number of points N is equal to 148 746 for the TE ($r = 1$) (figure 15(a)), 90 638 for TE ($r = 3$) (figure 15(b)) and 86 880 for TE $r = 3$ with ENO selection (figure 15(c)). Note that the pattern of the point is symmetric for the first two cases, but, due to the application of the TE selection of the stencil loses its symmetric pattern, with a slightly increasing of the errors, as evident in figures 11, 12, 13 and 14.



(a) (b) (c)

Fig. 15 Distribution, in the combined time-stochastic space $T-\Xi$, of the application of the discretization operator $\mathscr{D}_k$ for the TE algorithm with $r=1$ (a), $r=3$ (b) and with $r=3$ and the ENO selection of the stencil (c).

The TE algorithm is very efficient in dealing with multi-scale problems in the context of stochastic differential equations. The aim of the next sections is to show how the extension ot the TE algorithm, namely spatial-TE method (sTE), allows to obtain the same efficiency also for stochastic partial differential problems.

5.3 A (spatial) linear pde: the linear advection equation

In this section, the TE algorithm is applied to partial stochastic differential equations. In this kind of problem, the solution depends on both physical and stochastic spaces.

The first problem is a linear advection equation in one spatial dimension x and affected by the presence of one uncertain parameter ξ

$$\frac{\partial u(x,t,\xi)}{\partial t}+\frac{\partial f(u(x,t,\xi))}{\partial x}=0, \tag{78}$$

where the flux is formulated as follows

$$\begin{cases} f(u(x,t,\xi))=a(\xi)u(x,t,\xi) \\ \qquad a(\xi)=\dfrac{1}{40}e^{5\xi^2}+\dfrac{1}{5} \end{cases} \tag{79}$$

with the random parameter ξ uniformly distributed, *i.e.* $\xi\sim\mathscr{U}[\frac{1}{5},\frac{4}{5}]$ and the (physical discontinuous) initial condition

$$u(x,\xi,0)=\begin{cases} 1 & \text{if} \quad \dfrac{2}{5}\le x\le\dfrac{3}{5} \\ 0 & \text{if} \quad \text{otherwise}. \end{cases} \tag{80}$$

Physically, the problem (78) and (79) represents a linear advection equation, where the velocity a is uncertain.

The deterministic solver is a second order MUSCL-Hancock method (MHM) with a Roe superbee slope limiter [24]. More details are given in B. Now, let us focus on the physical assembling algorithm in this case (see Section §4.4). As in a classical FV scheme, the MHM deterministic scheme, once fixed a point x_i in the physical space, requires the solution at the two adjacent points x_{i-1} and x_{i+1}. Moreover, the MHM scheme requires the slope associated to the three points at the physical coordinates x_{i-1}, x_i and x_{i+1}. Each of these slope can be computed only basing on the values of adjacent points. Finally, to compute the solution $u_i^n(\xi_j)=u(x_i,t_n,\xi_j)$ in the point x_i, the overall stencil is constituted by five points $\{u_{i-2}^n,u_{i-1}^n,u_i^n,u_{i+1}^n,u_{i+2}^n\}$. The physical assembling procedure is then performed by means of the following algorithm:

Algorithm 5: Physical assembling algorithm applied to the MUSCL-Hancock method for the function $u(x_i,t_n,\xi_j)$ at level kth: $(v^n_{ij})^k = \mathscr{D}_k u(x_i,t_n,\xi^k_j)$.

```
for $l = i-2,\dots,i+2$ do
    if $k < \bar{k}$ then
        $(v^n_{lj})^k = (\mathrm{D}^k_{k+1}\mathrm{D}^{k+1}_{k+2}\cdots\mathrm{D}^{\bar{k}-1}_{\bar{k}}(v^n_l)^{\bar{k}})_j$ ;
    else
        $(v^n_{lj})^k = (\mathrm{P}^k_{k-1}\mathrm{P}^{k-1}_{k-2}\cdots\mathrm{P}^{\bar{k}+1}_{\bar{k}}(v^n_l)^{\bar{k}})_j$ ;
    end
end
```

At the end of the algorithm 5, the physical vector $\mathbb{PV}^k_{ij} = \{(v^n_{i-2\,j})^k,(v^n_{i-1\,j})^k,(v^n_{ij})^k,(v^n_{i+1\,j})^k,(v^n_{i+2\,j})^k\}$ is obtained with the generic discrete function equal to $(v^n_{ij})^k = \mathscr{D}_k u(x_i,t_n,\xi^k_j)$. Using the discretization operator $\mathscr{D}_k u(x_i,t_n,\xi_j)$ can be interpreted as the application of the physical vector assembling in order to obtain $\mathbb{PV}^k_{ij}$ and the consequent application of the deterministic scheme as follows

$$(v^{n+1}_{ij})^{k+1} = \mathrm{MHM}((\mathbb{PV}^k_{ij})^n,\Delta t), \tag{81}$$

where MHM indicates the sequence of operations of the MUSCL-Hancock method (see B).

The linear advection equation (78) is solved here for a fixed spatial mesh of $N_x = 101$ cells equally spaced in $\Omega = [0,1]$. The time space is discretized with $N_t = 250$ time steps of length $\Delta t = 4\times 10^{-3}$. The reference solution is chosen as the solution obtained with $N_\xi = 2^{21}+1 = 2\,097\,153$ equally spaced points in the stochastic space with the same physical space and time discretization.

In this case, the error norms for the statistics are computed on the whole physical-time space using the following relations

$$\begin{cases} \mathrm{err}_{\mu^m}\big|_{L_p} = ||\mu^m - \mu^m_{\mathrm{ref}}||_{L_p} = \left(\dfrac{1}{N_t\times N_x}\displaystyle\sum_{i=1}^{N_t}\sum_{j=1}^{N_x}\left|\mu^m_{ij} - \mu^m_{\mathrm{ref},ij}\right|^p\right)^{1/p} \\ \mathrm{err}_{\mu^m}\big|_{L_\infty} = ||\mu^m - \mu^m_{\mathrm{ref}}||_{L_\infty} = \max\limits_{ij}\left|\mu^m_{ij} - \mu^m_{\mathrm{ref},ij}\right|, \end{cases} \tag{82}$$

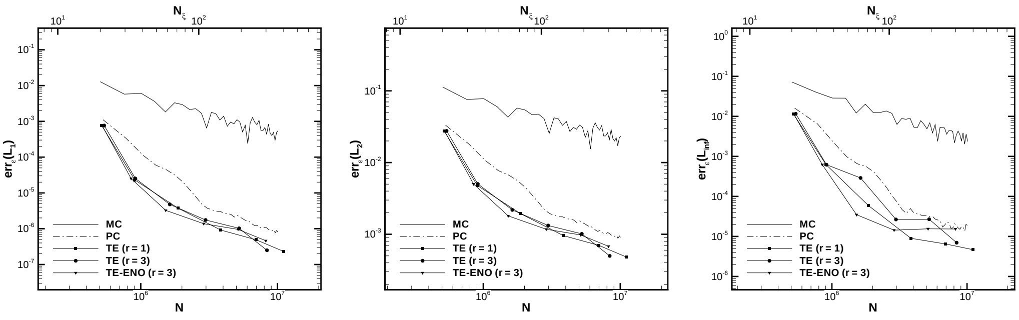
with $\mu_{ij} = \mu(x_j,t_i)$ the generic statistical moment.

In figure 16, the error norms of the expectancy $\mathbb{E}$ of u (78) are reported. The sTE algorithm is applied with $m_0 = 3$, m_L between 5 and 15 (with increment of 2) and $\varepsilon = 10^{-4}$. The MC (PC) curves are obtained with a number of points (degrees) between 20 and 400 with increments of 10. For each norm, the error for MC is higher than for sTE and PC, and sTE displays the best performances. In this case the advantage of an high-order reconstruction operator $\mathscr{R}_k$ are less evident because the solution is formed by a series of plateaux. As a consequence, the compression capability of the scheme is already accomplished with a linear reconstruction operator, while it remains useful to introduce the ENO stencil selection when higher-order reconstruction are employed.

In figure 17, the error norms are computed for the variance of the solution u (78). Also in this case the qualitative results are the same described for the expectancy (the sTE algorithm performs better than PC).

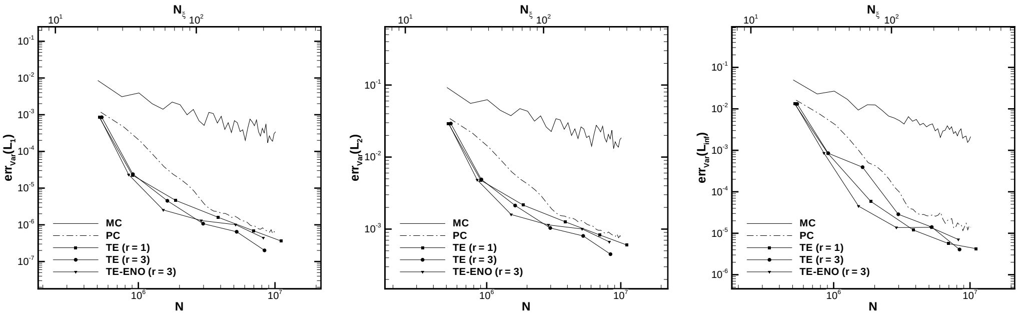
The pattern of the points in the combined physical-stochastic $\Omega - \Xi$ space are reported in figure 18 ($m_0 = 3$, $m_L = 9$ and $\varepsilon = 10^{-4}$). A very good compression, associated to an high-order reconstruction operator employing an ENO stencil selection, is limited to a narrow region relative to the moving hat (see equation (78)). This could reduce the advantage of an high-order TE scheme compared to a linear scheme due to the presence of the constant regions where the linear reconstruction is already accurate.

In the next section, the sTE is applied to a non-linear partial differential equation where the advantage of the application of the sTE scheme over the MC and PC is more evident. At the same time, the introduction of high-order reconstruction operators $\mathscr{R}_k$ with an ENO selection of the interpolation stencil improves the quality of the results.



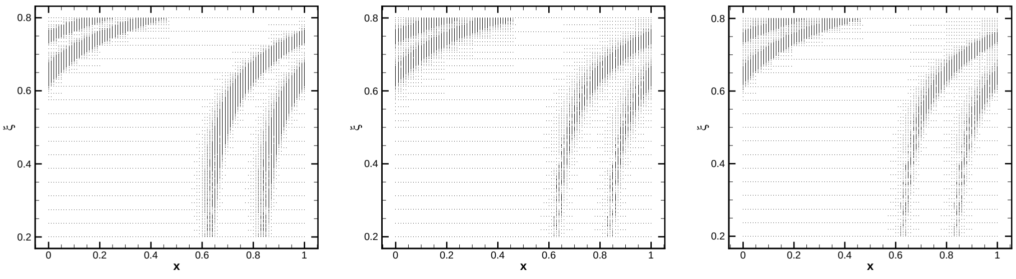
(a) (b) (c)

Fig. 16 Errors for the expectancy $\mathbb{E}$ of the solution of Eq. (78), with norms L_1 (a), L_2 (b) and L_∞ (c) following (82). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.



(a) (b) (c)

Fig. 17 Errors for the variance $\mathbb{V}$ar of the solution u (78), with norms L_1 (a), L_2 (b) and L_∞ (c) following (82). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.



(a) (b) (c)

Fig. 18 Distribution, in the combined physical-stochastic space $\Omega - \Xi$, of the application of the discretization operator $\mathscr{D}_k$ for the TE algorithm with $r = 1$ (a), $r = 3$ (b) and $r = 3$ (ENO selection of the stencil) (c) for the solution u (78) at a time $t = 2$.

5.4 A fully non-linear pde: the inviscid Burgers equation

In this section, the sTE scheme is applied to non-linear Burgers equations considering an uncertain smooth initial solution.

The problem is formulated as follows

$$\begin{cases} \dfrac{\partial u(x,t,\xi)}{\partial t} + \dfrac{\partial f(u(x,t,\xi))}{\partial x} = 0 \\ \qquad u(x,0,\xi) = u_0(x,\xi) = sin(x\pi\xi), \end{cases} \tag{83}$$

where the flux function is the non-linear Burgers flux $f(u(x,t,\xi)) = \frac{1}{2}u^2(x,t,\xi)$ and the uncertain parameter ξ is uniformly distributed $\xi \sim \mathscr{U}[\frac{3}{2}, \frac{5}{2}]$.

In this case, a first order Godunov scheme is used with an explicit Euler formula for the integration in time. The physical space $\Omega = [0,1]$ is constituted by $N_x = 101$ cells constructed around 101 equally spaced points. Periodic boundary conditions are also used. The time space $T = [0,2]$ is divided in $N_t = 500$ equal intervals of length $\Delta t = 4 \times 10^{-3}$.

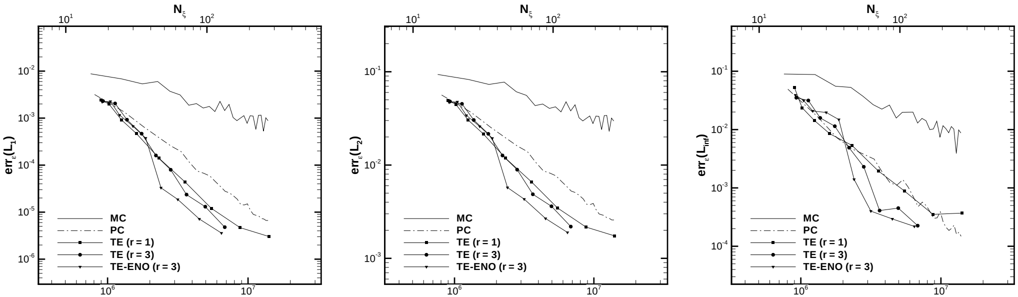
The physical assembling algorithm is the same presented for the linear advection equation in §5.3. The unique difference is constituted by the stencil $\mathbb{P}V^k_{ij}$ that is constituted by only three points: $\mathbb{P}V^k_{ij} = \left\{(v^n_{i-1\,j})^k, (v^n_{i\,j})^k, (v^n_{i+1\,j})^k\right\}$. Recalling the notation of section §5.3, the deterministic scheme is

$$(v^{n+1}_{ij})^{k+1} = \mathrm{GO}((\mathbb{P}V^k_{ij})^n, \Delta t), \tag{84}$$

where GO represents the ensemble of the operations of the first order Godunov deterministic scheme.

The reference solution in this case is the solution obtained on an uniform grid in the stochastic space with $N_\xi = 2^{21} + 1$ points with the same physical and time resolution. The error norms are computed as done for the linear case, see definitions (82).

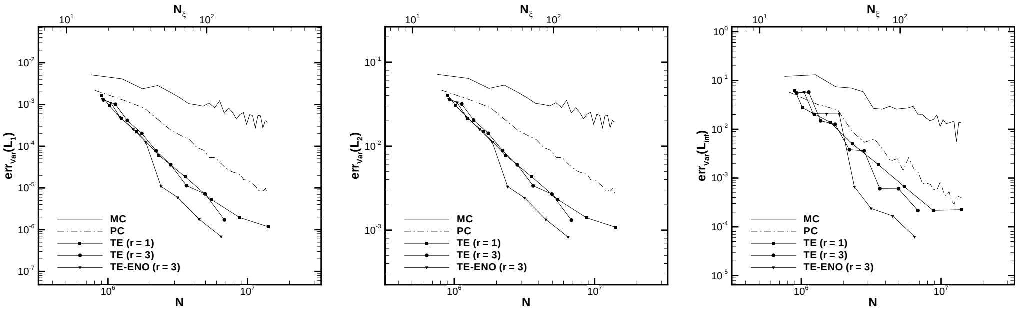
In figure 19, the error norms (82) for the expectancy of the solution u are reported (parameters for the sTE algorithm: $m_0 = 3$, m_L between 5 and 21 (with increment equal to 2) and threshold equal to $\varepsilon = 10^{-1}$). The MC and PC are applied with a number of simulations for the MC and a degree for the PC between 15 and 275 (with an increment of 10). The sTE algorithm performs better than both MC and PC. The MC method produces always the worst results, while the PC only in the case of L_∞ norms displays results comparable with the sTE algorithm and with the lower ($r = 1$) interpolation reconstruction $\mathscr{R}_k$. The higher reconstructions perform better than the lower ones. Moreover, advantages are remarkable in the case of ENO selection of the stencil.



(a) (b) (c)

Fig. 19 Errors for the expectancy $\mathbb{E}$ of the solution u (83), with norms L_1 (a), L_2 (b) and L_∞ (c) following (82). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

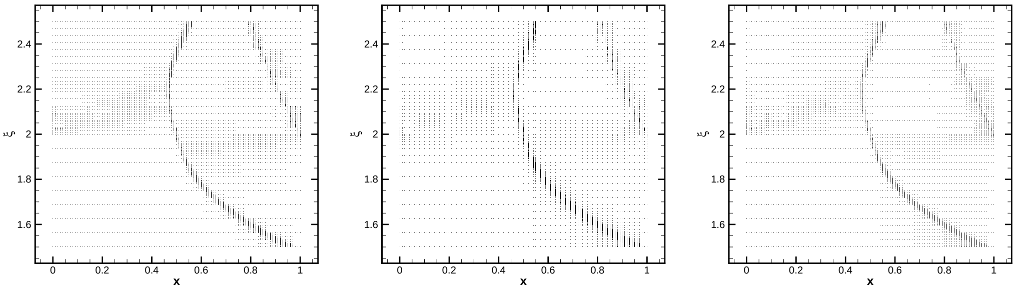
In figure 20, the error norms (82) for the variance of the solution u are reported. The parameters are the same already described for the expectancy. The sTE algorithm performs always better than the MC and PC for all the norms and reconstruction. The higher reconstruction operator $\mathscr{R}_k$ allows to attain a better accuracy with a lower number of simulation. The sTE algorithm with high reconstruction and ENO selection displays an error reduction of more than one order of magnitude with respect to the PC method at the same number of points.



(a) (b) (c)

Fig. 20 Errors for the expectancy $\mathbb{V}\mathrm{ar}$ of the solution u (83), with norms L_1 (a), L_2 (b) and L_∞ (c) following (82). The overall number of discretization N via the application of the operator $\mathscr{D}_k$ (lower axis) and the number of stochastic points N_ξ (upper axis) for each time step are reported.

In figure 21, the pattern for the distribution of points where the discretization operator is applied are reported for the three cases, $r = 1$ (Fig. 21(a)), $r = 3$ (Fig. 21(b)) and $r = 3$ with ENO selection of the stencil (see figure 21(c)) at a time $t = 2$ (parameters: $m_0 = 3$, $m_L = 15$ and $\varepsilon_k = 10^{-1}$). Main differences between the three patterns are associated to the presence of large derefined zones for higher order scheme, but consequently to a larger pollutted region near the discontinuity line. The polluted region reduces, as for the $r = 1$ case, with the introduction of the ENO stencil selection as evident in figure 21(c).



(a) (b) (c)

Fig. 21 Distribution, in the combined physical-stochastic space $\Omega - \Xi$, of the application of the discretization operator $\mathscr{D}_k$ for the TE algorithm with $r = 1$ (a), $r = 3$ (b) and with $r = 3$ and the ENO selection of the stencil (c) for the solution of the stochastic problem (83) at the time $t = 2$.

6 Concluding remarks

This paper illustrates a general extension of the Multi-resolution framework proposed by Harten to take into account uncertainty quantification in differential equations. The TE algorithms allows to obtain a multi-resolution representation of the solution not related to the knowledge of the solution at the finest level. This is well fitted to the UQ framework, where the issue is to reduce the global number of points in the stochastic space. Two formulations in terms of cell-average and point values have been provided. While in the physical space, it is common to deal with spatially cell averaged value, in the stochastic space it is more common a point value setting. Moreover, high-order reconstruction with ENO selection of the stencil in the stochastic space has been illustrated.

The HO reconstruction efficiency has been verified, first on some algebraic functions, and then on some ordinary differential equations, thus using the TE algorithm with HO. The HO-TE performs better than MC and PC, in terms of number of simulations for a prescribed level of accuracy. Moreover, the effect of HO reconstruction increases the convergence speed. Concerning the ODE, the use of an ENO selection of the stencil has been shown to slightly degrades the performances, with respect to a third-order reconstruction, for the Kraichnan-Orszag test case. This is due to a deterioration of the interpolation due to the employment of non-symmetric stencils.

Finally, the sTE algorithm has been presented in this paper for solving sPDE. This algorithm features a high-order reconstruction in the stochastic space, a weak coupling in physical and stochastic space, *i.e.* the number of points in the stochastic space is adaptive with respect to time and to the location in the physical space. Moreover, ENO selection of the stencil permits to treat properly the discontinuities propagating in the coupled physical/stochastic space.

The sTE has been applied on two sPDE, the linear advection equation and the Burgers equation. In particular, a deterministic scheme, a second order MUSCL-Hancock method (MHM) with a Roe superbee slope limiter has been coupled with sTE for the linear advection equation. For the advection, sTE works systematically better than MC and PC, but the advantage of an high-order reconstruction are less evident because the solution is formed by a series of plateaux. For Burgers equation, the sTE algorithm with high reconstruction and ENO selection displays an error reduction of more than one order of magnitude with respect to the PC method at the same number of points, thus making evident the interest of using this kind of approach for discontinuities propagating in sPDE.

Future developments will be focused on the coupling between the sTE strategy and the SI finite-volume scheme proposed in [1], and on the number of dimensions in the stochastic space.

Acknowledgements

Remi Abgrall has been partially supported by the ERC Advanced Grant ADDECCO N. 226316, while Gianluca Geraci has been fully supported by the ERC Advanced Grant ADDECCO N. 226316.

References

1. Abgrall, R., Congedo, P.M.: A semi-intrusive deterministic approach to uncertainty quantifications in non-linear fluid flow problems. Journal of Computational Physics (235), 828–845 (2013)
2. Abgrall, R., Congedo, P.M., Galéra, S., Geraci, G.: Semi-intrusive and non-intrusive stochastic methods for aerospace applications. In: 4TH EUROPEAN CONFERENCE FOR AEROSPACE SCIENCES, Saint Petersburg, Russia, July 4th-8th, 2011, 1, pp. 1–8 (2011)
3. Abgrall, R., Congedo, P.M., Geraci, G.: A One-Time Truncate and Encode Multiresolution Stochastic Framework. Journal of Computational Physics **257**, 19–56 (2014). DOI http://dx.doi.org/10.1016/j.jcp.2013.08.006
4. Abgrall, R., Harten, A.: Multiresolution Representation in Unstructured Meshes. SIAM Journal on Numerical Analysis **35**(6), 2128–2146 (1998). DOI 10.1137/S0036142997315056
5. Abgrall, R., Sonar, T.: On the use of Mühlbach expansions in the recovery step of ENO methods. Numerische Mathematik (1997), 1–25 (1997). URL http://www.springerlink.com/index/LTXHR8P0MC3QBQA7.pdf
6. Arandiga, F., Chiavassa, G., Donat, R.: Harten framework for multiresolution with applications: From conservation laws to image compression. Boletín SEMA **31**(31), 73–108 (2009). URL http://www.sema.org.es/ojs/index.php?journal=sema&amp;page=article&amp;op=view&amp;path[]=174 http://www.sema.org.es/ojs/index.php?journal=sema&page=article&op=view&path%5B%5D=174
7. Arandiga, F., Donat, R.: Nonlinear multiscale decompositions: The approach of A. Harten. Numerical Algorithms **23**, 175–216 (2000). URL http://www.springerlink.com/index/N363R0747675J70L.pdf

8. Aràndiga, F., Donat, R., Harten, A.: Multiresolution based on weighted averages of the hat function I: Linear reconstruction techniques. SIAM journal on numerical analysis **36**(1), 160–203 (1998). URL http://epubs.siam.org/doi/pdf/10.1137/S0036142996308770
9. Babuška, I., Nobile, F., Tempone, R.: A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. SIAM Review **52**(2), 317 (2010). DOI 10.1137/100786356. URL http://link.aip.org/link/SIREAD/v52/i2/p317/s1&Agg=doi
10. Bellman, R.E., Richard, B.: Adaptive Control Processes: A Guided Tour. Princeton University Press (1961). URL http://books.google.com/books?id=POAmAAAAMAAJ&pgis=1
11. Getreuer, P., Meyer, F.G.: ENO multiresolutions Schemes with General Discretizations. SIAM Journal on Numerical Analysis **46**(6), 2953–2977 (2008). URL http://epubs.siam.org/doi/pdf/10.1137/060663763
12. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements. A spectral approach. Springer Verlag (1991)
13. Graham, I., Kuo, F., Nuyens, D., Scheichl, R., Sloan, I.: Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. Journal of Computational Physics **230**(10), 3668–3694 (2011). DOI 10.1016/j.jcp.2011.01.023. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999111000489
14. Harten, A.: Eno schemes with subcell resolution. Journal of Computational Physics **83**(1), 148 – 184 (1989). DOI 10.1016/0021-9991(89)90226-X. URL http://www.sciencedirect.com/science/article/pii/002199918990226X
15. Harten, A.: Discrete multi-resolution analysis and generalized wavelets. Applied Numerical Mathematics **12**(13), 153 – 192 (1993). DOI 10.1016/0168-9274(93)90117-A. URL http://www.sciencedirect.com/science/article/pii/016892749390117A
16. Harten, A.: Adaptive multiresolution schemes for shock computations. Journal of Computational Physics **135**(2), 260–278 (1994). DOI 10.1006/jcph.1997.5713. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999197957132 http://www.sciencedirect.com/science/article/pii/S0021999184711995
17. Harten, A.: Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. Communications on Pure and Applied Mathematics **48**(12), 1305–1342 (1995). URL http://onlinelibrary.wiley.com/doi/10.1002/cpa.3160481201/abstract
18. Harten, A.: Multiresolution Representation of Data : A General Framework. SIAM Journal on Numerical Analysis **33**(3), 1205–1256 (1996)
19. Harten, A., Engquist, B., Osher, S.: Uniformly high order accurate essentially non-oscillatory schemes, III. Journal of Computational Physics **71**(2), 231–303 (1987). DOI 10.1016/0021-9991(87)90031-3. URL http://linkinghub.elsevier.com/retrieve/pii/0021999187900313
20. Le Maître, O., Knio, O.: Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics. Springer Verlag (2010)
21. Leveque, R.J.: Finite volume methods for conservation laws and hyperbolic systems. Cambridge University Press (2002)
22. Orszag, S.: Dynamical properties of truncated Wiener-Hermite expansions. Physics of Fluids **10**(12), 2603–2613 (1967)
23. Quarteroni, A., Sacco, R., Saleri, F.: Matematica Numerica. Springer (2008)
24. Toro, E.F.: Riemann solvers and numerical methods for fluid mechanics. Springer, Berlin (1997)
25. Tryoen, J.: Methodes de Galerkin stochastiques adaptatives pour la propagation d'incertitudes parametriques dans les systemes hyperboliques. Ph.D. thesis, Université Paris-Est (2011)
26. Tryoen, J., Le Maître, O., Ern, A.: Adaptive Anisotropic Spectral Stochastic Methods for Uncertain Scalar Conservation Laws. SIAM Journal Scientific Computing **34**, A2459–A2481 (2012)
27. Tryoen, J., Le Maître, O., Ndjinga, M., Ern, A.: Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems q. Journal of Computational Physics **229**, 6485–6511 (2010). URL http://dx.doi.org/10.1016/j.jcp.2010.05.007
28. Wan, X., Karniadakis, G.E.: Beyond WienerAskey Expansions: Handling Arbitrary PDFs. Journal of Scientific Computing **27**(1-3), 455–464 (2005). DOI 10.1007/s10915-005-9038-8. URL http://www.springerlink.com/index/10.1007/s10915-005-9038-8
29. Xiu, D., Karniadakis, G.E.: Modeling uncertainty in flow simulations via generalized polynomial chaos. Journal of Computational Physics **187**(1), 137–167 (2003). DOI 10.1016/S0021-9991(03)00092-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999103000925

A Cubic interpolation coefficients

In this section, some further details on the cubic interpolation used in the present work are introduced. Recalling the equation (42)

$$q_j(\xi; f,r,s) = \sum_{m=-s}^{-s+r} v_{j+m} L_m\left(\frac{\xi-\xi_j}{h}\right), \tag{85}$$

here a nested sequences of mesh is considered (see figure 2).

The aim is to evaluate the polynomial $q_j^{k-1}(\xi; f,r,s)$ of degree $r=3$, defined on the interval $[\xi_{j-1}^{k-1}, \xi_j^{k-1}]$, in the stochastic point ξ_{2j-1}^k in order to interpolate the function $f = f(\xi)$.

The first case is relative to the centered stencil $\mathscr{S}_j^{k-1} = \mathscr{S}_j^{k-1}(r=3, s=2) = \left\{\xi_{j-2}^{k-1}, \xi_{j-1}^{k-1}, \xi_j^{k-1}, \xi_{j+1}^{k-1}\right\}$.

The polynomial (85) becomes

$$q_j^{k-1}(\xi_{2j-1}^k; f,3,2) = \sum_{m=-2}^{1} v_{j+m}^{k-1} L_m \left(\frac{\xi_{2j-1}^k - \xi_j^{k-1}}{\Delta\xi} \right), \tag{86}$$

where $f(\xi_{j+m}^{k-1}) = v_{j+m}^{k-1}$ and L_m indicates the Lagrange polynomial (see (41)).

The values for the Lagrange polynomials can be evaluated for $y = \frac{\xi_{2j-1}^k - \xi_j^{k-1}}{\Delta\xi} = -\frac{1}{2}$ as follows:

$$\begin{aligned} L_{-2}(y) &= -\frac{1}{6}y(y+1)(y-1) \to L_{-2}(-\frac{1}{2}) = -\frac{1}{16} \\ L_{-1}(y) &= \frac{1}{2}y(y-1)(y+2) \to L_{-1}(-\frac{1}{2}) = \frac{9}{16} \\ L_0(y) &= -\frac{1}{2}(y-1)(y+1)(y+2) \to L_0(-\frac{1}{2}) = \frac{9}{16} \\ L_1(y) &= \frac{1}{6}y(y+1)(y+2) \to L_1(-\frac{1}{2}) = -\frac{1}{16} \end{aligned} \tag{87}$$

obtaining

$$q_j^{k-1}(\xi_{2j-1}^k; f,3,2) = \frac{1}{16}\left(-v_{j-2}^{k-1} + 9v_{j-1}^{k-1} + 9v_j^{k-1} - v_{j+1}^{k-1}\right). \tag{88}$$

The cubic polynomial $q_j^{k-1}(\xi_{2j-1}^k; f,3,2)$ (88) is used to build the reconstruction operator $\mathscr{R}_k$ for the TE algorithm (and its sTE extension) with $r = 3$. The ENO selection is used to obtain less oscillatory interpolants. For the cubic case here of interest the stencil selection is performed, employing the algorithm 1, between the stencils $\mathscr{S}_j^{k-1} = \mathscr{S}_j^{k-1}(r,s) = \left\{\xi_{j-s}^{k-1}, \ldots, \xi_{j-s+r}^{k-1}\right\}$. In the case of $s = 1$, the Lagrange polynomials evaluated in $y = \frac{\xi_{2j-1}^k - \xi_j^{k-1}}{\Delta\xi} = -\frac{1}{2}$, are the following

$$\begin{aligned} L_{-1}(y) &= -\frac{1}{6}y(y-1)(y-2) \to L_{-1}(-\frac{1}{2}) = \frac{5}{16} \\ L_0(y) &= \frac{1}{2}(y+1)(y-1)(y-2) \to L_0(-\frac{1}{2}) = \frac{15}{16} \\ L_1(y) &= -\frac{1}{2}y(y+1)(y-2) \to L_1(-\frac{1}{2}) = -\frac{5}{16} \\ L_2(y) &= \frac{1}{6}y(y-1)(y+1) \to L_2(-\frac{1}{2}) = \frac{1}{16}. \end{aligned} \tag{89}$$

Moreover, the interpolation polynomial becomes

$$q_j^{k-1}(\xi_{2j-1}^k; f,3,1) = \frac{1}{16}\left(5v_{j-1}^{k-1} + 15v_j^{k-1} - 5v_{j+1}^{k-1} + v_{j+2}^{k-1}\right). \tag{90}$$

For symmetry, it is simple to obtain the polynomial in the case $s = 3$:

$$q_j^{k-1}(\xi_{2j-1}^k; f,3,3) = \frac{1}{16}\left(v_{j-3}^{k-1} - 5v_{j-2}^{k-1} + 15v_{j-1}^{k-1} + 5v_j^{k-1}\right). \tag{91}$$

Note that in the stochastic space the periodicity of the function f cannot be considered, then, at the boundaries the stencil is always modified to be admissible: for $[\xi_0^{k-1}, \xi_1^{k-1}]$ the only possible choice is $s = 1$; for $[\xi_1^{k-1}, \xi_2^{k-1}]$ the ENO selection concerns the stencils $s = 1,2$; while symmetric considerations hold for the two intervals at the other boundary.

B The high-order MUSCL Hancock Method (MHM)

In this section the MUSCL Hancock Method (MHM) is briefly recalled. A very interesting and exhaustive presentation of this method could be found in [24].

Let us consider a 1D scalar conservation law

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f(u(x,t))}{\partial x} = 0, \tag{92}$$

where $x \in \Omega \subset \mathscr{R}$ is the physical space and $t \in T \subset \mathscr{R}^+$ is the time space. In the context of finite volume scheme the physical space is divided in a set of non-overlapping cells $\mathscr{C}_i$ with $\Omega = \bigcup_i \mathscr{C}_i$. The classical first order Godunov scheme, applied to (92), is obtained introducing the so-called cell-average $\overline{u_i}$ on each cell $\mathscr{C}_i$:

$$\overline{u_i}(t) = \frac{1}{|\mathscr{C}_i|}\int_{\mathscr{C}_i} u(x,t)\mathrm{d}x, \tag{93}$$

where $|\mathscr{C}_i|$ indicated the volume of the cell.

After the integration on each cell $\mathscr{C}_i$, it can be written

$$|\mathscr{C}_i|\frac{\mathrm{d}\bar{u}_i}{\mathrm{d}t} + f(u(x_L,t)) - f(u(x_R,t)) = 0, \tag{94}$$

with $\mathscr{C}_i = [x_{i_L}, x_{i_R}]$ and where x_{i_L} and x_{i_R} indicate the left and right interfaces.

Integrating in time the equation (94) between t_n and $t_{n+1} = t_n + \Delta t$, it follows that

$$\begin{aligned} |\mathscr{C}_i|(\bar{u}_i^{n+1} - \bar{u}_i^n) + \int_{t_n}^{t_{n+1}} f(u(x_L,t))\mathrm{d}t - \int_{t_n}^{t_{n+1}} f(u(x_R,t))\mathrm{d}t = \\ |\mathscr{C}_i|(\bar{u}_i^{n+1} - \bar{u}_i^n) + \Delta t(F_{i_L}^n - F_{i_R}^n) = 0, \end{aligned} \tag{95}$$

where a numerical approximation for the flux along the interface x_{i_L} (and x_{i_R}) holds

$$F_{i_L}^n \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{i_L},t))\mathrm{d}t. \tag{96}$$

The final form for the first order Godunov scheme is

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{|\mathscr{C}_i|}\left(F_{i_L}^n - F_{i_R}^n\right). \tag{97}$$

As pointed out by LeVeque in [21] for hyperbolic problem the information propagates with finite speed and it is reasonable to suppose that each numerical flux, at the interface, is function of the solution on the two cells to which it belongs: $F_{i_L}^n = F_{i_L}^n(\bar{u}_{i-1}^n, \bar{u}_i^n)$ and $F_{i_R}^n = F_{i_R}^n(\bar{u}_i^n, \bar{u}_{i+1}^n)$.

In this work, an exact Riemann solver is used to compute the numerical flux. In particular, given two constant state $\bar{u}_{i-1}^n$ and $\bar{u}_i^n$ separated by the interface, the exact solution of the problem, the so-called Riemann problem, can be found and the solution at the interface computed. The numerical flux is then equal to the flux function evaluated knowing the exact solution at the interface. In the following, the numerical flux function obtained via the solution of the exact Riemnn problem is indicated as $F_{i_L}^n = \mathscr{F}^{\mathrm{RM}}(u_{i-1}^n, u_i^n)$ for the left interface or $F_{i_R}^n = \mathscr{F}^{\mathrm{RM}}(u_i^n, u_{i+1}^n)$ for the right interface.

The first order Godunov scheme introduces a great amount of numerical diffusion and then val Leer [21, 24] proposed to consider non constant data on each cell to achieve a higher accuracy. From this idea, the so-called Monotone Upstream-centred Scheme for Conservation Laws (MUSCL) has been proposed. In this work, a piecewise linear approximation is used for the solution $u(x,t)$ on the cell $|\mathscr{C}_i|$:

$$u(x,t_n) = \bar{u}_i^n + \sigma_i^n(x - x_i) \quad \text{with} \quad x_{i_L} \leq x \leq x_{i_R}, \tag{98}$$

in which σ_i^n is the slope. The choice of $\sigma_i^n = 0$ lead to the Godunov scheme. Computing the slope σ_i^n as a function of only the cell averaged solution in the neighboring cells, *i.e.* $\sigma_i^n = \sigma_i^n(\bar{u}_{i-1}^n, \bar{u}_{i+1}^n)$, is not the best choice. If centered slope are used, spurious oscillations can be introduced in discontinuous solution. In practice, a slope limiter should be introduced near the discontinuity to avoid oscillations. In this work the Roe's superbee limiter is employed in which

$$\begin{cases} \sigma_i^n = \mathrm{maxmod}\left(\sigma_{(1)}^n, \sigma_{(2)}^n\right) \\ \sigma_{(1)}^n = \mathrm{minmod}\left(\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathscr{C}_i|}\right), 2\left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathscr{C}_i|}\right)\right) \\ \sigma_{(2)}^n = \mathrm{minmod}\left(2\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathscr{C}_i|}\right), \left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathscr{C}_i|}\right)\right), \end{cases} \tag{99}$$

where the minmod and maxmod are defined as follows

$$\mathrm{minmod}(a,b) = \begin{cases} a & \text{if} \quad |a| < |b| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |a| > |b| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab <= 0 \end{cases} \qquad \mathrm{maxmod}(a,b) = \begin{cases} a & \text{if} \quad |a| > |b| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |a| < |b| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab <= 0. \end{cases} \tag{100}$$

The fully discrete second order MHM, to compute the cell averaged solution $\bar{u}_i^{n+1}$, consists of the following three steps:

- For each cell $\mathscr{C}_\ell \in \{\mathscr{C}_{i-1}, \mathscr{C}_i, \mathscr{C}_{i+1}\}$ the solution at the interface is computed according to

$$\begin{cases} u_{\ell_L}^n = \bar{u}_\ell^n - \sigma_\ell^n \dfrac{|\mathscr{C}_\ell|}{2} \\ u_{\ell_R}^n = \bar{u}_\ell^n + \sigma_\ell^n \dfrac{|\mathscr{C}_\ell|}{2} \end{cases} \tag{101}$$

- On each cell $\mathscr{C}_\ell \in \{\mathscr{C}_{i-1}, \mathscr{C}_i, \mathscr{C}_{i+1}\}$ the solution is evolved using an half time step:

$$\begin{cases} u^{\Uparrow}_{\ell_R} = \bar{u}_{\ell_R} + \dfrac{1}{2}\dfrac{\Delta t}{|\mathscr{C}_\ell|}\left(f(u^n_{\ell_L}) - f(u^n_{\ell_R})\right) \\ u^{\Uparrow}_{\ell_L} = \bar{u}_{\ell_L} + \dfrac{1}{2}\dfrac{\Delta t}{|\mathscr{C}_\ell|}\left(f(u^n_{\ell_L}) - f(u^n_{\ell_R})\right) \end{cases} \tag{102}$$

- The cell averaged value on the cell $\mathscr{C}_i$ is evolved following

$$\bar{u}^{n+1}_i = \bar{u}^n_i - \frac{\Delta t}{|\mathscr{C}_i|}\left(\mathscr{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right) - \mathscr{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right)\right). \tag{103}$$

The time advancing formula is then limited to a stencil of only three cells $\mathscr{C}_{i-1}$, $\mathscr{C}_i$ and $\mathscr{C}_{i+1}$ but the computation of the slopes for the cells $\mathscr{C}_{i-1}$ and $\mathscr{C}_{i+1}$ requires (see (99)) also to know the solution on the two sourrounding cells $\mathscr{C}_{i-2}$ and $\mathscr{C}_{i+2}$. The average solution $\bar{u}^{n+1}_i$, on each cell $\mathscr{C}_i$ at time $t_{n+1} = t_n + \Delta t$, can be computed knowing the solution on the augmented stencil $\left\{\bar{u}^n_{i-2}, \bar{u}^n_{i-1}, \bar{u}^n_i, \bar{u}^n_{i+1}, \bar{u}^n_{i+2}\right\}$ that constituted the stecil obtained via the physical assembling algorithm (see algorithm 5).

Paper *P4*

International Journal for Numerical Methods in Fluids

An adaptative multiresolution semi-intrusive for UQ in compressible fluid problems

Journal:	*International Journal for Numerical Methods in Fluids*
Manuscript ID:	FLD-13-0296
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	27-Jul-2013
Complete List of Authors:	Abgrall, Remi; Universite Bordeaux I, MAB and Projet Scalapplix Congedo, Pietro; INRIA, Team Bacchus Geraci, Gianluca; Universite de Bordeaux, Institut de Mathematiques de Bordeaux; INRIA, Equipe Bacchus Iaccarino, Gianluca; Stanford University, Mechanical Engineering
Keywords:	Chaos, Compressible flow, Finite volume, Mesh adaptation, Partial differential equations, Probabilistic methods, Uncertainty quantification

INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2013; **00**:1–38
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/fld

An adaptive multiresolution semi-intrusive scheme for UQ in compressible fluid problems

R. Abgrall*[1], P.M. Congedo[1], G. Geraci[1] and G. Iaccarino[2]

[1] *INRIA Bordeaux–Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talcence CEDEX, FRANCE*
[2] *Mechanical Engineering and Institute for Computational Mathematical Engineering, Stanford University, 500 Escondido Road, Stanford, USA.*

SUMMARY

This paper deals a multiresolution strategy applied to a semi-intrusive scheme recently introduced by the authors in the context of uncertainty quantification (UQ) analysis for compressible fluids problems. The mathematical framework of the multiresolution framework is presented for the cell-average setting and the coupling with the existing semi-intrusive scheme is described from both the theoretical and practical point-of-view. Some reference test-cases are performed to demonstrate the convergence properties and the efficiency of the overall scheme: the linear advection problem for both smooth and discontinuous initial conditions, the inviscid Burgers equation and the 1D Euler system of equations to model an uncertain shock tube problem obtained by the well-known Sod shock problem. For all the cases presented, the convergence curves are computed with respect to semi-analytical solutions obtained for the stochastic formulation of the test cases. In the case of the shock tube problem, an original technique to obtain a reference high-accurate numerical stochastic solution has also been developed. Copyright © 2013 John Wiley & Sons, Ltd.

Received . . .

KEY WORDS: Finite volume, Uncertainty Quantification, linear advection, Burgers equation, Euler equations, shock tube

1. INTRODUCTION AND MOTIVATION

In recent years, the scientific numerical community faced a new challenge, the effect and propagation of uncertain parameters in the numerical models. Nowadays, the attention is focused not only on the accurate solution of the equations, but also on the effect of uncertain parameters in boundary or initial conditions and in the model.

Among the non-intrusive approaches, *i.e.* where uncertainties are quantified practically by making multiple calls to a deterministic code, several methods are commonly employed: Monte Carlo family of techniques [1], the collocation family [2] and the non-intrusive Galerkin projection methods. This last family of methods has been introduced for the first time by Ganem and Spanos [3] for the analysis of structural dynamics systems and has been generalized by Xiu and Karniadakis [4] to general probability distributions. Actually, the non-intrusive Galerkin projection represents the state-of-the art of the stochastic analysis for systems with a smooth response surface due to its spectral convergence property.

The Galerkin projection is also the most important technique in order to manage intrusively the uncertainty propagation into a numerical code. In practice, this means that it is possible to obtain an equivalent set of governing equations for the coefficients of a truncated polynomial representation of the quantities of interest [5]. Then, the number of equations is related to the number of coefficients

*Correspondence to: INRIA Bordeaux–Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talcence CEDEX, FRANCE employed in the polynomial expansion, and the numerical code should be deeply modified. In many cases, this leads to complex problems regarding the generality of the approach when *ad hoc* solvers are proposed [6]. More recently, Abgrall and Congedo proposed a novel semi-intrusive approach that extend in a straightforward and natural way, the representation of the variables in the physical space also along the stochastic space [7]. This approach leads to a very flexible scheme able to handle whatever form of probability density function even time varying and discontinuous. One of the prominent advantage of this kind of approach is the possibility to extend in an easier way an existing deterministic code to its stochastic counterparts.

Following the general idea of a semi-intrusive propagation of the uncertainties, recently, Abgrall et al. [8, 9, 10] introduced a point-value setting in the multiresolution (MR) framework to represent data in the stochastic space. The multiresolution representation of data permits to increase the efficiency of the numerical code for the solution of stochastic partial differential equations. The idea of introducing the MR representation of data, in the context of stochastic problem, is not totally new. In [5], a multiresolution basis is employed to represent the solution of a partial differential equations after fixing the physical coordinate. This representation is very efficient but limited to the case where the stochastic representation is used at a fixed physical location. To overcome this issue, more recently, Tryoen et al. introduced in [6] a multiresolution wavelets representation in the context of intrusive Galerkin projections. However, the Galerkin approach presented remains very problem-dependent. In fact, using a Roe-type solver demands the computation of the eigenstructure of the Roe matrix explicitly; this can be very complex. Moreover, *ad hoc* entropy fix should be adopted, thus increasing the numerical cost associated to the representation of discontinuous solution [11]. This original approach has been further improved to obtain a more efficient scheme employing a multiresolution adaptive strategy [12]. However, this approach is limited by the spatial and time discretization accuracy (only first order) that could dominate the overall accuracy. Moreover, the approach proposed by Abgrall et al [8, 9, 10] has the advantage to remain very general, not limited from the order of the spatial and time discretization, from the probability density function (that can be even discontinuous and time varying) and, eventually, from the geometry of the stochastic space in the case of multidimensional problems.

In this paper, the MR is extended to the cell-average framework and the representation is implemented in the semi-intrusive scheme [7]. Thanks to its intrinsic capability to manage discontinuous responses, the semi-intrusive methods represents a promising alternative to the Galerkin projection techniques for all the applications where the system is dominated by shocks, as for example in computational fluid dynamics for transonic flows.

In this paper, we demonstrate the advantages of the introduction of a real-time adaptivity in the stochastic space, by following the evolution of the solution in the overall physical and stochastic space. This is shown by comparing the accuracy, at a fixed computational cost, with and without the adaptivity based on the MR framework on the original SI scheme. Different reference test-cases are performed for which the reference solution can be obtained in an analytical or semi-analytical approach.

This paper is organized as follows. In section 2, the mathematical setting for the stochastic differential equation is given. Section 3 illustrates the multi-resolution framework of Harten, generalized for the stochastic space, where a cell-average setting is chosen. In particular the Truncate and Encode algorithm is presented in section 3.1 where the representation of the discrete data is obtained from the coarsest level towards the finest. The semi-intrusive scheme is briefly sketched in section 4 where the formulation is detailed for the MUSCL-Hancock method. The overall formulation of the adaptive semi-intrusive scheme is presented in 5. Several numerical results are presented in section 6. In particular, the introduction of the adaptive representation of data in the stochastic space is demonstrated to improve the spatial convergence and to cure the staircase approximation phenomenon with respect to an equivalent not adapted solution. The linear advection equation, the inviscid Burgers equation and an uncertain version of the Sod shock tube are performed as test-cases. Concluding remarks are reported in §7.

DOI: 10.1002/fld

2. UNCERTAINTY QUANTIFICATION FOR PARTIAL DIFFERENTIAL EQUATIONS

In this section, we introduce the mathematical setting, used for the UQ analysis in the context of partial differential equations. Let us consider an output of interest $u(\mathbf{x}, t, \boldsymbol{\xi}(\omega))$ depending from the physical space $\mathbf{x} \in \Omega \subset \mathbb{R}^{n_d}$, the time $t \in T$ and a vector of parameters $\boldsymbol{\xi} \in \Xi$ where Ξ is the sample space. The output of interest u can be a conserved (or primitive, or another flow variable) variable of a system of conservation laws.

We suppose that the output of interest is governed by an algebraic or differential operator $\mathcal{L}$ with a source term $\mathcal{S}$:

$$\mathcal{L}(\mathbf{x}, t, \boldsymbol{\xi}(\omega); u(\mathbf{x}, t, \boldsymbol{\xi}(\omega))) = \mathcal{S}(\mathbf{x}, t, \boldsymbol{\xi}(\omega)). \tag{1}$$

Initial and boundary conditions, that could depend from the parameter vector $\boldsymbol{\xi}$, should be provided for a well-posed problem. Both the operators $\mathcal{L}$ and the source term $\mathcal{S}$ are defined on the domain $\Omega \times T \times \Xi$.

Let us define a measurable space (Ξ, Σ, p) where Σ is its $\sigma-$algebra of events and p a probability measure with the following properties:

- $p(A) \geq 0$ for all $A \in \Sigma$;
- Countable additivity: if $A_i \in \Sigma$ are disjoint sets then $p(\bigcup_i A_i) = \sum_i p(A_i)$;
- as probability measure p is normalized on Ξ: $p(\Xi) = 1$.

The $\mathbb{R}^d-$valued random variable $\boldsymbol{\xi}$ specifies a set of events with a corresponding probability. More formally, the random variable $\boldsymbol{\xi}$ is a measurable function that maps the measurable space (Ξ, Σ, p) to another measurable space, the Borel $\mathcal{B}^d$ $\sigma-$algebra of the real space $(\mathbb{R}^d, \mathcal{B}^d, \mathbb{P})$. There is some set of events ω, that $\boldsymbol{\xi}$ maps to an output event $A \in \mathcal{B}^d$ with the probability of occurrence of A, $\mathbb{P}(A)$ equal to the probability of ω:

$$\mathbb{P}(A) = p(\boldsymbol{\xi}^{-1}(A)) = p(\omega : \boldsymbol{\xi}(\omega) \in A). \tag{2}$$

As usual in the literature, we consider that $\mathbb{P}(A) = p(\boldsymbol{\xi} \in A) = p(\boldsymbol{\xi})$.

The aim of UQ analysis is to find statistical quantities of the solution $u(\mathbf{x}, t, \boldsymbol{\xi})$, the statistical moments or the probability distribution.

Assuming $u(\boldsymbol{\xi}) \in L_2(\Xi, p)$, mean and variance can be computed as follows:

$$\begin{aligned} \mathcal{E}(u, \mathbf{x}, t) &= \int_{\Xi} u(\mathbf{x}, t, \boldsymbol{\xi}) p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} \\ \mathrm{Var}(u, \mathbf{x}, t) &= \int_{\Xi} \left(u(\mathbf{x}, t, \boldsymbol{\xi}) - \mathcal{E}(u)\right)^2 p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}. \end{aligned} \tag{3}$$

3. THE CELL-AVERAGE MULTIRESOLUTION SETTING

In this section, the multiresolution framework in a cell-averaged representation of data is presented. The original Harten's framework [13, 14, 15, 16] is here modified to allow an efficient representation of data with respect to a general weighted function. In the context of UQ, the weighted function is easily identified as the probability distribution of the input parameters.

In this paper, only the cell-average framework is analyzed and this choice allows a straightforward extension of the finite volume representation of data in the coupled physical/stochastic space as already shown in [7] employing only uniform meshes in both spaces (see later §4).

The Harten framework can be considered, as pointed out by Aràndiga and Donat in [17], as a rearrangement of the information in a set of discrete data representing different resolution levels. This rearrangement of data with the addition of a truncation procedure could yield a reduction of the computational cost and of the memory usage associated to the representation/calculation and memorization of discrete data.

The Harten framework can be viewed as a more general framework with respect to the classical wavelets framework in which the hierarchical representation of data is obtained by means of a functional basis based on a dilating equation and a so called mother wavelets. As presented in [18] the dilating equation in a general space can be difficult to solve, especially for domains of complex geometries. The Harten framework is capable to avoid the solution of a dilating equation obtaining a local polynomial basis for general geometries with, eventually, data-dependent strategies for the representation of data. All this features makes the Harten framework, an optimal starting point for the development of a general framework for the representation of data.

Two building blocks exist: a discretization operator $\mathcal{D}_k$ and a reconstruction operator $\mathcal{R}_k$. Both operators operates between the continuous space to represent (the stochastic space in this context) and one of its discrete representation, for instance the resolution level $k-$th. The knowledge of the these two operators allow to define in an unique way two other operators working on data rearrangement between different resolution levels. These discrete operators between consecutive levels k (higher resolution) and $k-1$ (lower resolution) are the operators of decimation D_k^{k-1} and prediction P_{k-1}^k.

In this paper, we consider the cell-average framework. Let us consider a function $f = f(\boldsymbol{\xi})$, $f : \Xi \subset \mathbb{R}^d \rightarrow \mathbb{R}$ with d the number of uncertain parameters. In the classical MR cell-average framework, $f \in \mathcal{F}$ where $\mathcal{F}$ is the functional space of the absolutely integrable functions $\mathcal{F} = L^1(\Xi)$. However, in the context of UQ, $\mathcal{F}$ is identified with L^2 to deal with function with finite variance. Let us consider the probability density function $p(\boldsymbol{\xi})$ and let us define the following measure:

$$\mathrm{d}\mu(\boldsymbol{\xi}) = p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}. \tag{4}$$

If the stochastic space is represented by means of a non-overlapping tessellation

$$\Xi = \bigcup_{j=1}^{N_\xi} \Xi_j, \quad \text{with} \quad \Xi_i \cap \Xi_j = 0 \quad \text{if} \quad i \neq j. \tag{5}$$

the measure of each element of the tessellation can be found as follows

$$\mu(\Xi_j) = \int_{\Xi_j} \mathrm{d}\mu(\boldsymbol{\xi}). \tag{6}$$

Let us consider a set of discrete operators of discretization $\{\mathcal{D}_k\}_{k=0}^L$, each of them defined on a vectorial space of finite dimension

$$\mathcal{D}_k : \quad \mathcal{F} \rightarrow V_k \quad \text{with} \quad \dim(V_{k+1}) > \dim(V_k) = J_k. \tag{7}$$

The sequence $\{\mathcal{D}_k\}_{k=0}^L$ has to be nested according to the following properties:

- $\mathcal{D}_k$ is onto
- the null space of each level include the null space associated to the previous resolution level $\mathcal{N}(\mathcal{D}_k) \subset \mathcal{N}(\mathcal{D}_{k+1})$.

These properties reflect in the following relation between discretization operators

$$\mathcal{D}_{k+1}(f) = 0 \Rightarrow \mathcal{D}_k(f) = 0 \quad \forall f \in \mathcal{F}. \tag{8}$$

A such operator on the k-th level can be defined over the j-th cell Ξ_j^k as

$$(\mathcal{D}_k f)_j \overset{\text{def}}{=} \frac{1}{\mu(\Xi_j^k)} \int_{\Xi_j^k} f(\boldsymbol{\xi})\mathrm{d}\mu(\boldsymbol{\xi}) = v_j^k. \tag{9}$$

Thanks to the onto property of each operator $\mathcal{D}_k$, the reconstruction operator $\mathcal{R}_k$ can be defined as its right-inverse

$$\mathcal{R}_k : \quad V_k \rightarrow \mathcal{F}. \tag{10}$$

The reconstruction operator is not required to be linear and this makes the Harten's multiresolution more general with respect to the wavelets framework [19].

The reconstruction operator $\mathcal{R}_k$ for the cell average setting originally has been introduced by Harten in the 1D case employing the concept of reconstruction via primitive function. In practice, the cell-averaged function is replaced by a point valued function that corresponds to its primitive in the nodes of the mesh. A more convenient approach can be adopted, following Abgrall and Sonar [20], even for multidimensional problems on unstructured meshes [16]. Fixed a polynomial degree of reconstruction r, a stencil $\mathcal{S}_j^k$ of cells with cardinality $s = s(r) = \text{card}(\mathcal{S}_j^k)$ can be fixed. On each stencil $\mathcal{S}_j^k$, a polynomial $\mathcal{P}_j^k(\boldsymbol{\xi}; f)$ of degree r can be constructed. The admissibility of this stencil obeys to a Vandermonde condition (see for further details [20]). Supposing the stencils admissible, the conditions to satisfy for the computation of $\mathcal{P}_j^k$ is

$$\mathcal{D}_k(\mathcal{P}_j^k(\boldsymbol{\xi}; f))_l = \mathcal{D}_k(f)_l, \quad \forall l \in \mathcal{S}_j^k. \tag{11}$$

The reconstruction operator $\mathcal{R}_k$ in this case is exactly equal to the union of all the polynomial $\mathcal{P}_j^k$ defined on all the cells Ξ_j^k.

The two operators $\mathcal{D}_k$ and $\mathcal{R}_k$ should satisfy a consistency relationship between them

$$(\mathcal{D}_k\mathcal{R}_k)(v) = v \quad \forall v \in V_k, \tag{12}$$

thus implying $\mathcal{D}_k\mathcal{R}_k = \mathrm{I}_k$ where I_k is the identity operator on V_k.

For the nested sequence whose elements are defined in (7), the decimation operator D_k^{k-1} can be defined, which is a linear mapping between V_k onto V_{k-1}:

$$\mathrm{D}_k^{k-1}: \quad V_k \to V_{k-1}, \tag{13}$$

where

$$\mathrm{D}_k^{k-1}v^k = \mathcal{D}_{k-1}f \in V_{k-1} \quad \forall v^k = \mathcal{D}_k f \in V_k. \tag{14}$$

The decimation operator, independent from the particular f, is employed to generate recursively the set of discrete data from the highest resolution level ($k = L$) to the lowest ($k = 0$) $\{v^k\}_{k=0}^{L-1}$,

$$v^{k-1} = \mathrm{D}_k^{k-1}v^k \quad \forall k = L, L-1, \dots, 1. \tag{15}$$

By an agglomeration (splitting) procedure, for a generic mesh, even non structured, it is always possible to obtain a less (higher) resolution level. To each cell Ξ_j^k at the lower resolution level corresponds a number of cell ($\bar{l}_c$) at the higher resolution level. To preserve the nested character between levels, the following properties between meshes should hold:

$$\Xi_j^k = \sum_{l}^{\bar{l}_c} \Xi_l^{k+1}. \tag{16}$$

In the following, without loss of generality, $\bar{l}_c = 2$. This happens naturally for the 1D case of equally splitted cells between levels in the case of regular nested meshes.

In this case, the decimation operator (see figure 1) could be obtained as follows

$$\begin{aligned}(\mathrm{D}_k^{k-1}v^k)_j = (\mathrm{D}_k^{k-1}\mathcal{D}_k f)_j = (\mathcal{D}_{k-1}f)_j &= \frac{1}{\mu(\Xi_j^{k-1})}\int_{\Xi_j^{k-1}} f(\boldsymbol{\xi})\mathrm{d}\mu(\boldsymbol{\xi}) \\ &= \frac{1}{\mu(\Xi_j^{k-1})}\left(\mu(\Xi_{2j}^k)(\mathcal{D}_k f)_{2j} + \mu(\Xi_{2j-1}^k)(\mathcal{D}_k f)_{2j-1}\right).\end{aligned} \tag{17}$$

Moreover, the prediction P_{k-1}^k allows to approximate the set of data v^k from v^{k-1}

$$v^k = \mathcal{D}_k f \approx \mathcal{D}_k(\mathcal{R}_{k-1}v^{k-1}). \tag{18}$$

DOI: 10.1002/fld

This leads to the definition of the prediction operator P^k_{k-1} between discrete data on successive resolution level as

$$\mathrm{P}^k_{k-1} \overset{\text{def}}{=} \mathcal{D}_k \mathcal{R}_{k-1} : \quad V^{k-1} \to V^k. \tag{19}$$

The prediction operator P^k_{k-1} is obtained following the definition (19) and using first the reconstruction procedure (11) for the level $k-1$th, and then applying the discretization operator $\mathcal{D}_k(\mathcal{P}^{k-1}_j)$ relative to the level k.

A consistency property can be defined, $\mathrm{D}^{k-1}_k \mathrm{P}^k_{k-1} = \mathrm{I}_k$, that follows from

$$v^{k-1} = \mathrm{D}^{k-1}_k v^k = \mathrm{D}^{k-1}_k \mathcal{D}_k f = \mathrm{D}^{k-1}_k \mathcal{D}_k \mathcal{R}_{k-1} v^{k-1} = \mathrm{D}^{k-1}_k \mathrm{P}^k_{k-1} v^{k-1}. \tag{20}$$

The last element of the MR framework is constituted by the prediction error e^k

$$e^k \overset{\text{def}}{=} v^k - \mathrm{P}^k_{k-1} v^{k-1} = (\mathrm{I}_k - \mathrm{P}^k_{k-1}\mathrm{D}^{k-1}_k) v^k. \tag{21}$$

The prediction error satisfies (from the consistency property (20))

$$\mathrm{D}^{k-1}_k e^k = \mathrm{D}^{k-1}_k (v^k - \mathrm{P}^k_{k-1} v^{k-1}) = v^{k-1} - v^{k-1} = 0, \tag{22}$$

then it is in the null space of the decimation operator $e^k \in \mathcal{N}(\mathrm{D}^{k-1}_k)$. Using the definition (13) and applying the rank theorem, it is possible to write

$$\dim(V_k) = \dim(\mathcal{N}(\mathrm{D}^{k-1}_k)) + \dim(V_{k-1}) \to \dim(\mathcal{N}(\mathrm{D}^{k-1}_k)) = \dim(V_k) - \dim(V_{k-1}) = J_k - J_{k-1}. \tag{23}$$

The linear independent coordinates of e^k are called wavelets or details d^k. Two operators can be defined to link the prediction error to the details, E^k and G^k, as follows

$$e^k \overset{\text{def}}{=} E^k d^k, \quad d^k \overset{\text{def}}{=} G^k e^k \quad \text{with} \quad E^k G^k : V^k \to \mathcal{N}(\mathrm{D}^{k-1}_k). \tag{24}$$

Using all the operators described in this section, a multi-resolution representation of data can be defined.

This is obtained by two procedure: the *encoding* and the *decoding*. The *encoding* moves from the highest resolution level to the lowest one applying recursively (for all $k = L, \dots, 1$) the decimation operator and computing the details

$$\begin{cases} v^{k-1} &= \mathrm{D}^{k-1}_k v^k \\ d^k &= G_k(\mathrm{I}_k - \mathrm{P}^k_{k-1}\mathrm{D}^{k-1}_k) v^k. \end{cases} \tag{25}$$

The multi-resolution representation v_{MR} refers to the possibility to obtain a one-to-one correspondence between the highest resolution level v^L and the sequence of the details d^k in addition to the lowest resolution level v^0:

$$v_{\mathrm{MR}} \overset{\text{def}}{=} \{v^0, d^1, \dots, d^L\}. \tag{26}$$

The *decoding* procedure is the dual procedure with respect to the *encoding*: recursively moves from the lowest resolution level v^0 together with the prediction error e^k for all the levels $k = 1, \dots, L$

$$v^k = \mathrm{P}^k_{k-1} v^{k-1} + E_k d^k. \tag{27}$$

Ideally, *decoding* and *encoding* permit an ideal exchange of information among different resolution levels. In order to be useful, these operations are coupled with an operator of data truncation. This additive operator allows, under a certain tolerance, to eliminate the over abundant information. The compression capability opens several possibilities to the application of the multi-resolution framework to compress the data as, for instance, in the signal/image representation schemes [18] or as a fundamental brick in the solution of intrinsically multi scales problems, as demonstrated already in the first seminal works of Harten [14, 15].

The truncation is instead based on the elimination of the wavelets d^k under a prescribed tolerance. The problem statement is the following: given a sequence of scale coefficients or wavelets for a fixed level d^k and assigned a level dependent tolerance criterion ε_k, we should generate $\hat{d}^k = \left\{\hat{d}_j^k\right\}_{j=1}^{J_k - J_{k-1}}$ according to

$$\hat{d}_j^k = \mathrm{tr}(d_j^k, \varepsilon_k) = \begin{cases} 0 & |d_j^k| \leq \varepsilon_k \\ d_j^k & \text{otherwise.} \end{cases} \tag{28}$$

Different choices exist in literature for the threshold parameter ε_k: a level independent choice $\varepsilon_k = \varepsilon$ or a dependent criterion $\varepsilon_k = \varepsilon/2^{L-k}$. Since the original work of Harten, the stability of the MR representation of the data has been studied. Harten proposed [13] to modify the *encoding* procedure in order to preserve the following condition

$$||v^L - \hat{v}^L|| \leq C\varepsilon, \tag{29}$$

with a constant C and measured in some norms as the L^1 and L^∞.

In this work, the main contribution is to adapt this framework performing the one-time encoding and truncated procedure in order to obtain a compact representation of the data in the stochastic space. This fundamental brick of the algorithm is described in the following section.

3.1. A one-time truncate and encode cell-average representation

In this section, the truncate and encode TE algorithm is described in the case of cell-average quantities. The pivotal idea of the algorithm is to identify in the prediction error e^k at a certain $k-$th level, a measure of the quality of the predictor operator P_{k-1}^k.

From classical interpolation results (see for instance [21]), note that the interpolation error diminishes, moving from a coarser level to a finer one, with respect to the local regularity of the function and to the local polynomial order of the interpolation. On the contrary, in presence of discontinuities, the error remains constant and of the order $\mathcal{O}[1]$. This means that, starting from the knowledge of a fine level k (using the discretization operators $\mathcal{D}_k$), the recursive combinations of prediction operations via the operators P_{k-1}^k and evaluations of the error e^k permits to determine the region, where the solution respects a certain accuracy criterion. In particular, if the criterion is equal to the truncation operation described above, at the end of the algorithm, the discretized set of data $\{v^k\}_{k=0}^L$ is directly related to the data $\{\hat{v}^k\}_{k=0}^L$ obtained under the same truncation criterion by the classical MR framework.

The algorithm starts with the definition of the coarsest level of resolution $k = 1$. On this level the discretization operator is applied obtaining the discrete data v^1: $v^1 = \mathcal{D}_1 f$. By decimation, it is also possible to obtain the discrete data on the level $k = 0$ knowing only v^1:

$$v^0 = \mathrm{D}_1^0 v^1. \tag{30}$$

An *encoding* step (analogous to what is normally done in the classical MR (see (25))) is then completed, by computing the linear independent coefficients d^k of e^k for $k = 1$:

$$d^k = G_k(\mathrm{I}_k - \mathrm{P}_{k-1}^k \mathrm{D}_k^{k-1}) v^k. \tag{31}$$

The truncation is applied on d^1 with respect to the threshold ε, defined by the user, and to the relation $\varepsilon_k = \varepsilon_k(\varepsilon, k)$:

$$\hat{d}^1 = \mathrm{tr}(d^1, \varepsilon_k). \tag{32}$$

This operation relies on the knowledge of the finest level ($k = L$), where the threshold is always equal to ε (see (28)). The integer $k = L$ is assigned to the finest level if the coarsest is marked as $k = 0$ and at each refinement k is increased by one.

The data d^1 are analyzed in order to locate the region of the domain, where the accuracy of the prediction, via P_{k-1}^k, is not adequate. This is accomplished in a very simple way after the truncation,

DOI: 10.1002/fld

by identifying the non-zero wavelets d_j^1. At each non-zero (truncated) wavelets, corresponds a region where the knowledge of the solution is not sufficient under the criterion used in the truncation (32). Then, further information are added. In particular, after the generation of the mesh on the level $k=2$, on all the cells/points inside the regions (at level $k=0$) used to generate the corresponding wavelets d_j^1, the discretization operator $\mathcal{D}_2$ is applied. On the contrary, in the region marked as well-described, the *decoding* procedure is performed:

$$v^2 = \mathrm{P}_1^2 v^1 + E_2 d^2 \simeq \mathrm{P}_1^2 v^1. \tag{33}$$

The assumption in the equation (33) means that for every null wavelets at a level $k-1$, the corresponding wavelets at level k are null too. In the case of non null details, the equation (33) is not applied, but substituted by a direct (exact) discretization of the function by means of the operator $\mathcal{D}_k$ for $k=2$.

Knowing v^2 and v^1, the *encoding* is performed by computing d^2 and their truncated counterpart $\hat{d}^2$ by (28). The algorithm is then repeated until reaching the finest level L or a full satisfactory prediction, *i.e.* $d_j^k = 0$ for all $j = 1, \dots, J_k - J_{k-1}$.

To make things clear, the algorithm is now presented in the case of 1D stochastic space. Some preliminary operation are first performed:

- Generation of a nested set of meshes $\mathcal{G}^k$ for $k = 0, \dots, L$ (0 is the coarsest mesh):

$$\mathcal{G}^k = \left\{\Xi_j^k\right\}_{j=1}^{J_k} \quad \text{where} \quad \Xi_j^k = [\xi_{j-1}^k, \xi_j^k]. \tag{34}$$

  In this case the case of bounded probability density function is addressed and a topological tessellation for the mesh can be obtained, *i.e.* each cell has the same Lebesgue measure equal to $1/J_k$. Otherwise, in the case of unbounded pdf, the set of meshes can be built on a nested sequence of cells with the same probability measure $\mathrm{d}\mu$.

- Definition of the operator $\mathcal{D}_k$, $\mathcal{R}_k$, D_k^{k-1} and P_{k-1}^k according to §3:

$$\begin{cases} (\mathcal{D}_k f(\xi))_j = \dfrac{1}{\mu(\Xi_j^k)} \displaystyle\int_{\Xi_j^k} f(\xi) p(\xi) \mathrm{d}\xi \\ \mathcal{R}_k : (\mathcal{D}_k \mathcal{R}_k v^k)_l = (\mathcal{D}_k f(\xi))_l \quad \text{with} \quad l \in \mathcal{S}_j^k \\ (\mathrm{P}_{k-1}^k v^{k-1})_j = (\mathcal{D}_k \mathcal{R}_{k-1} v^{k-1})_j = \dfrac{1}{\mu(\Xi_j^k)} \displaystyle\int_{\Xi_j^k} \mathcal{R}_{k-1} v^{k-1} p(\xi) \mathrm{d}\xi. \end{cases} \tag{35}$$

  The decimation operator can be defined when the topological relation between the cells at two different resolution levels is known. Let us consider the situation sketched in figure 1. We assume that the cells generated by the splitting of Ξ_j^{k-1}, are named as Ξ_{2j-1}^k and Ξ_{2j}^k even if this numeration does not correspond to the index j of the generating stochastic cell at the lower resolution level. The indexes numeration in figure 1 is exactly matched only if all the cells are splitted from a resolution level to the higher one. In that case, the dimensions of the spaces of the two levels $k-1$ and k are related by the following relation, $J_k/J_{k-1} = 2$. In the following, the abstract indexes $2j$ and $2j-1$ are employed to make evident the dependence of the two cells, at level k, from the generating cell Ξ_j^{k-1}. However, the indexes should always be intended in the sense described above. When a cell is split to obtain the higher resolution level (see figure 1), the measure $\mathrm{d}\mu$ is defined as follows:

$$\begin{cases} \mu(\Xi_j^{k-1}) = \mu(\Xi_{2j-1}^k) + \mu(\Xi_{2j}^k) \\ \mu(\Xi_{2j-1}^k) = \mu(\Xi_{2j}^k). \end{cases} \tag{36}$$

  Then, the decimation operator is simply obtained as

$$(\mathrm{D}_k^{k-1} v^k)_j = v_j^{k-1} = \frac{1}{\mu(\Xi_j^{k-1})} \left(\mu(\Xi_{2j}^k) v_{2j}^k + \mu(\Xi_{2j-1}^k) v_{2j-1}^k\right) \tag{37}$$

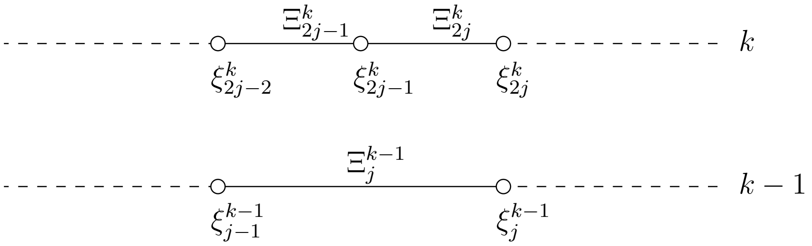


Figure 1. Example of 1D stochastic nested meshes for the cell-average setting decimation procedure.

- Setting a proper threshold ε and a proper relation for $\varepsilon_k = \varepsilon_k(\varepsilon, k; L)$
- Discretization of the level $k = 1$: $(v^1) = (\mathcal{D}_1 f)$;
- Decimation of the discrete data v^1 to obtain $(v^0) = (\mathrm{D}^0_1 v^1)$.

The TE algorithm for cell-average setting in 1D stochastic space can be explicitly written as:

Algorithm 1: Truncate and Encode algorithm for the cell average setting in 1D stochastic space.

```
while $2 \leq k \leq L$ do
    for $j = 1, \dots, J_{k-2}$ do
        Encoding:
        $(d^{k-1})_j = v^{k-1}_{2j} - (\mathrm{P}^{k-1}_{k-2} v^{k-2})_{2j} = v^{k-1}_{2j} - \left( \frac{1}{\mu(\Xi^{k-1}_{2j})} \int_{\Xi^{k-1}_{2j}} \mathcal{R}_{k-2} v^{k-2} p(\xi) \mathrm{d}\xi \right)$ ;
        Truncation:   $\hat{d}^{k-1}_j = \mathrm{tr}(d^{k-1}_j, \varepsilon_{k-1})$ ;
    end
    for $j = 1, \dots, J_{k-1}$ do
        if $\hat{d}^{k-1}_j > 0$ then
            Discretization: $v^k_{2j} = (\mathcal{D}_k f)_{2j} = \frac{1}{\mu(\Xi^k_{2j})} \int_{\Xi^k_{2j}} f(\xi) p(\xi) \mathrm{d}\xi$ ;
            Discretization: $v^k_{2j-1} = (\mathcal{D}_k f)_{2j-1} = \frac{1}{\mu(\Xi^k_{2j-1})} \int_{\Xi^k_{2j-1}} f(\xi) p(\xi) \mathrm{d}\xi$ ;
        end
    end
end
```

At this level, remark that the sequence of discretization operators should be nested and $\mathcal{N}(\mathcal{D}_k) \subset \mathcal{N}(\mathcal{D}_{k+1})$. This means that the error vector e^k can be represented by means of only its independent components, the wavelets d^k, thanks to the relation (24). It is always possible to write, recalling the definition of the error vector e_k (21) and the nested property of the discretization operator (17), as follows

$$
\begin{aligned}
e^k_{2j-1} &= v^k_{2j-1} - (\mathrm{P}^k_{k-1} v^{k-1})_{2j-1} \\
&= \frac{1}{\mu(\Xi^k_{2j-1})} \left(\mu(\Xi^{k-1}_j) v^{k-1}_j - \mu(\Xi^k_{2j}) v^k_{2j} \right) - \frac{1}{\mu(\Xi^k_{2j-1})} \left(\mu(\Xi^{k-1}_j) v^{k-1}_j - \mu(\Xi^k_{2j}) (\mathrm{P}^k_{k-1} v^{k-1})_{2j} \right) \\
&= \frac{\mu(\Xi^k_{2j})}{\mu(\Xi^k_{2j-1})} \left((\mathrm{P}^k_{k-1} v^{k-1})_{2j} - v^k_{2j} \right) = -\frac{\mu(\Xi^k_{2j})}{\mu(\Xi^k_{2j-1})} d^k_j .
\end{aligned}
\tag{38}
$$

The first loop should be performed in order to compute all the *wavelets* d^k_j, while the second loop is performed over the whole set of cells belonging to the resolution level. In particular, the error vector component is compared with the threshold for deciding whether the discretization via the model evaluation is necessary. In the second loop, in the case of a nested sequence, with splitting based on the probability measure, the local error is equal to *wavelet* computed over the same cell Ξ^{k-1}_j (see equation (38)). Therefore, the truncated wavelet is exactly equal to the truncated component of the error.

DOI: 10.1002/fld

In the classical framework, the first step is the *encoding* procedure moving from the finest level to the coarsest. In this case, the explicit evaluation of the function f is performed at the finest level while the other levels are obtained by agglomeration. In the present paper, the encoding is performed proceeding from the coarsest level. Each time a higher resolution level is added, *i.e* k, the function is explicitly evaluated via the discretization operator $\mathcal{D}_k$. Due to numerical errors, the relation (14) could not hold. In such a case, the wavelets d^k are not the linear independent components of the error vector e^k. For representing the error vector in terms of its independent components d^k, the Discetrize Agglomerate Decimate (DAD) algorithm is introduced. The DAD algorithm consists in the following operations

Algorithm 2: DAD algorithm.

```
Discretization:
    $v_{2j}^k = \frac{1}{\mu(\Xi_{2j}^k)} \int_{\Xi_{2j}^k} f(\xi)p(\xi)\mathrm{d}\xi$ ;
    $v_{2j-1}^k = \frac{1}{\mu(\Xi_{2j-1}^k)} \int_{\Xi_{2j-1}^k} f(\xi)p(\xi)\mathrm{d}\xi$ ;
Agglomeration:
    $\mu(\Xi_j^{k-1}) = \mu(\Xi_{2j-1}^k) + \mu(\Xi_{2j}^k)$ ;
Decimation:
    $(\mathrm{D}_k^{k-1} v^k)_j = v_j^{k-1} = \frac{1}{\mu(\Xi_j^{k-1})} \left(\mu(\Xi_{2j}^k) v_{2j}^k + \mu(\Xi_{2j-1}^k) v_{2j-1}^k\right)$
```

The DAD algorithm should be always performed before the *Encoding* in the TE algorithm 2. The introduction of the DAD algorithm is a peculiarity of the cell-average framework, while the point-value setting does not require any similar procedure because two successive levels are constituted by a set of points in the intersection of the two spaces.

Another peculiarity of the cell average framework is the presence of integral quantities that requires different evaluation in each cell, according to the numerical rule used to obtain the integrals in the discretization operator (9). The family of Newton-Cotes formula, employing only equally spaced points, is the best choice in term of computational cost; this family of quadrature rule is both nested and based on equally spaced points. The three point quadrature rule of Newton Cotes, known also as the Cavalieri-Simpson rule, is employed in this work:

$$\int_a^b f(\xi)\mathrm{d}\mu(\xi) \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{39}$$

When a cell is split, it is easy to see that only three of the six points required (three for each cells) should be computed again. On the contrary, the points employed at the previous level can be re-employed thanks to the nested nature of the meshes. This makes the sequence of cell evaluations from the coarsest resolution level to the finest one, only a hierarchical representation without extra computational effort. For instance, if a Gauss (two points) quadrature rule would be employed, the point of a previous level could not be used for the evaluation, of the integrals, at successive resolution levels. This feature is a key aspect when the MR framework is coupled (see section §5) with the semi-intrusive scheme presented in the following section.

3.2. ENO polynomial reconstruction for the MR setting

In this section, further details on the polynomial interpolation are provided. The $\mathcal{R}_k$ operator, from a practical point of view, can be obtained by the union of all the polynomial obtained by the conservative interpolation techniques described by the equation (11). Two different operations are relative to the piecewise polynomial approximation $\mathcal{P}_j$. The first is to obtain $\mathcal{P}_j^k$ from the mesh at the resolution level k and, of course, from the cell average quantities at this resolution level. The second operation is the prediction of a cell average value (for a cell entirely contained in the support of the polynomial $\mathcal{P}_j^k$) at the successive resolution level (see equation (19)). To make things clearer, the case of uniform probability distribution is here addressed. The first task is to define the polynomial representation for a second order polynomial piecewise approximations ($r = 2$), over the stochastic cell Ξ_j:

$$\mathcal{P}_j = a(\xi - \xi_j)^2 + b(\xi - \xi_j) + c, \tag{40}$$

Copyright © 2013 John Wiley & Sons, Ltd.
Prepared using fldauth.cls

where ξ_j is the coordinate of the center of the stochastic cell.

To obtain the coefficients a, b and c, the conditions (11) must be fulfilled for a certain stencil. In the case of centered reconstruction, the stencil is fixed and equal to $\mathcal{S}_j = \{\Xi_{j-1}, \Xi_j, \Xi_{j+1}\}$. A linear system can be obtained as

$$\begin{cases} \mathbb{E}(\mathcal{P}_j^k \,|\, \Xi_{j-1}) = \dfrac{1}{\mu(\Xi_{j-1}^k)} \displaystyle\int_{\Xi_{j-1}^k} \mathcal{P}_j^k \mathrm{d}\xi = \mu(\Xi_{j-1}^k) v_{j-1}^k \\ \mathbb{E}(\mathcal{P}_j^k \,|\, \Xi_{j}) = \dfrac{1}{\mu(\Xi_{j}^k)} \displaystyle\int_{\Xi_{j}^k} \mathcal{P}_j^k \mathrm{d}\xi = \mu(\Xi_{j}^k) v_{j}^k \\ \mathbb{E}(\mathcal{P}_j^k \,|\, \Xi_{j+1}) = \dfrac{1}{\mu(\Xi_{j+1}^k)} \displaystyle\int_{\Xi_{j+1}^k} \mathcal{P}_j^k \mathrm{d}\xi = \mu(\Xi_{j+1}^k) v_{j+1}^k, \end{cases} \tag{41}$$

where the linear operator $\mathbb{E}(\bullet \,|\, \Xi)$ becomes (on the generic cell Ξ_j)

$$\mathbb{E}(\mathcal{P}_j^k \,|\, \Xi_j) = a\, \mathbb{E}((\xi - \xi_j)^2 \,|\, \Xi_j) + b\, \mathbb{E}((\xi - \xi_j) \,|\, \Xi_j) + c. \tag{42}$$

If the integration is performed analytically, with respect to the parameter $(\xi - \xi_j)$, the system becomes

$$\begin{pmatrix} \mathbb{E}((\xi - \xi_j)^2 \,|\, \Xi_{j-1}) & \mathbb{E}((\xi - \xi_j) \,|\, \Xi_{j-1}) & 1 \\ \mathbb{E}((\xi - \xi_j)^2 \,|\, \Xi_{j}) & \mathbb{E}((\xi - \xi_j) \,|\, \Xi_{j}) & 1 \\ \mathbb{E}((\xi - \xi_j)^2 \,|\, \Xi_{j+1}) & \mathbb{E}((\xi - \xi_j) \,|\, \Xi_{j+1}) & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = A(\xi - \xi_j) \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \mu(\Xi_{j-1}^k) v_{j-1}^k \\ \mu(\Xi_{j}^k) v_{j}^k \\ \mu(\Xi_{j+1}^k) v_{j+1}^k \end{pmatrix}, \tag{43}$$

where the matrix $A = A(\xi - \xi_j)$ is dependent from the stochastic cell Ξ_j via its coordinate ξ_j. From a practical point of view, when the polynomial reconstruction should be performed over a cell Ξ_j, the matrix $A^{-1}(\xi - \xi_j)$ is first evaluated and then the vector of coefficients is obtained by the matrix vector product with the right hand side that depends from both the resolution level k and the stencil $\mathcal{S}_j$.

However, the procedure described above should be modified if the ENO interpolation is required. The only modification concerns the choice of the stencil: the procedure select the less oscillatory one between the following $\{\Xi_{j-2}, \Xi_{j-1}, \Xi_j\}$, $\{\Xi_{j-1}, \Xi_j, \Xi_{j+1}\}$ and $\{\Xi_j, \Xi_{j+1}, \Xi_{j+2}\}$. The smoothest one is selected choosing the one with $\min(|a|)$ following [22]. Obviously, at the boundaries of the domain, the stencil is always modified to be inside the domain. This is a key aspect if the higher accuracy is desired. With the modification of the stencil, the scheme preserves its maximal accuracy as it is shown for the solution of the stochastic linear advection equation with smooth solution in §6.1.

The second task to solve is the prediction of a cell average v_j^{k+1} at the next following resolution level, if the polynomial $\mathcal{P}_j^k$ reconstruction at the previous resolution level is available (the cell $\Xi_j^{k+1} \subset \Xi_j^k$ as required by the nested character of the discretization procedure). This task is accomplished analytically in the following way. The expectancy operator is applied to the polynomial $\mathcal{P}_j^k$ over the stochastic cell $\Xi_j^{k+1} \subset \Xi_j^k$

$$\mathbb{E}(\mathcal{P}_j^k \,|\, \Xi_j^{k+1}) = a\, \mathbb{E}((\xi - \xi_j^k)^2 \,|\, \Xi_j^{k+1}) + b\, \mathbb{E}((\xi - \xi_j^k) \,|\, \Xi_j^{k+1}) + c, \tag{44}$$

where the terms $\mathbb{E}((\xi - \xi_j^k)^2 \,|\, \Xi_j^{k+1})$ and $\mathbb{E}((\xi - \xi_j^k) \,|\, \Xi_j^{k+1}) + c)$ can be analytically evaluated when the cell Ξ_j^{k+1} is defined.

The procedure described in this section is used in SI scheme in order to obtain the polynomial representation of the functions along the stochastic space to evaluate the expectancy of the flux function.

DOI: 10.1002/fld

4. THE SEMI-INTRUSIVE FINITE VOLUME FORMULATION FOR PDE

In this section, the semi-intrusive (SI) method of Abgrall and Congedo is sketched. The interested reader can refer to [7] for a complete presentation of the numerical scheme and its application to different test cases.

The SI method is an intrusive scheme for the propagation of uncertainties, that requires only a limited number of modification to an existing numerical code. In particular, the SI scheme relies on the deterministic formulation available in a numerical code. Moreover, the number of equations remains equal to the original deterministic formulation, not as in the intrusive polynomial chaos method (see [5]). This feature permits to obtain intrusive stochastic formulation even for high order schemes. In the present work, a second order MUSCL-Hancock method (MHM) is employed to formulate the deterministic part of the scheme. This result, to the best of our knowledge, is the first adaptive intrusive scheme of high-order. Another adaptive intrusive strategy based on data-independent wavelets limited only to first order in time and space is the work of Tryoen et al. [11]. This work is the first to introduce wavelets adaptivity into an intrusive stochastic formulation by means of the polynomial chaos technique, but remain very limited in its generality requiring for each case *ad hoc* modifications.

4.1. MUSCL-Hancock deterministic numerical formulation

The MHM is a slightly different approach with respect to the classical predictor-corrector MUSCL approach. It requires only the computation of slopes in the predictor step. Moreover, it does not require the solution of Riemann problems in the predictor step. The corrector step is based on the evolution of cell-average quantities, taking into account their contribution related to the flux at interfaces obtained by the solution of a Riemann problem. Let us consider a 1D scalar conservation law

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f(u(x,t))}{\partial x} = 0, \tag{45}$$

where $x \in \Omega \subset \mathbb{R}$ is the physical space and $t \in T \subset \mathbb{R}^+$ is the time space. The physical space is divided in a set of non-overlapping cells $\mathcal{C}_i$ with $\Omega = \bigcup_i \mathcal{C}_i$. The classical first order Godunov scheme, applied to (45), is obtained introducing the so-called cell-average $\bar{u}_i$ on each cell $\mathcal{C}_i$:

$$\bar{u}_i(t) = \frac{1}{|\mathcal{C}_i|} \int_{\mathcal{C}_i} u(x,t)\mathrm{d}x, \tag{46}$$

where $|\mathcal{C}_i|$ indicates the volume of the cell. Van Leer [23, 24] proposed to consider non-constant data on each cell to achieve a higher accuracy in the so-called Monotone Upstream-centred Scheme for Conservation Laws (MUSCL). The piecewise linear approximation is used for the solution $u(x,t)$ on the cell $|\mathcal{C}_i|$:

$$u(x,t_n) = \bar{u}_i^n + \sigma_i^n (x - x_i) \quad \text{with} \quad x_{i_L} \leq x \leq x_{i_R}, \tag{47}$$

with σ_i^n the so-called slope. Of course, the choice of $\sigma_i^n = 0$ leads to the Godunov scheme. A slope limiter should be introduced near the discontinuity to avoid oscillations. In this work, both the Roe's superbee limiter and the van Leer limiters are employed. The superbee limiter in its limited slope form is

$$\begin{cases} \sigma_i^n = \mathrm{maxmod}\left(\sigma_{(1)}^n, \sigma_{(2)}^n\right) \\ \sigma_{(1)}^n = \mathrm{minmod}\left(\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathcal{C}_i|}\right), 2\left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathcal{C}_i|}\right)\right) \\ \sigma_{(2)}^n = \mathrm{minmod}\left(2\left(\dfrac{\bar{u}_{i+1}^n - \bar{u}_i^n}{|\mathcal{C}_i|}\right), \left(\dfrac{\bar{u}_i^n - \bar{u}_{i-1}^n}{|\mathcal{C}_i|}\right)\right), \end{cases} \tag{48}$$

where the minmod and maxmod functions are defined as follows

$$\mathrm{minmod}(a,b) = \begin{cases} a & \text{if} \quad |a| < |b| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |a| > |b| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab <= 0 \end{cases}$$

$$\operatorname{maxmod}(a,b)=\begin{cases} a & \text{if} \quad |a|>|b| \quad \text{and} \quad ab>0 \\ b & \text{if} \quad |a|<|b| \quad \text{and} \quad ab>0 \\ 0 & \text{if} \quad ab<=0. \end{cases}$$

The van Leer limiter, in the form of slope limiter, is defined as (see Toro [24] for further details)

$$\sigma_i^n=\begin{cases} \text{MIN}\left(\dfrac{2R}{1+R},\dfrac{2}{1+R}\right)\dfrac{\bar{u}_{i+1}^n-\bar{u}_{i-1}^n}{2\Delta x} & \text{if} \quad R>0 \\ 0 & \text{if} \quad R\leq 0, \end{cases} \tag{49}$$

where R is the ratio between successive slopes $R=(\bar{u}_i^n-\bar{u}_{i-1}^n)/(\bar{u}_{i+1}^n-\bar{u}_i^n)$.

The MHM is then introduced in order to avoid the problem related to the solution of the so-called generalized Riemann problem, in which the two states are not constant. The fully discrete second order MHM, for computing the cell averaged solution $\bar{u}_i^{n+1}$, consists of the following three steps:

- Step 1 - For each cell $\mathcal{C}_\ell \in \{\mathcal{C}_{i-1},\mathcal{C}_i,\mathcal{C}_{i+1}\}$, the solution at the interface is computed according to

$$\begin{cases} u_{\ell_L}^n=\bar{u}_\ell^n-\sigma_\ell^n\dfrac{|\mathcal{C}_\ell|}{2} \\ u_{\ell_R}^n=\bar{u}_\ell^n+\sigma_\ell^n\dfrac{|\mathcal{C}_\ell|}{2} \end{cases} \tag{50}$$

- Step 2 - On each cell $\mathcal{C}_\ell \in \{\mathcal{C}_{i-1},\mathcal{C}_i,\mathcal{C}_{i+1}\}$, the solution evolved of a half time step employing the flux function $f=f(u)$:

$$\begin{cases} u_{\ell_R}^{\Uparrow}=\bar{u}_{\ell_R}+\dfrac{1}{2}\dfrac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u_{\ell_L}^n)-f(u_{\ell_R}^n)\right) \\ u_{\ell_L}^{\Uparrow}=\bar{u}_{\ell_L}+\dfrac{1}{2}\dfrac{\Delta t}{|\mathcal{C}_\ell|}\left(f(u_{\ell_L}^n)-f(u_{\ell_R}^n)\right) \end{cases} \tag{51}$$

- Step 3 - The cell-averaged value on the cell $\mathcal{C}_i$ evolves following

$$\bar{u}_i^{n+1}=\bar{u}_i^n-\frac{\Delta t}{|\mathcal{C}_i|}\left(\mathcal{F}^{\text{RM}}\left(u_{i-1_R}^{\Uparrow},u_{i_L}^{\Uparrow}\right)-\mathcal{F}^{\text{RM}}\left(u_{i_R}^{\Uparrow},u_{i+1_L}^{\Uparrow}\right)\right). \tag{52}$$

The symbol $\mathcal{F}^{\text{RM}}$ is employed to indicate the flux evaluated at the interface, after the solution of the Riemann problem defined by two constant states based on the evolved extrapolated values. For the linear advection §6.1 and Burgers equation §6.2, an exact Riemann solver is used. Moreover, in the case of the Euler system of equations §6.3, the Roe-Pike method is employed with the Harten-Hyman entropy fix following [24].

The time advancing formula is then limited to a stencil of only three cells $\mathcal{C}_{i-1}$, $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ but the computation of the slopes for the cells $\mathcal{C}_{i-1}$ and $\mathcal{C}_{i+1}$ requires (see (48) and (49)) also to know the solution on the two sourrounding cells $\mathcal{C}_{i-2}$ and $\mathcal{C}_{i+2}$. The average solution $\bar{u}_i^{n+1}$, on each cell $\mathcal{C}_i$ at time $t_{n+1}=t_n+\Delta t$, can be computed knowing the solution on the augmented stencil $\left\{\bar{u}_{i-2}^n,\bar{u}_{i-1}^n,\bar{u}_i^n,\bar{u}_{i+1}^n,\bar{u}_{i+2}^n\right\}$. In the following, the notation $\bar{u}_i^{n+1}=\text{MHM}\left(\bar{u}_{i-2}^n,\bar{u}_{i-1}^n,\bar{u}_i^n,\bar{u}_{i+1}^n,\bar{u}_{i+2}^n,\Delta t\right)$ is used to identify the ensemble of the operation described above. The aim is to evaluate the updated value in time of a certain cell $\bar{u}_i^{n+1}$, knowing the solution at the previous time step.

4.2. Semi-intrusive formulation for the MHM

The SI version of the MHM (here presented in the 1D stochastic case without loss of generality) can be obtained adding one dimension more (the stochastic space) with a finite-volume like representation. In particular, the conditional expectancy operator, defined on the stochastic cell Ξ_j, is introduced according to the following definition:

$$\mathbb{E}(\bullet\,|\,\Xi_j)=\frac{1}{\mu(\Xi_j)}\int_{\Xi_j}\bullet(x,\xi,t)\,p(\xi,t)\,\mathrm{d}\xi. \tag{53}$$

If the conditional expectancy operator is applied to the step three of the MHM scheme (52), the following scheme is obtained:

$$\mathbb{E}(u_i^{n+1}\,|\,\Xi_j) = \mathbb{E}(u_i^{n}\,|\,\Xi_j) - \frac{\Delta t}{|\mathcal{C}_i|}\left(\mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right)|\,\Xi_j) - \mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right)|\,\Xi_j)\right). \tag{54}$$

The evaluation of the updated conditional expectancy value on the cell Ξ_j, is obtained by evaluating the conditional expectancy contribution related to the numerical fluxes $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right)|\,\Xi_j)$ and $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right)|\,\Xi_j)$. To evaluate this integral contribution, a polynomial representation of the physical averaged solution with respect to the stochastic dimensions, has to be obtained. The conservative interpolation procedure, already presented in §3 to obtain the reconstruction operator $\mathcal{R}_k$, can be adopted requiring for the polynomial $\mathcal{P}_j(\xi)$:

$$\mathbb{E}(P_\ell(\xi)\,|\,\Xi_\ell) = \mathbb{E}(u\,|\,\Xi_\ell) \quad \forall \Xi_\ell \in \mathcal{S}_j \tag{55}$$

If the stencil $\mathcal{S}_j$ is chosen with a cardinality $s = s(r) = \mathrm{card}(\mathcal{S}_j) = r+1$ (for a 1D space), a polynomial $\mathcal{P}_j(\xi)$ of degree r can be built.

The polynomial representation $\mathcal{P}_j(\xi)$ can be injected into the steps 1 (50) and 2 (51) of the MHM. If the Cavalieri-Simpson rule (using three quadrature points $ng = 3$) is adopted for the quadrature, the SI scheme for the MHM can be recasted in a form that makes easy the use of MR stochastic representation of data.

We assume a uniform tessellation for the physical and stochastic space, with a number of cells equal to N_x and N_ξ, respectively and a constant time step Δt. The first step is to evaluate the initial condition in terms of conditional expectancies. This can be obtained easily via a tensorization of the quadrature rule and evaluating the analytical value of the function $u(x, \xi, 0)$. This step yields the stochastic initial condition $\mathbb{E}(u_i(x, \xi, 0)\,|\,\Xi_j)$ for all $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_\xi$.

The SI algorithm becomes:

Algorithm 3: Semi-intrusive version of the MUSCL-Hancock method for a 1D stochastic space.

```
for $n = 1, \ldots, N_t$ do
    for $i = 1, \ldots, N_x$ do
        for $j = 1, \ldots, N_\xi$ do
            Polynomial reconstruction (via (55)) over $\Xi_j = [\xi_{j-1}, \xi_j] \Rightarrow \mathcal{P}_j(\xi)$ ;
            for $ng = 1, \ldots, 3$ do
                $\xi_{ng} = \xi_{j-1} + \frac{\xi_j - \xi_{j-1}}{2}(ng-1)$ ;
                Step 1 (see (50)) $\Rightarrow \forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1},\} \to \left\{u^n_{\ell_L}(\xi_{ng}), u^n_{\ell_R}(\xi_{ng})\right\}$ ;
                Step 2 (see (51)) $\Rightarrow \forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1},\} \to \left\{u^{\Uparrow}_{\ell_L}(\xi_{ng}), u^{\Uparrow}_{\ell_R}(\xi_{ng})\right\}$ ;
            end
            Flux expectancy computation:
            $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}_L\,|\,\Xi_j) = \sum_{ng=1}^{3} w_{ng}\ \mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}(\xi_{ng}), u^{\Uparrow}_{i_L}(\xi_{ng}), \xi_{ng}\right)$ ;
            $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}_R\,|\,\Xi_j) = \sum_{ng=1}^{3} w_{ng}\ \mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}(\xi_{ng}), u^{\Uparrow}_{i+1_L}(\xi_{ng}), \xi_{ng}\right)$ ;
            Time update:
            $\mathbb{E}(\bar{u}_i^{n+1}\,|\,\Xi_j) = \mathbb{E}(\bar{u}_i^{n}\,|\,\Xi_j) - \frac{\Delta t}{|\mathcal{C}_i|}\left(\mathbb{E}(\mathcal{F}^{\mathrm{RM}}_L\,|\,\Xi_j) - \mathbb{E}(\mathcal{F}^{\mathrm{RM}}_R\,|\,\Xi_j)\right)$
        end
    end
end
```

where $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}_L\,|\,\Xi_j) = \mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i-1_R}, u^{\Uparrow}_{i_L}\right)|\,\Xi_j)$ and $\mathbb{E}(\mathcal{F}^{\mathrm{RM}}_R\,|\,\Xi_j) = \mathbb{E}(\mathcal{F}^{\mathrm{RM}}\left(u^{\Uparrow}_{i_R}, u^{\Uparrow}_{i+1_L}\right)|\,\Xi_j)$.

DOI: 10.1002/fld

5. THE OVERALL MULTIRESOLUTION ADAPTIVE-SI SCHEME

In the previous section, the SI scheme applied to the MHM is presented. In this section, the adaptive version of the numerical algorithm (aSI) is described. The main difference, referring to the algorithm 3 is in the internal loop, on j, concerning the stochastic cells. This loop should be substituted by the application of the TE algorithm 1. The discretization step is performed by the application of the MHM, as presented in the internal loop (on j), in the algorithm 3. The complete aSI scheme is:

Algorithm 4: Semi-intrusive version of the MUSCL-Hancock method for a 1D stochastic space.

```
for $n = 1, \dots, N_t$ do
    for $i = 1, \dots, N_x$ do
        while $2 \leq k \leq L$ do
            for $j = 1, \dots, J_{k-2}$ do
                Encoding:
                $d_j^{k-1} = v_{2j}^{k-1} - (\mathrm{P}_{k-2}^{k-1} v^{k-2})_{2j} = v_{2j}^{k-1} - \left( \frac{1}{\mu(\Xi_{2j}^{k-1})} \int_{\Xi_{2j}^{k-1}} \mathcal{R}_{k-2} v^{k-2} p(\xi) \mathrm{d}\xi \right)$ ;
                Truncation:    $\hat{d}_j^{k-1} = \mathrm{tr}(d_j^{k-1}, \varepsilon_{k-1})$ ;
            end
            for $j = 1, \dots, J_{k-1}$ do
                if $\hat{d}_j^{k-1} > 0$ then
                    Discretization:
                    for $\Xi_q \in \{\Xi_{2j-1}^k, \Xi_{2j}^k\}$ do
                        for $ng = 1, \dots, 3$ do
                            Polynomial evaluation: $\bar{u}(x, \xi_{ng}, t_n) \simeq (\mathcal{D}_k \mathcal{R}_L v^L(t_n))(\xi_{ng})$
                            Step 1 (see (50))
                            $\Rightarrow \forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}, \} \rightarrow \{u_{\ell_L}^n(\xi_{ng}), u_{\ell_R}^n(\xi_{ng})\}$
                            Step 2 (see (51))
                            $\Rightarrow \forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \mathcal{C}_i, \mathcal{C}_{i+1}, \} \rightarrow \left\{u_{\ell_L}^{\Uparrow}(\xi_{ng}), u_{\ell_R}^{\Uparrow}(\xi_{ng})\right\}$
                        end
                        Flux expectancy computation:
                        $\mathbb{E}(\mathcal{F}_L^{\mathrm{RM}} \,|\, \Xi_q) = \sum_{ng=1}^3 w_{ng} \, \mathcal{F}^{\mathrm{RM}} \left( u_{i-1_R}^{\Uparrow}(\xi_{ng}), u_{i_L}^{\Uparrow}(\xi_{ng}), \xi_{ng} \right)$
                        $\mathbb{E}(\mathcal{F}_R^{\mathrm{RM}} \,|\, \Xi_q) = \sum_{ng=1}^3 w_{ng} \, \mathcal{F}^{\mathrm{RM}} \left( u_{i_R}^{\Uparrow}(\xi_{ng}), u_{i+1_L}^{\Uparrow}(\xi_{ng}), \xi_{ng} \right)$
                        Cell agglomeration of $\mathbb{E}(u_i^n \,|\, \Xi_q)$ via equation (56)
                        Time update:
                        $\mathbb{E}(\bar{u}_i^{n+1} \,|\, \Xi_q) = \mathbb{E}(\bar{u}_i^n \,|\, \Xi_q) - \frac{\Delta t}{|\mathcal{C}_i|} \left( \mathbb{E}(\mathcal{F}_L^{\mathrm{RM}} \,|\, \Xi_q) - \mathbb{E}(\mathcal{F}_R^{\mathrm{RM}} \,|\, \Xi_q) \right)$
                    end
                end
            end
        end
        Reconstruction: $(\mathcal{D}_L \mathcal{R}_L v^L)_l = (\mathcal{D}_L \bar{u}(x_i, \xi, t_{n+1}))_l$   with   $l \in \mathcal{S}_j^L$ ;
    end
end
```

The reconstruction operator $\mathcal{R}_k$ for each cell Ξ_j is the polynomial $\mathcal{P}_j$ reconstructed for the SI scheme. A link between the MR representation and the SI scheme exists since the polynomial representation of the data in the stochastic space is the same for the SI and TE. The polynomial reconstruction is carried out when the algorithm attain the highest resolution level (indicated in the algorithm by $k = L$) and the reconstruction operator $\mathcal{R}_L$ is then obtained and stored. The

DOI: 10.1002/fld

reconstruction operator is then used, for the polynomial evaluation before Step 1. The physical cell-averaged values are obtained, for each quadrature points ξ_{ng}, applying the discretization operator $\mathcal{D}_k$. Moreover, a conservative interpolation is also present into the MR algorithm, where the operator $\mathcal{R}_k$ is used to obtain the wavelets during the *encoding* procedure.

One important feature of the aSI algorithm is the possibility to locally refine/derefine the stochastic space, as a function of the variation of the solution during the computation. At the end of each time step, for each physical location, the algorithm produces a sequence of conditional expectancies $\mathbb{E}(u^n \,|\, \Xi_j)$ with different measures $\mu(\Xi_j)$, due to the local refinement/derefinement of the tessellation. The TE strategy starts from the coarsest level to the finest (until some cell have to be split or the maximum resolution level is reached). In practice, if a cell has not to be splitted, it is moved at the highest resolution level. The local variation of the cell measure yields a strong relation between the actual level of evaluation of the scheme, and the maximum level (locally) reached at the previous time step (and consequently the measure of each cell). Two problems exist: the agglomeration of a cell at a time n, and the splitting of a cell at a time $n+1$. The MR framework presented is based on a nested subdivision of the cell. Then, at the end of the TE algorithm, each cell belonging to the coarsest level $k=0$, will result in a set of cells. When the TE algorithm requires the application of the SI-MHM at a generic level k, an equivalent conditional expectancy $\mathbb{E}(u \,|\, \Xi_j^k)$ evaluated at time n is computed by applying the equation (54). This conditional expectancy should be obtained by the agglomeration of all the stochastic cells belonging to Ξ_j^k at time n, following the exact definition:

$$\mathbb{E}(u \,|\, \Xi_j^k) = \frac{1}{\mu(\Xi_j^k)} \sum_{\Xi_\ell \subseteq \Xi_j^k} \mu(\Xi_\ell)\mathbb{E}(u \,|\, \Xi_\ell). \tag{56}$$

Obviously, it is easy to verify that the limit case is the one with a cell not subdivided, then the equation (56) reduces to an identity. Due to the nested sequences of operators and meshes, a cell would be always constituted by an integer number of cells at the end of the TE algorithm (see algorithm 1). A sketch of a possible situation for the agglomeration of a cell Ξ_j^k is reported in figure 2.

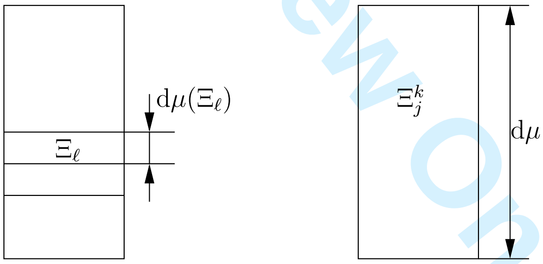


Figure 2. Example of the agglomeration procedure to obtain a coarser cell Ξ_j even if the TE algorithm yields a set of children cells.

The other issue is related to reduce the computational cost basing on the computed quantities, when a cell has to be split. For this reason, the quadrature rule of Newton-Cotes is adopted. In this case, the entire set of degrees of freedom (dof) can be saved, if the cell has to be split. Let us consider the figure 3, where the Cavalieri-Simpson rule is used. On the left, the cell at level k is represented with its dof, the circles are used for the value of $\bar{u}_i$ obtained via the polynomial $\mathcal{P}_j$ (the polynomial evaluation step in the algorithm 4), and squares for the fluxes obtained after the application of the step 1 and 2 of the MHM. When the cell is split in two cells, only three points have to be added (the numerical scheme has to be applied). On the contrary, the other points can be obtained directly from the mother cell at level k. In the figure 3, the black circle/squares represent the new points to compute. In practice, the black points are associated to the values for $\bar{u}_i$ obtained by interpolation and the fluxes are obtained via the Step 1 and 2; otherwise, they are only recovered from the mother cell. Finally, the fluxes conditional expectancy computation is performed easily combining the new fluxes (black) and the old ones (white) with the correct weights for the quadrature.

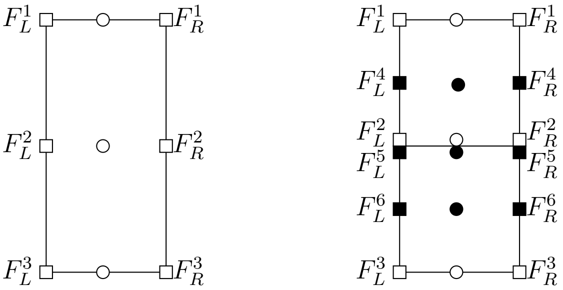


Figure 3. Example of a splitting procedure to save the computational cost associated to the degree of freedom already computed. On the left the cell at level of resolution k is reported while on the right the corresponding split cells are reported with the new points to explicitly add (black symbols).

The nested procedure described above allows to extend the accuracy of the quadrature rule even to high-order Newton-Cotes formula. Moreover, in the present work, the three points Cavalieri-Simpson rule (see (39)) is employed. The error is proportional to the fourth derivative of the integrand, so the rule is fully accurate to polynomial function of order equal or less than three (see [21] for further details). In the following, the variance of the outputs of interest is computed. In this case, the quadrature of the polynomial $\mathcal{P}_j$ squared has to be evaluated on each cell $\Xi_j = [\xi_a, \xi_b]$. In order to attain the exact integration of $\mathcal{P}_j^2$, the closed four points Newton-Cotes rule (also known as the Boole's rule) is employed

$$\int_{\Xi_j} f(\xi)\mathrm{d}\xi = \frac{\xi_b - \xi_a}{90}\left(7f(\xi_a) + 32f\left(\xi_a + \frac{\xi_b - \xi_a}{4}\right)\right. \\ \left. + 12f\left(\xi_a + \frac{\xi_b - \xi_a}{2}\right) + 32f\left(\xi_a + 3\frac{\xi_b - \xi_a}{4}\right) + 7f(\xi_b)\right). \tag{57}$$

The last five points rule has an error $\mathcal{O}(f^{(6)}(\eta))$, where $\eta \in]\xi_a, \xi_b[$, so it is able to integrate exactly polynomial function of order equal to five.

6. NUMERICAL RESULTS

In this section, the aSI scheme derived in §5 is applied to a set of test problems. The aim is to show the convergence properties and to provide some evidence of the advantage to employ an adaptive representation of the solution in the stochastic space. For all the problems, the expectancy and the variance of the some outputs are computed according to the definitions (3) with respect to their exact value. Different 1D-1D test cases are taken into account. The linear advection problem is solved for both smooth and discontinuous initial conditions in section §6.1. In the first case, the uncertainty is considered in the initial condition, while in the discontinuous case an uncertain advection velocity is considered. For this test case, both the convergence curves for the first order Godunov method and the MHM are reported to demonstrate the ability of the scheme to maintain the convergence properties of the deterministic scheme. The Burgers equation is then solved employing a smooth initial, but uncertain, initial condition (§6.2). This case is chosen to demonstrate the ability of the scheme to capture (refining the stochastic space) a discontinuous solution (along the stochastic dimension) even if the discontinuities form during the evolution of a smooth solution. This property is a key feature in the development of numerical schemes for UQ in compressible flows applications. The last test case is the stochastic analysis of the uncertain shock tube problem solving the Euler system of equations in section §6.3. In this case, the statistics of the density are compared to the semi-analytical solution of the Euler equations, considering an uncertain parameter on the initial conditions (see Appendix .1).

Systematically in this paper, the spatial norms are computed employing the following definitions

$$\begin{cases} \mathrm{err}_{\mathcal{E}^m}|_{L_p} = ||\mathcal{E}^m(x) - \mathcal{E}^m_{\mathrm{ref}}(x)||_{L_p} = \left(\dfrac{1}{N_x}\displaystyle\sum_{i=1}^{N_x} |\mathcal{E}^m(\bar{u}_i) - \mathcal{E}^m_{\mathrm{ref}}(\bar{u}_i)|^p\right)^{1/p} \\ \mathrm{err}_{\mathcal{E}^m}|_{L_\infty} = ||\mathcal{E}^m(x) - \mathcal{E}^m_{\mathrm{ref}}(x)||_{L_\infty} = \max_i |\mathcal{E}^m(\bar{u}_i) - \mathcal{E}^m_{\mathrm{ref}}(\bar{u}_i)|\,, \end{cases} \tag{58}$$

where the integer $p = 1, 2$ for the L_1 and L_2 norms in the physical space and $\mathcal{E}^m$ indicates a statistical moment, *i.e.* the expectancy or the variance.

6.1. *Linear advection*

The first test case is the linear advection problem here reported, for $\Omega = [0, 1]$, in its general stochastic formulation

$$\begin{cases} \dfrac{\partial u(x,\xi,t)}{\partial t} + a(\xi,t)\dfrac{\partial u(x,\xi,t)}{\partial x} = 0 \\ u(x,\xi,0) = u_0(x,\xi), \end{cases} \tag{59}$$

where both the advection velocity a and the initial condition u_0 can depend on a random parameter.

Let us consider first the smooth test-case with an initial condition equal to $u_0(x,\xi,t) = sin(4\pi x + 20\xi)$, with the random parameter uniformly distributed $\xi \sim \mathcal{U}[0,1]$. The problem is solved until the time $t = 1$ with a constant advection velocity equal to $a = 0.1$ and with periodic boundary conditions. The exact solution can be computed analytically as follows

$$u(x,\xi,1) = sin(4(x - 0.1t)\pi + 20\xi) \tag{60}$$

The exact statistics can be computed as function of the i–th cell $\mathcal{C}_i = [x_i - \frac{|\mathcal{C}_i|}{2}, x_i + \frac{|\mathcal{C}_i|}{2}]$, integrating first with respect to the stochastic space and then with respect to the space

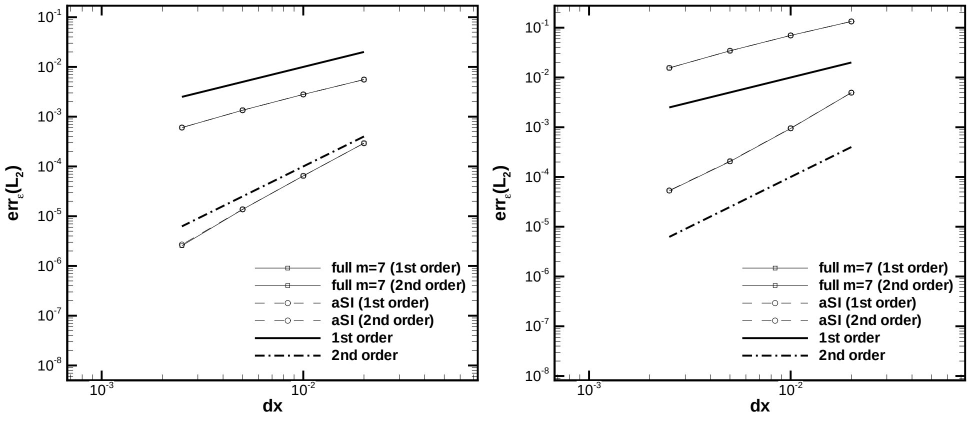
$$\begin{cases} \mathcal{E}(\bar{u}_i) = \dfrac{1}{|\mathcal{C}_i|}\displaystyle\int_{\mathcal{C}_i}\int_{\Xi} u(x,\xi,1)\,\mathrm{d}\xi\,\mathrm{d}x \\ \mathrm{Var}(\bar{u}_i) = \dfrac{1}{|\mathcal{C}_i|}\displaystyle\int_{\mathcal{C}_i}\int_{\Xi} u^2(x,\xi,1)\,\mathrm{d}\xi - \left(\int_{\Xi} u^2(x,\xi,1)\,\mathrm{d}\xi\right)^2\,\mathrm{d}x. \end{cases} \tag{61}$$

Expressions for both statistics are obtained using the MAPLE software. Numerical simulations are carried out on equally spaced spatial meshes of 51, 101, 201 and 401 points, with $N_t = 200$ time steps and $\Delta t = 5 \times 10^{-3}$.

In figure 4, both the expectancy of the solution 4(a) and the variance 4(a) for the linear advection problem (59) with smooth initial condition and constant advection velocity are reported. The continuous lines indicate the solution obtained via the scheme without compression, while with the dashed lines the solution obtained via the application of the aSI algorithm. In particular, the polynomial reconstruction is taken as a centered second-order polynomial except for the two boundary cells where the stencil is fully shifted into the numerical domain in order to maintain the order of accuracy. In particular, both the Godunov first order scheme and the MHM are reported to show that the numerical scheme is able to preserve the expected order of convergence even with compression. To preserve the formal second order of accuracy, the slope for the MHM is evaluated by a centered approximation without any limiter function. The full solution is obtained on an equally spaced mesh of 128 stochastic cells while the aSI algorithm is applied starting from a coarse level of 16 cell ($m_0 = 4$) to a higher resolution level of 128 cells ($m = 7$) and a threshold equal to $\varepsilon = 10^{-3}$. Note that the finest level is indicated as m. This case is reported in order to show the formal accuracy of the method because the solution is regular enough to minimize the gain associated to the compression of the solution. In particular, the average number of cells employed by the aSI scheme is 126 against the 128 of the full solution. Of course, the level of compression could be easily increased in this case employing a higher order polynomial $\mathcal{P}_j$ for the reconstruction. Remark that,

DOI: 10.1002/fld

looking at the accuracy, the stochastic reconstruction (quadratic polynomial) is sufficiently accurate with respect to the spatial and time accuracy (second order in the case of MHM). On the contrary, looking at the compression, a higher polynomial order can yield a stronger compression keeping the second order convergence rate.



(a) (b)

Figure 4. Spatial convergence for the linear advection problem with smooth initial condition (60). The statistics of the solution (mean (a) and variance (b)) obtained with (aSI) and without (full) compression are reported for both the Godunov first order scheme and the MHM method with a centered slope.

Let us consider now the linear advection problem (59), that is solved with an uncertain advection ($\xi \sim \mathcal{U}[\frac{1}{5}, \frac{4}{5}]$) velocity defined as

$$a(\xi) = \frac{1}{40} e^{5\xi^2} + \frac{1}{5}, \tag{62}$$

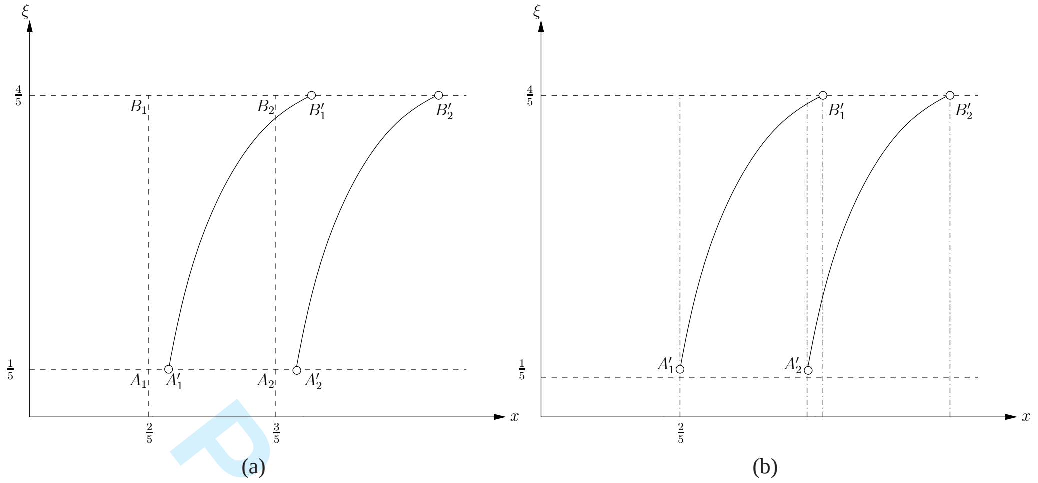
considering a discontinuous initial condition (in the physical space)

$$u(x, \xi, 0) = \begin{cases} 1 & \text{if} \quad \frac{2}{5} \leq x \leq \frac{3}{5} \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \tag{63}$$

In this case, the problem is solved until the final time of $t = 0.4$ with 200 equal steps of $\Delta t = 2 \times 10^{-3}$. The exact solution is derived for the first two statistical moments employing the following procedure. Referring to the figure 5, starting from the initial condition (defined by the points A_1, A_2, B_2, B_1) the new points (coordinates in the physical space) at the final time ($t = 0.4$) can be computed as follows

$$\begin{cases} A_1^{\prime,x} = A_1^x + a\left(\frac{1}{5}\right) t = \frac{12}{25} + \frac{1}{100} e^{\frac{1}{5}} \\ A_2^{\prime,x} = A_2^x + a\left(\frac{1}{5}\right) t = \frac{12}{25} + \frac{1}{100} e^{\frac{1}{5}} \\ B_1^{\prime,x} = B_1^x + a\left(\frac{1}{5}\right) t = \frac{12}{25} + \frac{1}{100} e^{\frac{16}{25}} \\ B_2^{\prime,x} = B_2^x + a\left(\frac{1}{5}\right) t = \frac{12}{25} + \frac{1}{100} e^{\frac{16}{25}}. \end{cases} \tag{64}$$

At the final time step, four different regions can be identified (see figure 5(b)). The solution in the external region, where $x \leq A_1^{\prime,x}$ and $x \geq B_2^{\prime,x}$, is easily identified as $u(x, \xi, t) = 0$. For the remaining



(a)

(b)

Figure 5. Schematic representation of the evolution between the initial condition (points A_1, A_2, B_1, B_2) and the final condition at time $t = 0.4$ (points A'_1, A'_2, B'_1, B'_2) (a). The regions in which the exact solution should be computed, at the final time $t = 0.4$, are reported in (b).

regions, the position of the discontinuity has to be computed. In particular, it is possible to define the two functions $\xi_d^1 = \xi_d^1(x)$ and $\xi_d^2 = \xi_d^2(x)$ as the positions of the discontinuities for each x belonging respectively to the intervals $[A_1'^{,x}, B_1'^{,x}]$ and $[A_2'^{,x}, B_2'^{,x}]$. If x belongs to the interval defined above, the following relations must hold

$$\begin{cases} x = A_1'^{,x} + a\left(\xi_d^1\right)t = \dfrac{12}{25} + \dfrac{1}{100}e^{(\xi_d^1)^2} & \text{if} \quad x \in [A_1'^{,x}, B_1'^{,x}] \\ x = A_2'^{,x} + a\left(\xi_d^2\right)t = \dfrac{17}{25} + \dfrac{1}{100}e^{(\xi_d^2)^2} & \text{if} \quad x \in [A_2'^{,x}, B_2'^{,x}]. \end{cases} \tag{65}$$

As a consequence, the position of the discontinuities, for a certain physical position can be derived

$$\begin{cases} \xi_d^1 = \xi_d^1(x) = \sqrt{\ln\left(100\left(x - \dfrac{12}{25}\right)\right)} \\ \xi_d^2 = \xi_d^2(x) = \sqrt{\ln\left(100\left(x - \dfrac{17}{25}\right)\right)}. \end{cases} \tag{66}$$

The exact statistics of the physical cell average $\bar{u}_i$ can be computed exactly for each cell $\mathcal{C}_i = \left[x_i - \frac{|\mathcal{C}_i|}{2}, x_i + \frac{|\mathcal{C}_i|}{2}\right]$ (in the limit of $|\mathcal{C}_i| \to 0$). For the mean, they are defined as

$$\mathcal{E}(\bar{u}_i) = \begin{cases} 0 & \text{if} \quad x_i \leq A_1'^{,x} \text{ or } x_i \geq B_2'^{,x} \\ \dfrac{5}{3}\left(\xi_d^1(x_i) - \dfrac{1}{5}\right) & \text{if} \quad x_i \in [A_1'^{,x}, A_2'^{,x}] \\ \dfrac{5}{3}\left(\xi_d^1(x_i) - \xi_d^2(x_i)\right) & \text{if} \quad x_i \in [A_2'^{,x}, B_1'^{,x}] \\ \dfrac{5}{3}\left(\dfrac{4}{5} - \xi_d^2(x_i)\right) & \text{if} \quad x_i \in [B_1'^{,x}, B_2'^{,x}]. \end{cases} \tag{67}$$

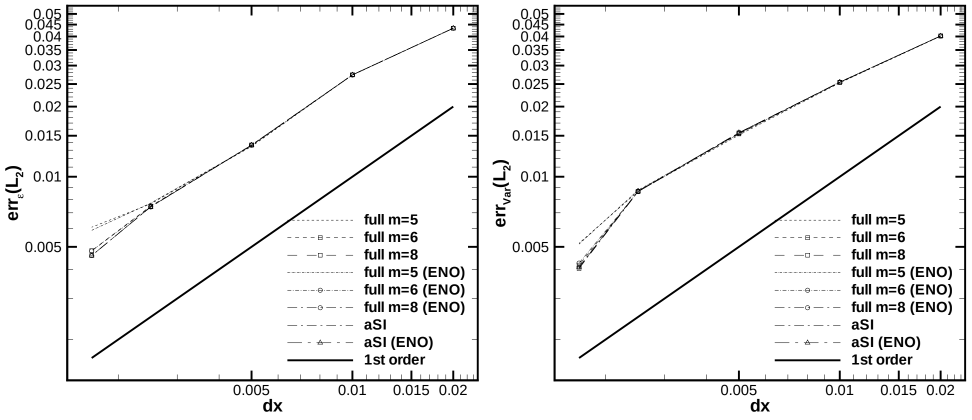
Concerning the variance, they can be obtained as (and not as $\text{Var} = \mathcal{E}((\bar{u}_i)^2) - (\mathcal{E}(\bar{u}_i))^2$)

$$\text{Var} = \mathcal{E}(\bar{u}_i) - (\mathcal{E}(\bar{u}_i))^2 \quad \forall x_i \in [0,1], \tag{68}$$

because in this specific case ($\bar{u}(x,\xi,t) = 1$)

$$\int_{\Xi} \bar{u}(x_i,\xi,t)^2 p(\xi)\mathrm{d}\xi = \int_{\Xi} \bar{u}(x_i,\xi,t)p(\xi)\mathrm{d}\xi = \mathcal{E}(\bar{u}_i). \tag{69}$$

In figure 6, the spatial convergence for the aSI scheme and for the full scheme, employing only the MHM with the superbee limiter (48), are reported for the mean 6(a) and the variance 6(b) (L_2 norms). Similar curves are obtained for L_1 and L_∞ norms but are not reported here for brevity. The computations are performed over equally spaced meshes in the physical space Ω with 51, 101, 201, 401 and 601 points. The aSI scheme is applied with a coarsest level of 16 cells ($m_0 = 4$), a finest level of 256 stochastic cells ($m = 4$) and a threshold equal to $\varepsilon = 10^{-3}$. The polynomial reconstruction is the quadratic polynomial with and without ENO selection of the stencil. The average number of stochastic cells employed is equal to 39 when the ENO selection is employed and 40 with the centered stencil.



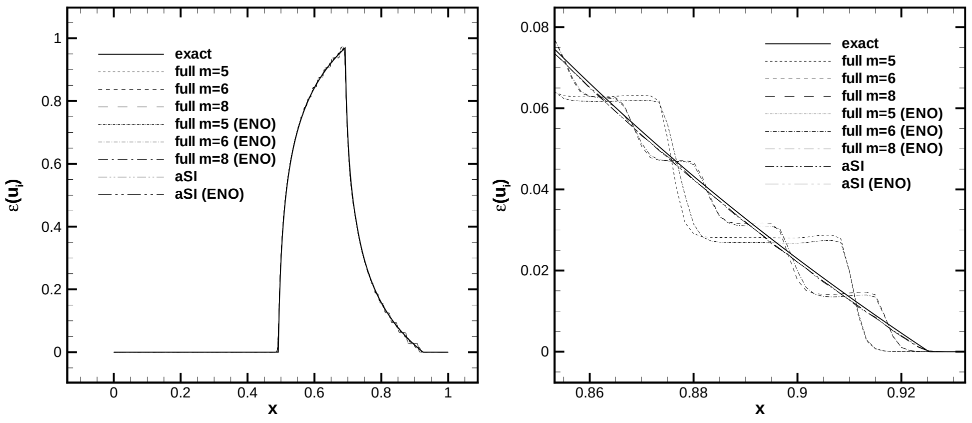
(a)

(b)

Figure 6. Spatial convergence for the linear advection problem with discontinuous initial condition (63). The statistics of the solution (mean (a) and variance (b)) obtained with (aSI) and without (full) compression are reported for the MHM method with the superbee limiter (48).

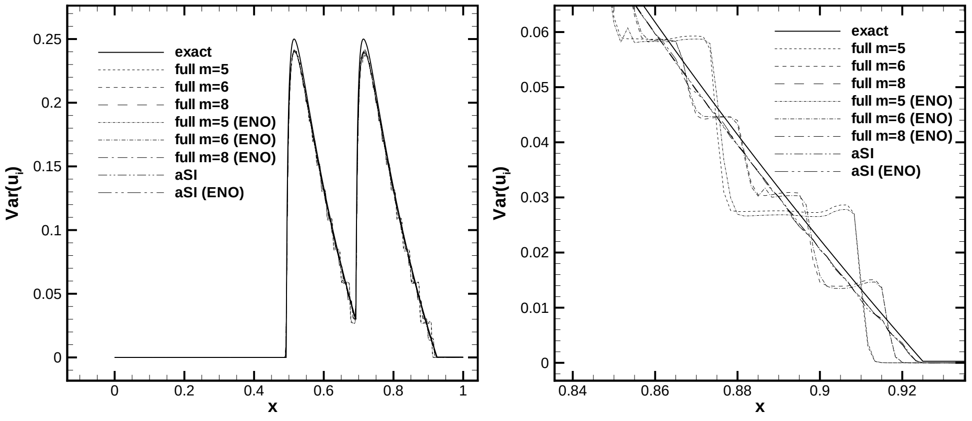
The figure 6 shows that the aSI scheme is able to preserve the accuracy and the order of convergence of the full scheme with a reduction of the computational cost with respect to the full solution obtained over a grid of 256 cells ($m = 8$). The aSI scheme requires a computational effort equivalent to a computation carried out on about 40 equally spaced stochastic cells. The full solutions on 32 ($m = 5$) and 64 ($m = 6$) cells are then reported in order to compare the efficiency of the scheme with respect to a solution obtained with a similar computational effort. However, the aSI scheme performs better with respect to both the full solution at 32 and 64 cells. Moreover, the quality with respect to the full solution of 256 cells is only slightly degraded. In figures 7 and 8, the statistics of the solution are reported over the entire physical space (the mesh of 601 points) and compared to the exact solution (see (67)) obtained on 2001 equally spaced points in the physical space. The solutions obtained with the full scheme with 32 and 64 stochastic cells exhibit the well-known staircase phenomenon, i.e. in presence of discontinuous solutions the statistics are constituted by a series of *plateau*. The presence of the plateau is due to the lower resolution associated to the discretization of the stochastic space with respect to the resolution of the physical space. The staircase phenomenon is more evident for the coarser case (32 cells), reduces slightly with 64 cells, and disappear with 256 cells. The aSI scheme automatically refines the space where a higher resolution is required. Remark that the staircase problem disappears by using aSI even if the (average) number of cells employed is lower than 64 (see figure 7(b) and 8(b)).

DOI: 10.1002/fld



(a) (b)

Figure 7. Expectancy for the cell averaged solution of the linear advection equation with discontinuous initial condition (63) at the final time $t = 0.4$. The whole physical domain is represented in (a), while in the figure (b) a zoom in the shock region is reported. The mesh is constituted by 601 equally spaced points.

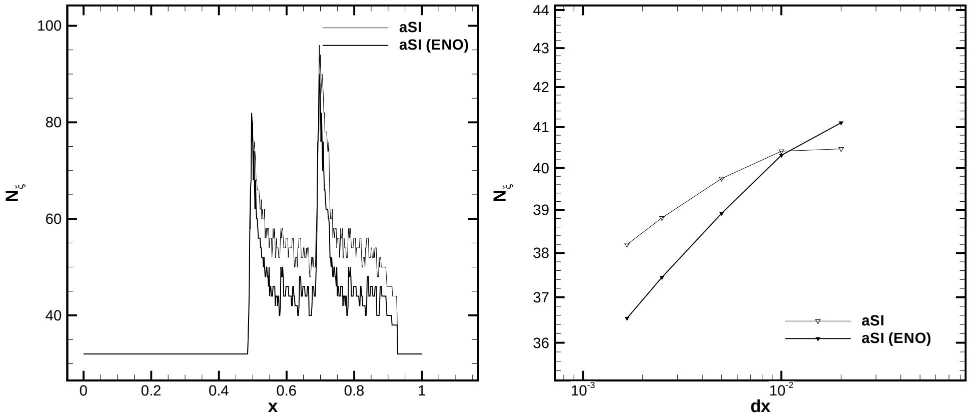


(a) (b)

Figure 8. Variance for the cell averaged solution of the linear advection equation with discontinuous initial condition (63) at the final time $t = 0.4$. The whole physical domain is represented in (a), while in the figure (b) a zoom in the shock region is reported. The physical mesh is constituted by 601 equally spaced points.

The ability of aSI scheme to refine only locally the space allows to increase locally the resolution along the stochastic space. In figure 9, the distribution of the stochastic cells over Ω at the final time step $t = 0.4$ is reported. It is evident that the higher computational effort is located in the region of the strong gradients; comparing the figure 8 and 9, it is evident that the two peaks associated to the local higher computational effort (in terms of stochastic cells) corresponds to the two peaks in the variance of the solution. In figure 9, the number of points employed by the aSI scheme with and without the ENO selection of the stencil are also reported. The ENO selection of the stencil reduces the number of cells employed. Morevoer, comparing the average number of stochastic cells employed for each computation, it is evident that the efficiency of the ENO selection increases with the spatial resolution. This is due to the global representation of the solution $u(x, \xi, t)$

over cells $\mathcal{C}_i \times \Xi_j$. Higher is the spatial resolution, sharper are the resulting discontinuities, so the ENO becomes more useful in order to gain in terms of accuracy (with the SI algorithm) and in terms of compression capabilities (with the TE algorithm). Figure 9(b) displays that for too coarse spatial resolution, the ENO selection of the stencil can be negative in terms of both accuracy and compression. The solution becomes smoother and smoother by decreasing the spatial resolution, so a centered stencil becomes the best choice.



(a) (b)

Figure 9. Evolution of the number of stochastic cells employed in each physical location for the aSI scheme with and without the ENO reconstruction (a) for the linear advection equation with discontinuous initial condition. The average number of stochastic cells employed by the aSI scheme as function of the physical space resolution is reported in (b).

6.2. *Inviscid Burgers equation*

In this section, the aSI algorithm is applied to the solution of the inviscid Burgers equation

$$\frac{\partial u(x,\xi,t)}{\partial t} + \frac{\partial f(u(x,\xi,t))}{\partial x} = 0 \quad x \in [0,1] \quad \text{and} \quad t \in [0,T], \tag{70}$$

where the flux function is defined as $f = f(u(x,\xi,t)) = \frac{1}{2}u^2(x,\xi,t)$.

We assume the following uncertain initial condition, with the random parameter uniformly distributed $\xi \sim \mathcal{U}[0,1]$,

$$u(x,\xi,0) = \begin{cases} H(\xi) & \text{if} \quad x \in [A_1^x, A_2^x] \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \tag{71}$$

The initial condition is represented by a hat function with a different amplitude dependent (non linearly) from the random parameter, $H(\xi) = \frac{1}{3}\xi^2 + \frac{1}{100}\xi + \frac{9}{10}$. To obtain the exact solution it is necessary to consider the two elementary solutions of the Riemann problem of the inviscid Burgers equation (see [23] for further details). The first case at the left of the hat function ($x = \frac{1}{10}$) is the Riemann problem with $u_l < u_r$ that admits as solution a rarefaction wave (depending on the uncertainty parameter) as follows

$$u(x,\xi,t) = \begin{cases} 0 & \text{if} \quad x \leq A_1^x \\ F(x) & \text{if} \quad x \in [A_1^x, A_1^x + H(\xi)t] \\ H(\xi) & \text{if} \quad x > A_1^x + H(\xi)t, \end{cases} \tag{72}$$

where the solution inside the rarefaction wave is $F(x) = (x + A_1^x)/t$.

Knowing the function $H(\xi)$, the exact solution for the uncertain rarefaction wave can be computed. Let us consider now the right of the hat initial function ($x = \frac{1}{2}$), where the solution of the Riemann problem is a shock wave traveling with an uncertain speed $s = H(\xi)/2$. The complete solution of the Riemann problem is then

$$u(x, \xi, t) = \begin{cases} H(\xi) & \text{if} \quad x < A_2^x + st \\ 0 & \text{if} \quad x > A_2^x + st. \end{cases} \tag{73}$$

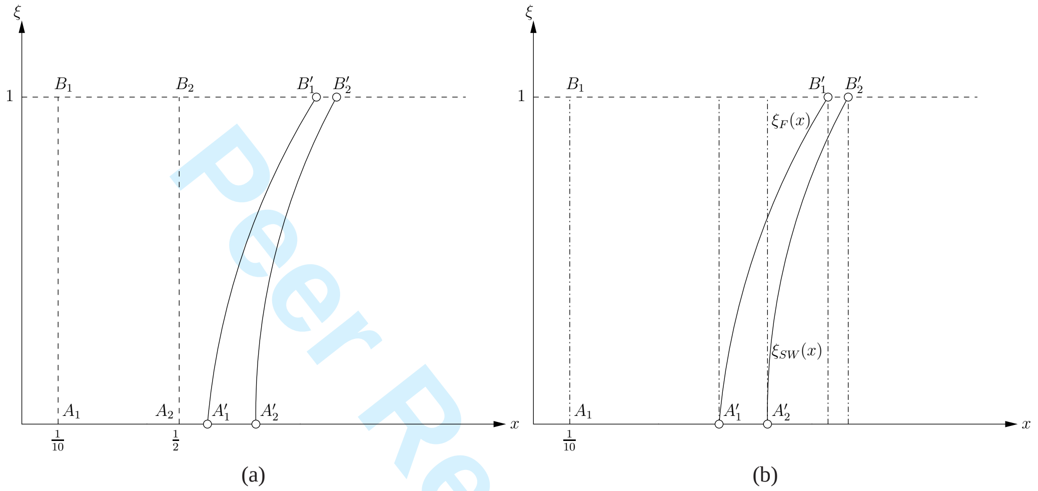


Figure 10. Schematic representation of the evolution between the initial condition (points A_1, A_2, B_1, B_2) and the final condition at time $t = 0.6$ (points A_1', A_2', B_1', B_2') (a). The regions in which the exact solution should be computed, at the final time $t = 0.6$, are reported in (b).

We solve the problem (70) until a time equal to $T = 0.6$, with the initial condition (71) defined by $A_1^x = B_1^x = \frac{1}{10}$ and $A_2^x = B_2^x = \frac{1}{2}$. The solution appears as sketched in figure 10, where the tail of the fan is at rest ($x = \frac{1}{10}$) while the position of the head is a function of the random parameter and its value is bounded between the slower moving fan ($A_1'^{,x} = \frac{1}{10} + H(0)t$) and the fast moving fan ($B_1'^{,x} = \frac{1}{10} + H(1)t$). The random parameter corresponding to a physical position $x \in [A_1'^{,x}, B_1'^{,x}]$ can be found after some algebraic manipulations analytically, by solving for ξ the equation $x = A_1^x + H(\xi)t$ for $A_1'^{,x} \leq x \leq B_1'^{,x}$, $\xi_F = \xi_F(x)$ (see figure 10 for the locus ξ_F). Following a similar procedure, the value of the random parameter corresponding to the shock position $\xi_{SW} = \xi_{SW}(x)$ can be found analytically, solving for ξ the equation $x = A_2^x + \frac{1}{2}H(\xi)t$ for $A_2'^{,x} \leq x \leq B_2'^{,x}$.

The statistics of the solution can be computed analytically for each cell $\mathcal{C}_i$ as follows. For the expectancy of the physical cell averaged value $\bar{u}_i$, it holds that

$$\mathcal{E}(\bar{u}_i) = \begin{cases} 0 & \text{if} \quad x_i \leq A_1^x \text{ or } x_i \geq B_2'^{,x} \\ F(x_i) & \text{if} \quad x_i \in [A_1, A_1'^{,x}] \\ \displaystyle\int_0^{\xi_F(x_i)} H(\xi)\mathrm{d}\xi + F(x_i)(1 - \xi_F(x_i)) & \text{if} \quad x_i \in [A_1'^{,x}, A_2'^{,x}] \\ \displaystyle\int_{\xi_F(x_i)}^{\xi_{SW}(x_i)} H(\xi)\mathrm{d}\xi + F(x_i)(1 - \xi_F(x_i)) & \text{if} \quad x_i \in [A_2'^{,x}, B_1'^{,x}] \\ \displaystyle\int_{\xi_{SW}(x_i)}^{1} H(\xi)\mathrm{d}\xi & \text{if} \quad x_i \in [B_1'^{,x}, B_2'^{,x}]. \end{cases} \tag{74}$$

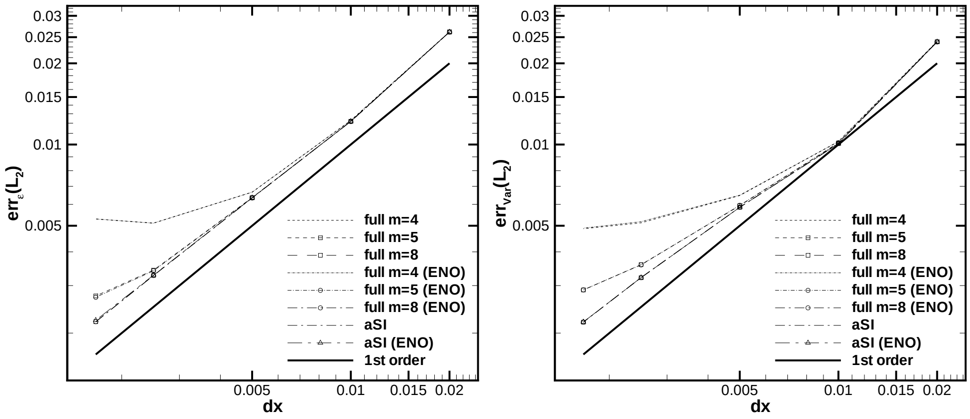
All the integrals in the equation (74) can be computed analytically.

DOI: 10.1002/fld

Moreover, the variance is easily analytically computed, due to the polynomial behavior of $H(\xi)$, as follows

$$\mathrm{Var}(\bar{u}_i) = \begin{cases} 0 & \text{if} \quad x_i \leq A_1^{\prime,x} \text{ or } x_i \geq B_2^{\prime,x} \\ \displaystyle\int_0^{\xi_F(x_i)} H^2(\xi)\mathrm{d}\xi + F^2(x_i)(1-\xi_F(x_i)) - \mathcal{E}^2(\bar{u}_i) & \text{if} \quad x_i \in [A_1^{\prime,x}, A_2^{\prime,x}] \\ \displaystyle\int_{\xi_F(x_i)}^{\xi_{SW}(x_i)} H^2(\xi)\mathrm{d}\xi + F^2(x_i)(1-\xi_F(x_i)) - \mathcal{E}^2(\bar{u}_i) & \text{if} \quad x_i \in [A_2^{\prime,x}, B_1^{\prime,x}] \\ \displaystyle\int_{\xi_{SW}(x_i)}^{1} H^2(\xi)\mathrm{d}\xi - \mathcal{E}^2(\bar{u}_i) & \text{if} \quad x_i \in [B_1^{\prime,x}, B_2^{\prime,x}]. \end{cases} \tag{75}$$

The (stochastic) inviscid Burgers problem (70) is solved over a set of equally spaced physical meshes with 51, 101, 201, 401 and 601 points. The time space is discretized using 600 time steps of constant length $\Delta t = 1 \times 10^{-3}$. The error norms in L_2, with respect to the exact stochastic solution (see equations (74) and (75)), are reported in figure 11. Similar results are obtained for L_1 and L_∞ norms, but are not reported here for brevity. The reference solution is the full computation performed with the SI scheme and a 256 ($m = 8$) equally spaced stochastic cells. This solution is compressed by means of the aSI scheme with a coarsest level of $m_0 = 4$ and a finest level of $m = 8$ with a threshold equal to $\varepsilon = 10^{-4}$. For both the full SI and the aSI schemes the computations are performed employing quadratic polynomial reconstruction with and without the ENO selection of the stencil. For each computation, the average number of stochastic cells is evaluated obtaining the equivalent number of equally spaced stochastic cells (with the same computational cost). The evolution of the number of stochastic cells associated to the different (physical) spatial resolutions are reported in figure 14(b) for the aSI scheme with and without the ENO procedure. Moreover, SI scheme is applied over 16 ($m = 4$) and 32 ($m = 5$) equally spaced stochastic cells. These resolutions are chosen because the average number of stochastic cells employed by the aSI scheme varies between these values. The SI scheme fails to converge with the expected first order slope both with and without the ENO, because of the appearance of the staircase phenomenon. The stochastic resolution is not high enough with respect to the physical resolution, as evident looking at the three last spatial resolutions in figure 11.



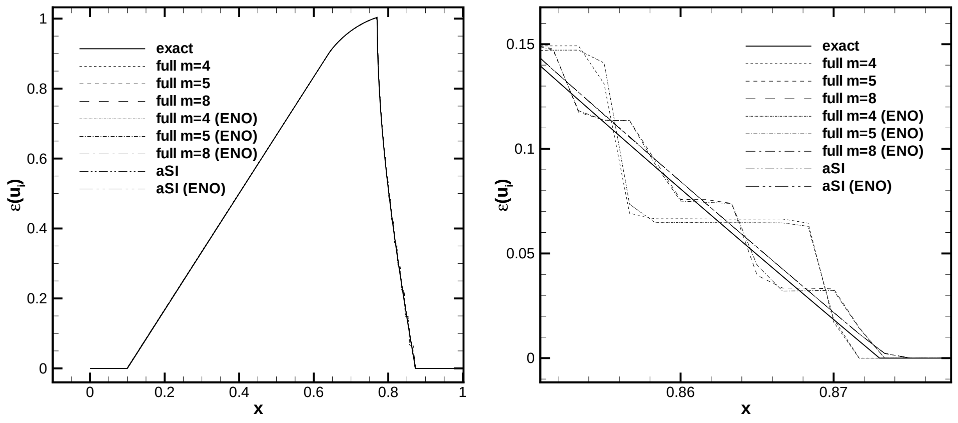
(a)

(b)

Figure 11. Spatial convergence for the Burgers equation with an uncertain hat initial condition (71). The statistics of the solution (mean (a) and variance (b)) obtained with (aSI) and without (full) compression are reported for the MHM method with superbee limiter (48).

DOI: 10.1002/fld

The staircase phenomenon is evident in figures 12 and 13, where the expectancy and the variance of the solution are reported over the 601 points physical mesh (the exact solution is evaluated over a mesh of 2001 equally spaced points). In particular, figures 12(b) and 13(b) show a zoom of the curves in the region, where the (uncertain) shock wave propagates (see figure 10). As expected, increasing the number of stochastic cells, even equally spaced, reduces the staircase phenomenon (from 16 to 32 cells). It disappears at 256 cells. Note that the aSI scheme, with an overall computational cost similar to the two coarse full simulations, produces better results (without the appearance of the staircase phenomenon) concentrating the computational effort, *i.e.* the number of cells, in the regions where the solution is less regular. The capability to refine and derefine during the simulation following the evolution of the solution in the physical/stochastic space makes the aSI scheme more efficient, yielding results that nearly coincide with the full reference solution.



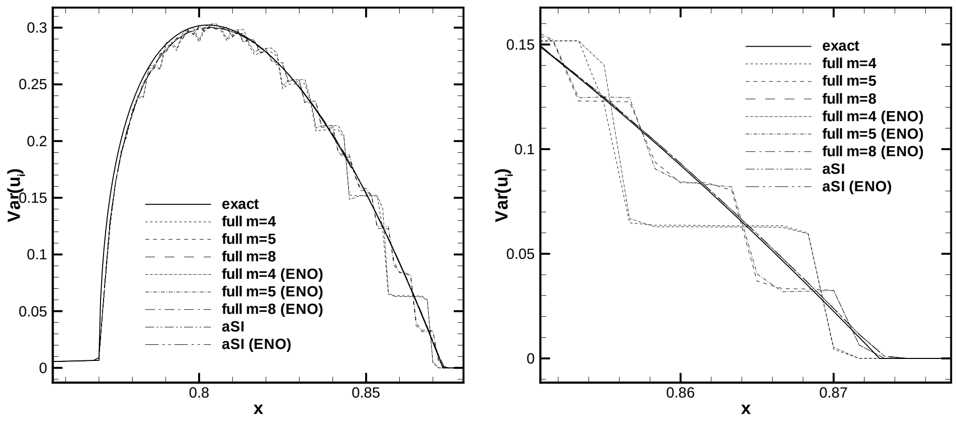
(a)

(b)

Figure 12. Expectancy for the cell-averaged solution of the inviscid Burgers equation at the final time $t = 0.6$. The whole physical domain is represented in (a), while in figure (b) a zoom in the shock region is reported. The physical mesh is constituted by 601 equally spaced points.

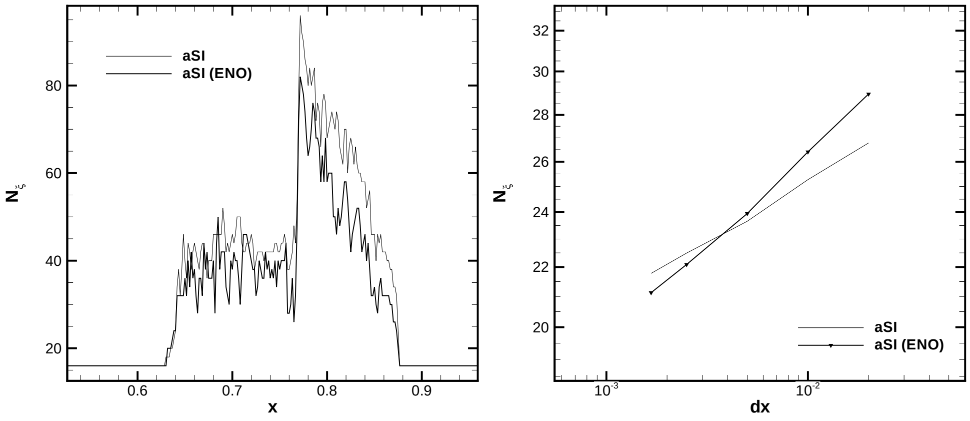
As already discussed for the solution of the linear advection equation with discontinuous initial condition, the presence of the ENO selection of the stencil makes the computations progressively more efficient increasing the physical resolution. This effect is evident in figure 14(b), where the (average) number of stochastic cells employed is reported as a function of the physical resolution. In figure 14(a), the direct comparison between the aSI scheme with and without the ENO selection of the stencil over the finest 601 points physical mesh is shown. With lower resolution meshes, there is no advantage in using the ENO procedure due to the representation of the solution over cells in the overall physical/stochastic space. However, the slope associated to the average number of stochastic cells shows that the solutions are represented by a narrow discontinuity (due to the increase of the spatial resolution). As a consequence, the non-oscillatory interpolation helps to avoid the so-called *pollution* of the stencil, *i.e.* the propagation of the interpolation error in the neighboring cells of a discontinuity. Again, the combination of the aSI scheme and the use of the ENO procedure for the polynomial interpolation, becomes even more efficient as the spatial resolution is increased. This is a desired property for any intrusive UQ scheme.

In the following section, the aSI scheme is applied to non linear system of stochastic partial differential equations.



(a) (b)

Figure 13. Variance for the cell-averaged solution of the inviscid Burgers equation at the final time $t = 0.6$. Two different zooms in the shock region are reported. The physical mesh is constituted by 601 equally spaced points.



(a) (b)

Figure 14. Evolution of the number of stochastic cells employed for each physical location for the aSI scheme with and without the ENO reconstruction (a) for the inviscid Burgers equation. The average number of stochastic cells employed by the aSI scheme as a function of the physical space resolution is reported in (b).

6.3. *Uncertain shock tube*

In this section, the solution of the uncertain shock tube problem is reported. The problem can be modeled by the well-known 1D Euler equations

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{0} \tag{76}$$

where the vector of conservative variables, the density ρ, the momentum $m = \rho u$ and the total Energy E^t, $\mathbf{u} \in \mathbb{R}^3$ and the flux vector $\mathbf{f}(\mathbf{u}) \in \mathbb{R}^3$ are

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ E^t \end{pmatrix} \qquad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} m \\ \dfrac{m^2}{\rho} + \Pi(\mathbf{u}) \\ \dfrac{m}{\rho}\left(E^t + \Pi(\mathbf{u})\right). \end{pmatrix} \tag{77}$$

The pressure $\Pi(\mathbf{u})$ (as function of the conservative variables) can be derived for a polytropic ideal gas as follows

$$\Pi(\mathbf{u}) = (\gamma - 1)\left(E^t - \frac{1}{2}\frac{|m^2|}{\rho}\right). \tag{78}$$

The initial condition for the uncertain shock tube problem is derived from the classical Sod test case [25], where an uncertainty of the density at the left state ($x < x_d$ for $t = 0$) is introduced:

$$\mathbf{u}_L(x, \xi, t) = \begin{pmatrix} \rho_L(\xi) \\ 0 \\ \dfrac{p_L}{\gamma - 1} \end{pmatrix} \qquad \mathbf{u}_R(x, \xi, t) = \begin{pmatrix} \rho_R \\ 0 \\ \dfrac{p_R}{\gamma - 1}. \end{pmatrix}, \tag{79}$$

In particular, the density on the left state is dependent from an uniformly distributed random parameter $\xi \sim \mathcal{U}[0, 1]$: $\rho_L(\xi) = 0.3 + 1.6\xi$. The values of the pressures are $p_L = 1$ and $p_R = 0.1$, while the right value of the density is $\rho_R = 0.125$. The total energy E^t is obtained (considering the gas at the rest in the whole domain) as a function of the local pressure and the ratio between specific heats, that for a diatomic gas can be assumed equal to $\gamma = 1.4$.

As pointed out by Toro [24], analyzing the eigenvalue structure of the Euler equations, the Riemann problem for the 1D Euler equations (see figure 15) generates (for $t > 0$) four states, where two are not known (variables are indicated with a star in the following). The Riemann problem for the solution of the 1D Euler equation can be reduced to the solution of a single non-linear algebraic equation for the pressure in the star region $p^\star$ from which the other quantities can be computed. With an uncertain shock tube problem, the dependence of $p^\star$ from the random parameter $p^\star = p^\star(\xi)$ should be considered. Unfortunately, this dependence cannot be computed explicitly. In this paper, only the case involving a left moving rerefaction fan and a right moving shock wave are considered. Moreover, initial conditions (79) produce this wave structure for all the random parameter taken into account. The problem is further complicated by the presence of complex functions that should be integrated to compute the exact statistics required. The solution strategy employed is the following. For each physical location, where the exact statistics should be computed, the solution along the stochastic space is divided into smooth regions (where the numerical quadrature with a large number of points produces fair well-converged results even for non-polynomial functions). The main issue is to determine the location of a discontinuity. This task can be accomplished solving an algebraic non-linear equation for the random parameter that can be formulated to involve all (but not only) the derivative available for the solution of the deterministic Riemann problem. After the subdivision of the random space in more regions, where the quadrature can be done numerically without accuracy loss (to the desired global accuracy), the statistics are computed in order to obtain the desired reference solutions.

Details of the numerical procedure to obtain the reference solution of the stochastic Riemann problem are reported in the Appendix .1.

Simulations are performed over a physical domain $\Omega = [-\frac{1}{5}, \frac{6}{5}]$ until a final time $t = 0.31$ with the position of the diaphragm equal to $x_d = 0.42$. The time space is divided in 6200 equal time steps of length $\Delta t = 5 \times 10^{-5}$. The simulations are carried out over equally spaced meshes of 201, 401, 801 and 1001 points employing the aSI scheme based on the MHM with a van Leer limiter (see equation (49)).

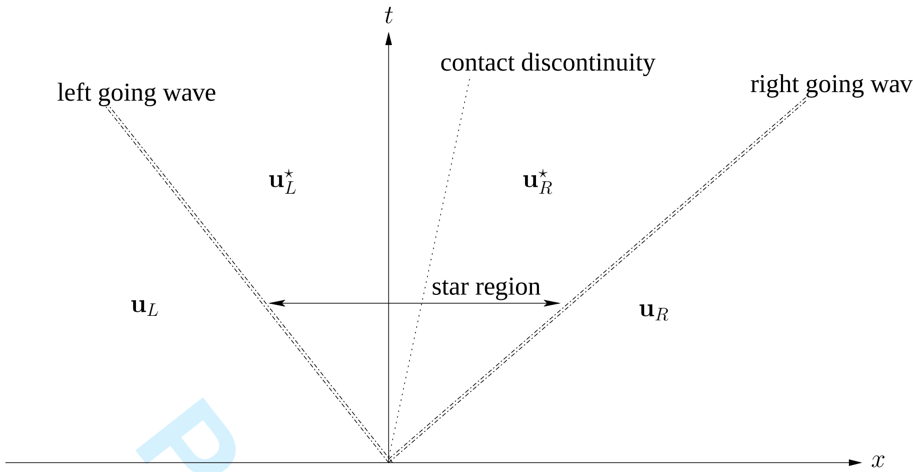
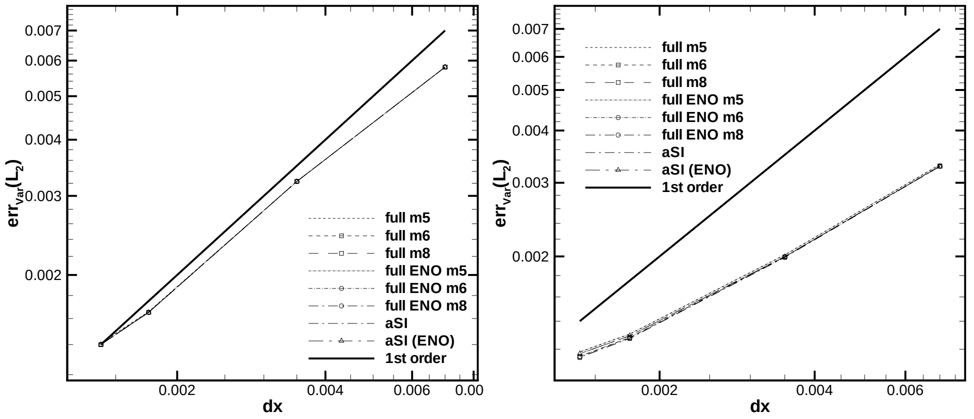


Figure 15. Riemann wave structure for the 1D Euler equation.

In figure 16, the spatial convergence is reported for both the mean (16(a)) and the variance (16(b)) in L_2 for the density ρ. The aSI method is obtained with a coarsest level of 4 ($m_0 = 2$) cells and a finest level of 256 ($m = 8$) stochastic cells with $\varepsilon = 10^{-4}$, while the reference solution is the full SI scheme with 256 cells. The aSI scheme has used an average number of stochastic cells between the two levels $m = 5$ and $m = 6$ (see figure 19(b)), so the other solutions are computed by means of the SI scheme for comparison. For all the schemes, both the centered second order polynomial reconstruction and the non-linear ENO one are used. The difference between the two polynomial reconstructions is difficult to appreciate because the spatial resolution is too poor for a sharp representation of the discontinuities. In this sense, there is no advantage in using the ENO reconstruction (for the aSI scheme and the SI scheme). The first order of convergence is attained for the expectancy of the density ρ, while, even with the SI scheme, the variance exhibits a lower rate of convergence 16(b). This behavior clearly indicates that even the solution employing 256 stochastic cells is not fully converged for moments higher than the expectancy.



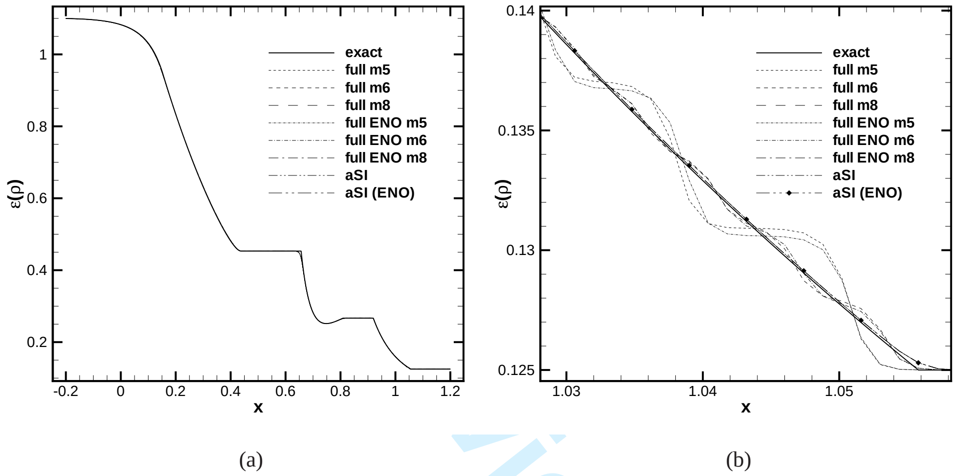
(a)

(b)

Figure 16. Spatial convergence for the stochastic shock tube problem equation with uncertain initial condition (79). The statistics of the solution (mean (a) and variance (b)) obtained with (aSI) and without (full) compression are reported for the MHM method with van Leer limiter (49).

DOI: 10.1002/fld

However, the aSI scheme displays the required properties: it saves the order of accuracy of the full SI scheme, both for mean and variance (see figure 16), and the degradation of the accuracy is strongly limited. Moreover, as already shown in the previous numerical results, the phenomenon of the staircase approximation of the statistics is prevented by the adaptation in the stochastic space. As shown in figure 17, note that all the numerical solutions are very similar to the exact solution 17(a) obtained over a mesh of 2001 equally spaced points in the physical space. By zooming (17(b)), the presence of the typical staircase phenomenon for both the SI scheme with 32 and 64 stochastic cells appears. The solution obtained with the aSI scheme agree very well with its full counterparts.



(a)

(b)

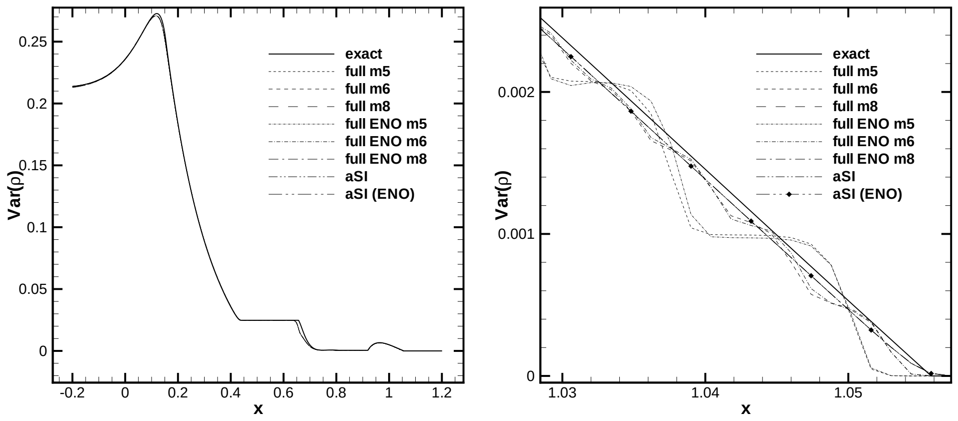
Figure 17. Density Expectancy for the cell averaged solution of the uncertain shock tube problem at the final time $t = 0.31$. The whole physical domain is represented in (a), while in the figure (b) a zoom in the shock region is reported. The physical mesh is constituted by 1001 equally spaced points.

As already demonstrated for the mean, even for the variance, the presence of the staircase approximation is prevented by the refinement of the stochastic space (see figure 18). Even if curves nearly coincide in figure 18(a), in the shock region the presence of the typical step pattern is evident for the full SI solution with 32 and 64 equally spaced stochastic cells (see 18(b)).

The lower order of convergence attained for the variance, even for the non compressed solution, highlights that the error in the stochastic space dominates the global error. As already demonstrated, the efficiency of the ENO selection of the stencil is related to the sharp representation of the discontinuities. In this case, the results with and without the ENO selection of the stencil are very similar. No advantages, even in term of compression, are observed. This issue is evident in figure 19(a), where the number of stochastic cells, along the physical domain, are reported. The region associated to the discontinuity spreads over a larger domain and, globally, the presence of non-centered stencils degradates the quality of prediction. This issue is well known in the ENO literature [26]. A possible cure, outside the scope of the present paper, would be the introduction of WENO type of interpolation. Employing a WENO type of interpolation, the correct centered stencil could be recovered without strong degradation of the prediction (the author already introduced a WENO interpolation in [10] in the context of the MR point-value setting).

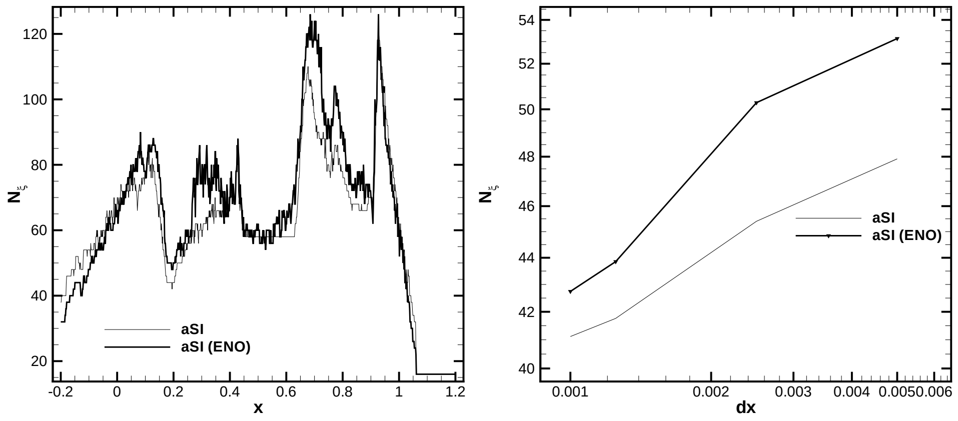
The evolution of the average number of stochastic cells employed by the aSI scheme with and without the ENO interpolation is reported in 19(b). In this case, there is no intersection between the two curves, revealing that in this case the ENO interpolation gives no advantage, even for high physical space resolutions.

DOI: 10.1002/fld



(a) (b)

Figure 18. Density Variance for the cell averaged solution of the uncertain shock tube problem at the final time $t = 0.31$. The whole physical domain is represented in (a), while in the figure (b) a zoom in the shock region is reported. The physical mesh is constituted by 1001 equally spaced points.



(a) (b)

Figure 19. Evolution of the number of stochastic cell employed for each physical location for the aSI scheme with and without the ENO reconstruction for the shock tube problem. The average number of stochastic cells employed by the aSI scheme as function of the physical space resolution has been reported in (b).

7. CONCLUDING REMARKS

In this paper, a novel adaptive intrusive numerical scheme for Uncertainty Quantification has been presented. The classical MR Harten framework, in its cell average setting, has been here extended to include the dependence from a generic probability density function. Moreover, an original algorithm has been developed to obtain the solution at the finest resolution level starting from the coarsest one. The aim is to obtain, at the same time, a saving in memory requirements and in the computational cost associated to the true model evaluation of the system. This general algorithm has been coupled with the Semi-Intrusive (SI) scheme for UQ proposed by Abgrall and Congedo [7]. The overall numerical scheme is the so-called adaptive-SI scheme. We demonstrated that it preserves the convergence properties of the original SI scheme with a strong saving in term of computational cost. Different test-cases have been presented to demonstrate the efficiency and the accuracy properties of the aSI scheme. The linear advection equation has been solved for initial smooth and discontinuous solution to demonstrate the capability of the stochastic scheme to preserve the accuracy related to the deterministic MUSCL-Hancock method (MHM). A second test-case has been focused on the inviscid Burgers equation. We demonstrated the capability of the method to automatically refine/derefine following the changes in the regularity of the solution in the coupled stochastic/physical space. In particular, a smooth solution has been considered, in the stochastic space, as initial condition, where shock waves velocities are directly related to the parameter in the stochastic space. The final test case proposed has been the Euler system of equation to solve an uncertain shock tube problem. The aSI scheme has been demonstrated to be efficient also in the case of vectorial problems. For the computation of the convergence curves, an original strategy for the semi-analytical solution of the stochastic shock tube problem has been also developed following and extending the classical numerical procedure for the solution of the Riemann problem for the Euler equations. This paper constituted the first effort to introduce a MR framework into the SI method. The generality of the approach is not limited to second order scheme, but can be easily extended to higher order numerical formulation for the physical space and time discretizations. In the present work, both the linear and non-linear MR framework have been presented in which the selection of the stencil to obtain the reconstruction operators can be obtained by a data-dependent procedure. The ENO selection of the stencil has been also introduced. Considering the numerical results presented, note that the advantages related to the non-linear schemes are very limited. This issue is related not to the non-linear procedure itself but to the peculiarity of the SI scheme that produces representations of the solution in a combined physical stochastic space. The representation of discontinuous solution along the stochastic space can recover a smoother behavior when the physical spatial resolution is not high enough. This has been demonstrated showing that the importance of the ENO scheme increases with the physical space resolution. To improve the global properties of the scheme, two further steps seem useful. The first is the introduction of the WENO reconstruction instead of the ENO interpolation recovering the correct stencil in all the regions in which the solution is smooth, as it has been already presented for the point-value setting [10]. The other step could be to increase the polynomial order for the reconstruction. This should improve both accuracy and compression capabilities. The extension and the analysis of the aSI scheme for a moderate number of dimension is actually underway.

ACKNOWLEDGEMENT

Rémi Abgrall and Gianluca Geraci have been supported by the ERC Advanced Grant ADDECCO N°226316. The authors greatly acknowledge also the Associated Team AQUARIUS (Joint team from INRIA and Stanford University) for funding the stay of Gianluca Geraci and Pietro Marco Congedo at Stanford University in which part of the present work has been conceived.

.1. Accurate numerical solution for the 1D stochastic Riemann problem for the Euler equations

In this section, let us illustrate the numerical procedure to obtain the reference solution for the stochastic shock tube problem (the interesting reader may refer to [24] for a complete description of the deterministic problem). Let us consider a deterministic Riemann problem for the 1D Euler equations, in particular the case of a left going rarefaction wave and a right moving shock wave. This assumption do not pose any limitation on the general procedure for the solution of the stochastic problem presented here.

The solution of the deterministic Riemann problem (for gas initially at the rest) consists in solving a non-linear equation for the pressure in the region between the shock and the contact discontinuity. Remark that each quantity is dependent on the random parameter. In the deterministic case, the random parameter is obviously assumed as a constant. In the following, the explicit dependence of each quantity with respect the random parameter ξ is explicitly reported for the uncertain initial left stat (see equation (79)). However the dependence of all the quantities from the random parameter must be considered redundant if the deterministic case is of interest because in that case all the uncertain parameters assume a fixed value.

The non-linear equation to solve for the pressure in the star region $p^\star$ is the following

$$\begin{aligned} f(p^\star(\xi), \mathbf{u}_L(\xi), \mathbf{u}_R) &= f_L(p^\star(\xi), \mathbf{u}_L(\xi)) + f_R(p^\star(\xi), \mathbf{u}_R) \\ &= \frac{2a_L(\xi)}{\gamma-1}\left[\left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} - 1\right] + \left(p^\star(\xi) - p_R\right)\left[\frac{\frac{2}{(\gamma+1)\rho_R}}{p^\star(\xi) + \frac{\gamma-1}{\gamma+1}p_R}\right]^{\frac{1}{2}} = 0, \end{aligned} \tag{80}$$

with the speed of sound $a_L(\xi) = \sqrt{\gamma \frac{p_L}{\rho_L(\xi)}}$.

The equation (80) is solved by means of an iterative Newton-Raphson scheme following [24]

$$\begin{cases} \Delta p^\star = -\left(\dfrac{\mathrm{d}f(p^\star(\xi), \mathbf{u}_L(\xi), \mathbf{u}_R)}{\mathrm{d}p^\star}\Big|_{p^\star(\xi)=p^\star_k(\xi)}\right)^{-1} f(p^\star_k(\xi), \mathbf{u}_L(\xi), \mathbf{u}_R) \\ p^\star_{k+1}(\xi) = p^\star_k(\xi) + \Delta p^\star. \end{cases} \tag{81}$$

The initial condition is systematically $p^\star_0 = \frac{p_L+p_R}{2}$) and $\Delta p^\star \leq 10^{-14}$ is chosen as convergence criterion. Derivative of the function $f(p^\star(\xi), \mathbf{u}_L(\xi), \mathbf{u}_R)$ with respect to $p^\star$ that can be computed as follows

$$\begin{cases} \dfrac{\mathrm{d}f(p^\star(\xi), \mathbf{u}_L(\xi), \mathbf{u}_R)}{\mathrm{d}p^\star(\xi)} = \dfrac{\mathrm{d}f_L(p^\star(\xi), \mathbf{u}_L(\xi))}{\mathrm{d}p^\star} + \dfrac{\mathrm{d}f_R(p^\star(\xi), \mathbf{u}_R)}{\mathrm{d}p^\star} \\ \dfrac{\mathrm{d}f_L(p^\star(\xi), \mathbf{u}_L(\xi))}{\mathrm{d}p^\star} = \dfrac{1}{\gamma p^\star(\xi)}\sqrt{\gamma\dfrac{p_L}{\rho_L}}\left(\dfrac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} \\ \dfrac{\mathrm{d}f_R(p^\star(\xi), \mathbf{u}_R)}{\mathrm{d}p^\star} = \dfrac{2}{(\gamma+1)\rho_R\left(p^\star(\xi) + \frac{\gamma-1}{\gamma+1}p_R\right)}\left[1 - \dfrac{(p^\star(\xi) - p_R)}{\left(p^\star(\xi) + \frac{\gamma-1}{\gamma+1}p_R\right)}\right]. \end{cases} \tag{82}$$

Once computed the pressure $p^\star$, the particle velocity $u^\star$ can be computed according to

$$u^\star(\xi) = \frac{1}{2}\left(f_R(p^\star(\xi), \mathbf{u}_R) - f_L(p^\star(\xi), \mathbf{u}_L(\xi))\right), \tag{83}$$

while the density in the star region is defined as

$$\begin{aligned} \rho^\star_L(\xi) &= \rho_L(\xi)\left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{1}{\gamma}} \\ \rho^\star_R(\xi) &= \rho_R\left[\frac{\frac{p^\star(\xi)}{p_R} + \frac{\gamma-1}{\gamma+1}}{\frac{\gamma-1}{\gamma+1}\frac{p^\star(\xi)}{p_R} + 1}\right]. \end{aligned} \tag{84}$$

Now, let us determine the positions of the rarefaction wave, of the contact discontinuity and of the shock wave. In the following, $\mathbb{HF}$, $\mathbb{TF}$, $\mathbb{CD}$ and $\mathbb{SW}$ are used to name the head and tail of the rarefaction fan, the contact discontinuity and the shock waves respectively. These coordinates can be computed only as a function of the variable in the star region, $p^\star$ and $u^\star$, and of the left and right states $\mathbf{u}_L$ and $\mathbf{u}_R$ at a certain time t:

$$\begin{cases} \mathbb{HF}(\xi,t) = x_d - a_L(\xi)t \\ \mathbb{TF}(\xi,t) = x_d - (u^\star(\xi) - a^\star_L(\xi))t \\ \mathbb{CD}(\xi,t) = x_d + u^\star(\xi)t \\ \mathbb{SW}(\xi,t) = x_d + a_R\left[\dfrac{\gamma+1}{2\gamma}\dfrac{p^\star(\xi)}{p_R} + \dfrac{\gamma-1}{2\gamma}\right] \end{cases} \quad \text{where} \quad \begin{cases} a_L(\xi)^\star = a_L(\xi)\left(\dfrac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} \\ a_R = \sqrt{\gamma\dfrac{p_R}{\rho_R}}. \end{cases} \tag{85}$$

The complete solution of the Riemann problem is then (see also figure 20)

$$\mathbf{u}(x,\xi,t) = \begin{cases} \mathbf{u}_L(x,\xi,t) & \text{if} \quad x < \mathbb{HF}(\xi,t) \\ \mathbf{u}_F(x,\xi,t) & \text{if} \quad \mathbb{HF}(\xi,t) < x < \mathbb{TF}(\xi,t) \\ \mathbf{u}^\star_L(x,\xi,t) & \text{if} \quad \mathbb{TF}(\xi,t) < x < \mathbb{CD}(\xi,t) \\ \mathbf{u}^\star_R(x,\xi,t) & \text{if} \quad \mathbb{CD}(\xi,t) < x < sw(\xi,t) \\ \mathbf{u}_R(x,\xi,t) & \text{if} \quad x > \mathbb{SW}(\xi,t), \end{cases} \tag{86}$$

DOI: 10.1002/fld

where the solution inside the rarefaction fan is as follows

$$\mathbf{u}_F(x,\xi,t) = \begin{bmatrix} \rho_L(\xi)\left[\dfrac{2}{\gamma+1} - \dfrac{\gamma-1}{a_L(\xi)(\gamma+1)}r(x,t)\right]^{\frac{2}{\gamma-1}} \\ \dfrac{2\rho_L(\xi)}{\gamma+1}\left[a_L(\xi)+r(x,t)\right] \\ \dfrac{p_L}{\gamma-1}\left[\dfrac{2}{\gamma+1} - \dfrac{\gamma-1}{a_L(\xi)(\gamma+1)}r(x,t)\right]^{\frac{2\gamma}{\gamma-1}} \end{bmatrix} = \begin{bmatrix} \rho_F(\xi) \\ m_F(\xi) \\ E^t(\xi), \end{bmatrix} \tag{87}$$

while in the star region

$$\mathbf{u}^\star_L(x,\xi,t) = \begin{bmatrix} \rho^\star_L(\xi) \\ \rho^\star_L(\xi)u^\star(\xi) \\ \dfrac{p^\star(\xi)}{\gamma-1} \end{bmatrix} \quad \text{and} \quad \mathbf{u}^\star_R(x,\xi,t) = \begin{bmatrix} \rho^\star_R(\xi) \\ \rho^\star_R(\xi)u^\star(\xi) \\ \dfrac{p^\star(\xi)}{\gamma-1} \end{bmatrix}. \tag{88}$$

A similarity variable $r(x,t)$, defined as $r(x,t) = \frac{x-x_d}{t}$, is introduced.

Note that if a value for the random parameter is fixed, the previous procedure coincide with the classical solution of the Riemann problem as reported in [24]. However, here the interest is the computations of the statistics, the expectancy and the variance, of the solution $u(x,\xi,t)$. To obtain the statistics, the solution $u(x,\xi,t)$ has to be integrated numerically splitting the random space. In particular, the integration is carried out by dividing the computational domain of the stochastic space according to (86). The complete solution of the stochastic Riemann problem for the Euler equation using the initial conditions (79), consists in capturing four structures: the region of points describing the position of the head and tail of the rarefaction wave, the contact discontinuity and the shock wave. For each zone, it is necessary to find the random parameter

$$\xi_d : x = g(\xi_d, t) \quad \forall (x,t) \in \Omega \supseteq D(t) \times T, \tag{89}$$

where the function $g(\xi,t)$ can be one of the region reported in (85). It is assumed here that functions g are monotone functions with respect to the random parameter. The subset of the physical space $D(t)$ can be defined considering the union of all the images of the functions describing the physical position of the discontinuities

$$D(t) = [\mathrm{HF}_{\min}(t), \mathrm{HF}_{\max}(t)] \cup [\mathrm{TF}_{\min}(t), \mathrm{TF}_{\max}(t)] \cup [\mathrm{CD}_{\min}(t), \mathrm{CD}_{\max}(t)] \cup [\mathrm{SW}_{\min}(t), \mathrm{SW}_{\max}(t)]. \tag{90}$$

Note that for each (x,t), more than one ξ_d corresponding to the intersections with different regions could exist, but not multiple intersections with the same region. The case of multiple intersections is determined by a non-null intersection between two or more images of the g functions. The monotonicity of the g function implies that the extrema of g correspond to the bounds of the stochastic space. This property is useful from a practical point of view because for each time step the domain $D(t)$ can be easily determined by (90).

Intersections should be computed solving the non-linear algebraic equations (89) by using Newton-Raphson techniques. Let us focus now on the four regions.

The intersection between the line x and the head fan $\mathbb{HF}(\xi,t)$ can be obtained as follows

$$x = \mathbb{HF}(\xi,t) = x_d - a_L t = x_d - \sqrt{\gamma\frac{p_L}{\rho_L(\xi)}} \quad \rightarrow \quad \rho_L(\xi) = \frac{\gamma p_L}{r^2(x,t)}. \tag{91}$$

If density ρ_L is linearly dependent on the random parameter $\xi \sim \mathcal{U}[0,1]$ (as presented in section 6.3) the value of the intersection is equal to

$$\xi_d = \frac{1}{\rho_L(1) - \rho_L(0)}\left(\frac{\gamma p_L}{r^2(x,t)} - \rho_L(0)\right). \tag{92}$$

Concerning the tail of the rarefaction wave, it follows that

$$x = x_d + \left(\frac{1}{2}\left[f_r(p^\star(\xi),\xi) - f_L(p^\star(\xi),\xi)\right] - a_L(\xi)\left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}}\right)t \tag{93}$$

where both functions f_L and f_R are dependent on the random parameter (omitting the dependence from the left $\mathbf{u}_L(\xi)$ and right states $\mathbf{u}_R$).

The problem is to find the root of the function $F(p^\star(\xi),\xi)$

$$\begin{aligned} F(p^\star(\xi),\xi) &= r(x,t) - f_R(p^\star(\xi)) + \sqrt{\gamma \frac{p_L}{\rho_L(\xi)}} \left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}} \\ &= r(x,t) - f_R(p^\star(\xi)) + C(p^\star(\xi),\xi) = 0, \end{aligned} \tag{94}$$

where the relation (83) is injected in the previous equation.

The iterative procedure for the solution of (94) is the following

$$\begin{cases} \Delta\xi = -\left(\left.\dfrac{\mathrm{d}F(p^\star(\xi),\xi)}{\mathrm{d}\xi}\right|_{\xi=\xi_k}\right)^{-1} F(p^\star(\xi_k),\xi_k)) \\ \xi_{k+1} = \xi_k + \Delta\xi, \end{cases} \tag{95}$$

where the differential is equal to

$$\begin{aligned} \frac{\mathrm{d}F(p^\star(\xi),\xi)}{\mathrm{d}\xi} &= \frac{\partial F(p^\star(\xi),\xi)}{\partial\xi} + \frac{\partial F(p^\star(\xi),\xi)}{\partial p^\star}\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi} \\ &= \frac{\partial C(p^\star(\xi),\xi)}{\partial\xi} + \left(-\frac{\mathrm{d}f_R(p^\star(\xi))}{\mathrm{d}p^\star} + \frac{\partial C(p^\star(\xi),\xi)}{\partial p^\star}\right)\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi}. \end{aligned} \tag{96}$$

The derivative of the function $C(p^\star(\xi),\xi)$ has to be computed as well as the derivative $\dfrac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi}$, while $\dfrac{\mathrm{d}f_R(p^\star(\xi))}{\mathrm{d}p^\star}$ is already available (see equation (82)).

The derivatives of $C(p^\star(\xi),\xi)$ are

$$\begin{aligned} \frac{\partial C(p^\star(\xi),\xi)}{\partial\xi} &= -\frac{1}{2}\frac{\left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}}\gamma p_L\left(\rho_L(1)-\rho_L(0)\right)}{\rho_L^2(\xi)\sqrt{\gamma\frac{p_L}{\rho_L(\xi)}}} \\ \frac{\partial C(p^\star(\xi),\xi)}{\partial p^\star} &= \frac{1}{2}\frac{\sqrt{\gamma\frac{p_L}{\rho_L(\xi)}}\left(\frac{p^\star(\xi)}{p_L}\right)^{\frac{\gamma-1}{2\gamma}}(\gamma-1)}{\gamma p^\star(\xi)}, \end{aligned} \tag{97}$$

while, at each time step, the derivative $\dfrac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi}$ can be approximated by means of a backward difference

$$\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi} \simeq \frac{p^\star(\xi_{k+1}) - p^\star(\xi_k)}{\xi_{k+1}-\xi_k}, \tag{98}$$

since $p^\star$ is not known explicitly.

From a practical point of view, the initial guess ξ_0 is chosen as the solution of the linear approximation for $\mathrm{TF}(\xi,t)$ between the extrema of the stochastic domain, with $\dfrac{\mathrm{d}p^\star(\xi_0)}{\mathrm{d}\xi} = 0.1$.

Considering the intersection with the contact discontinuity, it follows that

$$\begin{aligned} F(p^\star(\xi),\xi) &= r(x,t) - u^\star(\xi) \\ &= r(x,t) - \frac{1}{2}\big(f_R(p^\star(\xi)) - f_L(p^\star(\xi),\xi)\big) = r(x,t) - f_R(p^\star(\xi)) = F(p^\star). \end{aligned} \tag{99}$$

The iterative procedure is formally equal to (95) (even if here the dependence is not explicit with respect to ξ), with a different differential term

$$\begin{aligned} \frac{\mathrm{d}F(p^\star(\xi))}{\mathrm{d}\xi} &= \frac{\mathrm{d}F(p^\star(\xi))}{\mathrm{d}p^\star}\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi} \\ &= -\frac{\mathrm{d}f_R(p^\star(\xi))}{\mathrm{d}p^\star}\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi}. \end{aligned} \tag{100}$$

This differential can be computed according to (82) and (98).

Finally, the intersection with the shock waves is demanded. The non linear algebraic equation results

$$\begin{aligned} F(p^\star(\xi),\xi) &= r(x,t) - a_R\left[\frac{\gamma+1}{2\gamma}\frac{p^\star}{p_R} + \frac{\gamma-1}{2\gamma}\right]^{\frac{1}{2}} \\ &= r(x,t) - A(p^\star(\xi)) = F(p^\star(\xi)). \end{aligned} \tag{101}$$

Again, the formal iterative procedure (95) can be employed with $\frac{\mathrm{d}F(p^\star(\xi))}{\mathrm{d}\xi} = \frac{\mathrm{d}F(p^\star(\xi))}{\mathrm{d}p^\star}\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi} = -\frac{\mathrm{d}A(p^\star(\xi))}{\mathrm{d}p^\star}\frac{\mathrm{d}p^\star(\xi)}{\mathrm{d}\xi}$, where

$$\frac{\mathrm{d}A(p^\star(\xi))}{\mathrm{d}p^\star} = \frac{1}{4}\frac{\gamma+1}{\gamma p_R\sqrt{\frac{(\gamma+1)p^\star(\xi)}{2\gamma p_R} + \frac{\gamma+1}{2\gamma}}}. \tag{102}$$

Let us sketch the reference solution in the plan $\xi - x$ at a final time equal to $t = 0.31$, with the initial position of the diaphragm $x_d = 0.42$ in the figure 20.

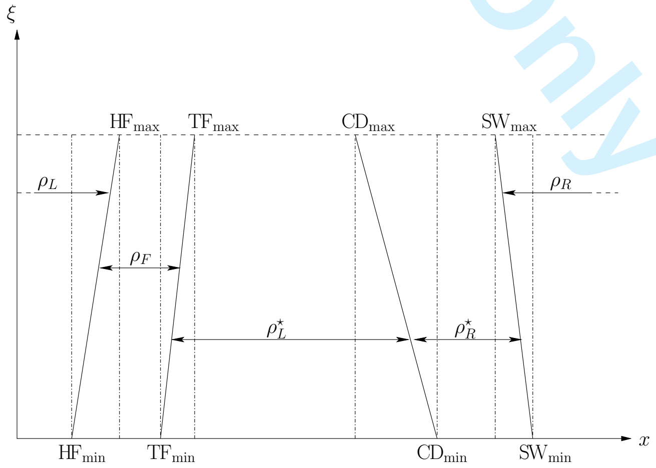


Figure 20. Schematic representation of the density of the uncertain shock tube problem in the combined physical/stochastic space. The regions where the solution should be subdivided are reported with the explicit identification of all the zones defining the variation of the solution as a function of the uncertain parameter ξ.

DOI: 10.1002/fld

The expectancy of the-cell averaged physical solution (at time $t = 0.31$) $u(x, \xi, t)$, for each cell $\mathcal{C}_i = [x_i - \frac{|\mathcal{C}_i|}{2}, x_i + \frac{|\mathcal{C}_i|}{2}]$ is computed as follows (here reported explicitly only for the density)

$$\mathcal{E}(\bar{u}_i) = \begin{cases} \dfrac{\rho_L(0) + \rho_L(1)}{2} & \text{if} \quad x_i \leq \mathrm{HF}_{\min} \\ \displaystyle\int_0^{\xi_d(x_i)} \rho_F(x,\xi)\mathrm{d}\xi + \rho_L(0)(1 - \xi_d(x_i)) + \dfrac{(\rho_L(1) - \rho_L(0))\,(1 - \xi_d^2(x_i))}{2} & \text{if} \quad \mathrm{HF}_{\min} \leq x_i \leq \mathrm{HF}_{\max} \\ \displaystyle\int_0^1 \rho_F(x,\xi)\mathrm{d}\xi & \text{if} \quad \mathrm{HF}_{\max} \leq x_i \leq \mathrm{TF}_{\min} \\ \displaystyle\int_0^{\xi_d(x_i)} \rho_L^\star(\xi)\mathrm{d}\xi + \int_{\xi_d(x_i)}^1 \rho_F(x,\xi)\mathrm{d}\xi & \text{if} \quad \mathrm{TF}_{\min} \leq x_i \leq \mathrm{TF}_{\max} \\ \displaystyle\int_0^1 \rho_L^\star(\xi)\mathrm{d}\xi & \text{if} \quad \mathrm{TF}_{\max} \leq x_i \leq \mathrm{CD}_{\min} \\ \displaystyle\int_0^{\xi_d(x_i)} \rho_L^\star(\xi)\mathrm{d}\xi + \int_{\xi_d(x_i)}^1 \rho_R^\star(\xi)\mathrm{d}\xi & \text{if} \quad \mathrm{CD}_{\min} \leq x_i \leq \mathrm{CD}_{\max} \\ \displaystyle\int_0^1 \rho_R^\star(\xi)\mathrm{d}\xi & \text{if} \quad \mathrm{CD}_{\max} \leq x_i \leq \mathrm{SW}_{\min} \\ \displaystyle\int_0^{\xi_d(x_i)} \rho_R^\star(\xi)\mathrm{d}\xi(1 - \xi_d(x_i))\rho_R & \text{if} \quad \mathrm{SW}_{\min} \leq x_i \leq \mathrm{SW}_{\max} \\ \rho_R & \text{if} \quad x_i \geq \mathrm{SW}_{\max}, \end{cases} \tag{103}$$

where the variance can be computed in a similar way (see what is done for the linear advection equation §6.1 and the Burgers equation (6.2)). All the numerical quadratures are performed over the stochastic (sub-domain discretized by means of 5000 equally spaced intervals employing a three points Gauss formula:

$$\int_a^b f(\xi)\mathrm{d}\xi = \frac{b-a}{2}\sum_{k=1}^{3} w_k f(\xi_k), \tag{104}$$

where $w_{1,3} = 5/9$, $w_2 = 8/9$, $\xi_{1,3} = \frac{b+a}{2} \pm \frac{b-a}{2}\sqrt{\frac{3}{5}}$ and $\xi_2 = \frac{b+a}{2}$.

REFERENCES

1. Graham I, Kuo F, Nuyens D, Scheichl R, Sloan I. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics* Feb 2011; **230**(10):3668–3694, doi:10.1016/j.jcp.2011.01.023.
2. Babuška I, Nobile F, Tempone R. A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. *SIAM Review* 2010; **52**(2):317, doi:10.1137/100786356.
3. Ghanem RG, Spanos PD. *Stochastic Finite Elements. A spectral approach*. Springer Verlag, 1991.
4. Xiu D, Karniadakis GE. Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics* May 2003; **187**(1):137–167, doi:10.1016/S0021-9991(03)00092-5.
5. Le Maître O, Knio O. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Springer Verlag, 2010.
6. Tryoen J, Le Maître O, Ndjinga M, Ern A. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems q. *Journal of Computational Physics* 2010; **229**:6485–6511.
7. Abgrall R, Congedo PM. A semi-intrusive deterministic approach to uncertainty quantifications in non-linear fluid flow problems. *Journal of Computational Physics* 2013; (235):828–845.
8. Abgrall R, Congedo PM, Geraci G. An adaptive multiresolution inspired scheme for solving the stochastic differential equations. *Proceedings of MASCOT 11, 11th Meeting on Applied Scientific Computing and Tools. Grid Generation, Approximation and Visualization. IMACS Series in Computational and Applied Mathematics Vol. 17, IAC - CNR, Rome, Italy*, 2011; 1–10.
9. Abgrall R, Congedo PM, Geraci G, Iaccarino G. Adaptive strategy in multiresolution framework for uncertainty quantification. *Center For Turbulence Research, Prooceedings of the Summer Program 2012*, 2012; 209–218.
10. Abgrall R, Congedo PM, Geraci G. A high-order non-linear multiresolution scheme for stochastic PDEs. *European Workshop on High Order Nonlinear Numerical Methods for Evolutionary PDEs: Theory and Applications (HONOM 2013)*, 2013.
11. Tryoen J. Methodes de Galerkin stochastiques adaptatives pour la propagation d'incertitudes parametriques dans les systemes hyperboliques. PhD Thesis, Univeriste Paris-Est.

Prepared using fldauth.cls
DOI: 10.1002/fld

12. Tryoen J, Le Maître O, Ern A. Adaptive Anisotropic Spectral Stochastic Methods for Uncertain Scalar Conservation Laws. *SIAM Journal Scientific Computing* 2012; **34**:A2459–A2481.
13. Harten A. Discrete multi-resolution analysis and generalized wavelets. *Applied Numerical Mathematics* 1993; **12**(13):153 – 192.
14. Harten A. Adaptive multiresolution schemes for shock computations. *Journal of Computational Physics* Aug 1994; **135**(2):260–278.
15. Harten A. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Communications on Pure and Applied Mathematics* 1995; **48**(12):1305–1342.
16. Abgrall R, Harten A. Multiresolution Representation in Unstructured Meshes. *SIAM Journal on Numerical Analysis* 1998; **35**(6):2128–2146.
17. Arandiga F, Donat R. Nonlinear multiscale decompositions: The approach of A. Harten. *Numerical Algorithms* 2000; **23**:175–216.
18. Arandiga F, Chiavassa G, Donat R. Harten framework for multiresolution with applications: From conservation laws to image compression. *Boletín SEMA* 2009; **31**(31):73–108.
19. Getreuer P, Meyer FG. ENO multiresolutions Schemes with General Discretizations. *SIAM Journal on Numerical Analysis* 2008; **46**(6):2953–2977.
20. Abgrall R, Sonar T. On the use of Mühlbach expansions in the recovery step of ENO methods. *Numerische Mathematik* 1997; (1997):1–25.
21. Quarteroni A, Sacco R, Saleri F. *Matematica Numerica*. Springer, 2008.
22. Abgrall R. On Essentially Non-oscillatory Schemes on Unstructured Meshes: Analysis and Implementation. *Journal of Computational Physics* Sep 1994; **114**(1):45–58.
23. LeVeque RJ. *Finite volume methods for conservation laws and hyperbolic systems*. Cambridge University Press, 2002.
24. Toro EF. *Riemann solvers and numerical methods for fluid mechanics*. Springer, Berlin, 1997.
25. Sod G. Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws. *Journal of Computational Physics* 1978; (27):1–31.
26. Aràndiga F, Belda AM. Weighted ENO interpolation and applications. *Communications in Nonlinear Science and Numerical Simulation* Apr 2004; **9**(2):187–195.

DOI: 10.1002/fld

Paper *P5*

Stochastic Discrete Equation Method (sDEM) for two-phase flows

R. Abgrall, P.M. Congedo, G. Geraci and M.G. Rodio[a]

[a] *INRIA Bordeaux Sud-Ouest, Talence, 33405 Cedex, France*

Abstract

A new scheme for the numerical approximation of a five-equations model taking into account uncertainty quantification (UQ) is presented. In particular, the Discrete Equation Method (DEM) for the discretization of the five-equations model is modified for including a formulation based on the adaptive Semi-intrusive (aSI) scheme, thus yielding a new intrusive scheme (sDEM) for simulating stochastic two-phase flows. Some reference test-cases are performed in order to demonstrate the convergence properties and the efficiency of the overall scheme. The propagation of initial uncertainties is evaluated in terms of mean and variance of several thermodynamic properties of the two phases.

Keywords: Uncertainty quantifications, adaptive Semi-Intrusive scheme (aSI), DEM (discrete equation method), multi-resolution, Two-phase compressible flows.

1. Introduction

This work is devoted to the numerical resolution of a stochastic two-phase flow, using an adaptive semi-intrusive scheme. The context of this work is in the interface problems characterized by the coexistence of two separated phases. In some particular conditions, heat and mass transfer between the two phases can appear, increasing the complexity of observed phenomenon. The two-phase flow problems have been addressed by many authors [1–8], because of their use in a large number of engineering devices. The prediction of this flow is particularly important for some specific physical problems, such as cavitation phenomena, wall corrosion, efficiency deterioration and so on.

Several studies have been focused on formulations yielding a good trade-off between physical accuracy and mathematical/numerical difficulties.

In this study, we deal with a class of methods based on compressible approach, treating the interface like a diffused zone (*i.e.* an artificial transition region where the thermodynamic conditions are unknown). This class of methods is principally affected by two important numerical issues: (i) how to define the closure laws for the average interfacial velocity and pressure and (ii) the approximation of the non-conservative terms, involving the volume fraction gradient, for shock interaction with volume fraction discontinuities.

In this field, Baer and Nunziato proposed a model [9] that was unconditionally hyperbolic and able to deal with a wide range of application. Many variants have been proposed [2, 10–12] and thanks to its ability to solve the interface problems, the model was extended to other interesting application as the evaporation fronts [13].

An original variant to Baer and Nunziato model has been proposed by Abgrall and Saurel [14]. Instead of following the most classical way, *i.e.* discretization of an averaged model, the authors developed a numerical scheme using the so-called *discrete equation method* (DEM): starting with a semi-discrete scheme for the compressible Navier-Stokes equations for each phase, a statistical average is performed in order to obtain an approximation of the mean quantities.

Anyway, the numerical complexity and implementation issues motivate the formulation of new simplified approaches. Kapila *et al.* [8] proposed a five-equations model supposing the pressure and velocity equilibrium between the phases. This model is unconditionally hyperbolic. This type of model has been used by many authors (see [1, 5, 15]) and extended to the numerical approximation of two-phase flow problem with viscous effects [16, 17].

Prediction and accuracy of these models are anyway strongly affected by the presence of numerous uncertainties. First, the models can be affected by some uncertainties, as, for example, the initial gas volume fraction (it is not possible to measure, with a good accuracy, the initial fraction during the experience) or some tuning coefficients that, for simplicity, are taken constant in the simulation (as the drag force coefficient in the drift-flux model or the heat exchange coefficient). Secondly, some uncertainties can be directly driven by the physics, geometric tolerances or experimental measurements.

Taking into account these uncertainties in the numerical simulation, is of fundamental importance for an accurate estimation of the simulation with respect to the experimental data. Anyway, this analysis is complicated by the crossing between the stochastic region (linked to the uncertainties) and shock-dominated multiphase flow.

Concerning uncertainty quantification methods, we can distinguish between non-intrusive approaches, *i.e.* where uncertainties are quantified practically by making multiple calls to a deterministic code (see the Monte Carlo family of techniques [18], the collocation family [19] and the non-intrusive Galerkin projection methods), and intrusive approaches, *i.e.* where the original deterministic code is completely modified in order to consider in the model the uncertainties and to quantify them. Concerning shock-dominated flows, the problem is to find an efficient representation of the stochastic solution, when the flow presents some discontinuities, thus producing a shock evolving in the coupled physical/stochastic space. Probabilistic uncertainty quantification (UQ) approaches represent the inputs as random variables and seek to construct a statistical characterization of few quantities of interest.

Wan and Karniadakis have introduced an adaptive class of methods for solving the discontinuity issues by using local basis functions, the multi-element generalized Polynomial Chaos (ME-gPC), see [20]. This strategy deals with an adaptive decomposition of the domain on which local basis are employed. In order to treat discontinuous response surfaces, [21, 22] applied a multiresolution analysis to Galerkin projection schemes. The intrusive Galerkin approach may lead to optimal representation of the solution, exhibiting an exponential convergence, if a proper basis is chosen. However the intrusive Galerkin approach results in a larger system of equations than in deterministic case with, in addition, a different structure that requires a new class of solver and numerical code. Despite this issue, the intrusive Galerkin approach can be demonstrated to have substantial advantages with respect non-intrusive approach, not only for idealized systems, but also for large-scale applications [21]. Advancements have been achieved in the Galerkin intrusive scheme where the wavelets formulation has been introduced in order to modify the basis of approximation [23]. It modifies the basis, by enriching the space with a hierarchical structure according to the regularity of the solution. However the Galerkin approach presented in [23] remains very problem-dependent. In fact, using a Roe-type solver requires to know the eigenstructure of the Roe matrix explicitly; this can be very complex. More over, *ad hoc* entropy fix should be adopted, thus increasing the numerical cost associated to the representation of discontinuous solution [24]. This original approach has been further improved to obtain a more efficient scheme employing a multiresolution adaptive strategy [23]. However actually this approach is limited by the spatial and time discretization accuracy that could dominate the overall accuracy of the global scheme. In [25], an intrusive formulation of the stochastic Euler equations based on Roe variables is presented. It is shown that the Roe variable formulation is robust for supersonic problems where the conservative variable formulation fails, but only for localized basis functions of the generalized chaos representation. For global Legendre polynomials, the discontinuities in stochastic space lead to oscillations and unphysical behavior of the solution and numerical instability. Wavelet functions are more robust in this respect, and do not yield oscillations around discontinuities in stochastic space, but need very regular grids.

More recently, in the context of uncertainty quantification studies, Abgrall and Congedo [26] proposed a novel semi-intrusive approach that extend in a straightforward and natural way, the representation of the variables in the physical space also along the stochastic space. This approach leads to a very flexible scheme able to handle whatever form of probability density function even time varying and discontinuous. One of the prominent advantage of this kind of approach is the possibility to extend in an easier way an existing deterministic code to its stochastic counterparts.

Recently, a cell-average setting multiresolution framework has been coupled with the SI scheme. Some reference test-cases are performed to demonstrate the convergence properties and the efficiency of the overall scheme: the linear advection problem for both smooth and discontinuous initial conditions, the inviscid Burgers equation and the 1D Euler system of equations to model an uncertain shock tube problem obtained by the well-known Sod shock problem [27].

Actually, for the stochastic investigation in a two-phase flow, the non-intrusive approach has been, clearly, favored, but the number of contributions is actually low [28–32]. However, the non-intrusive method result in a expensive computational cost, compared to intrusive method. To our knowledge, in literature there is only one contribution about an intrusive method applied to a two-phase flow investigation proposed by Petterson et al. [33]. They proposed a five equations model (one pressure and one velocity) coupled to a perfect gas equation of state for both the phases. Then, in order to obtain the stochastic formulation of the two-phase problem, they modified the fluxes, including the stochastic variable.

In this study, a new scheme for the numerical approximation of a five-equation model based on the DEM method using an adaptive semi-intrusive scheme for the uncertainty quantification is presented.

In particular, the MR framework with real-time adaptivity in the stochastic space, is adapted and coupled with the DEM scheme for the discretization of one dimensional two-phase five-equations model [14].

This paper is organized as follows. In section 2, at first, a description of the five equation model and of the semi-discrete equation obtained with the DEM method is explained. Then, in Section 3, main elements of the adaptive-semi-intrusive scheme (aSI) are presented. In particular, the key element of this new scheme, *i.e.* the expectancy flux computation with the link between the DEM formulation and the aSI scheme, is highlighted. Thermodynamic closure is addressed in section 4. In section 5, three test-cases are considered for the assessment of the proposed formulation.

2. Mathematical model

In this section, we illustrate the coupling of the two-phase flows resolution scheme with the adaptive multiresolution semi-intrusive scheme.

The two-phase model is based on a five-equation model with a single pressure and a single velocity. It is obtained imposing the asymptotic reduction of a seven equation model and it is discretized with a DEM approach, following Abgrall [5].

In this section, we recall briefly the governing equations and the principles of the DEM approach, since it has already been extensively explained in [5, 14, 34].

2.1. The five equations model

The well-known Baer & Nunziato [9] model is composed by the conservative equations of each phase and one transport equation for each volume fraction of phases (in this case no heat and mass transfer is considered):

$$
\begin{cases}
\dfrac{\partial \alpha_1}{\partial t} & = - \mathbf{u}_I \cdot \nabla \alpha_1 & + \mu(p_1 - p_2) \\
\dfrac{\partial \alpha_1 \rho_1}{\partial t} + div(\alpha_1 \rho_1 \mathbf{u}_1) & = \quad 0 & \\
\dfrac{\partial \alpha_1 \rho_1 \mathbf{u}_1}{\partial t} + div(\alpha_1 \rho_1 \mathbf{u}_1 \otimes \mathbf{u}_1) + \nabla(\alpha_1 p_1) & = \quad p_I \nabla(\alpha_1) & + \lambda(\mathbf{u}_2 - \mathbf{u}_1) \\
\dfrac{\partial \alpha_1 \rho_1 E_1}{\partial t} + div(\alpha_1(\rho_1 E_1 + p_1)\mathbf{u}_1) & = \quad p_I \mathbf{u}_I \cdot \nabla(\alpha_1) & + \lambda \mathbf{u}_I \cdot (\mathbf{u}_2 - \mathbf{u}_1) + \\
 & & - \mu p_I (p_1 - p_2) \\
\dfrac{\partial \alpha_2}{\partial t} + \mathbf{u}_I \cdot \nabla \alpha_2 & = & - \mu(p_1 - p_2) \\
\dfrac{\partial \alpha_2 \rho_2}{\partial t} + div(\alpha_2 \rho_2 \mathbf{u}_2) & = \quad 0 & \\
\dfrac{\partial \alpha_2 \rho_2 \mathbf{u}_2}{\partial t} + div(\alpha_2 \rho_2 \mathbf{u}_2 \otimes \mathbf{u}_2) + \nabla(\alpha_2 p_2) & = \quad p_I \nabla(\alpha_2) & - \lambda(\mathbf{u}_2 - \mathbf{u}_1) \\
\dfrac{\partial \alpha_2 \rho_2 E_2}{\partial t} + div(\alpha_2(\rho_2 E_2 + p_2)\mathbf{u}_2) & = \quad p_I \mathbf{u}_I \cdot \nabla(\alpha_2) & - \lambda \mathbf{u}_I \cdot (\mathbf{u}_2 - \mathbf{u}_1) + \\
 & \quad \underbrace{\hphantom{p_I \mathbf{u}_I \cdot \nabla(\alpha_2)}}_{Non\ conservative\ terms} & \underbrace{+ \mu p_I (p_1 - p_2)}_{Relaxation\ terms}
\end{cases}
\tag{1}
$$

where the subscripts 1 and 2 refer to the two phases k. Quantities α_k, ρ_k, $\mathbf{u}_k$, p_k, E_k are the volume fraction, the density, the velocity vector, the pressure and the total energy, respectively for each phase k. The last one is defined as $E_k = e_k + 0.5 u_k^2$. The interface velocity and the pressure are indicated with u_I and p_I, respectively. These ones are defined in [9] as $u_I = u_2$ and $p_I = p_1$, with 1 and 2 corresponding to the gas and the liquid phases, respectively. Other possible definitions of interface variables are given in [4, 14].

Parameters λ and μ represent the dynamic compaction viscosity and the relaxation velocity parameter, respectively.

The system (1) can be expressed in vectorial form as follows:

$$
\frac{\partial U}{\partial t} + \frac{\partial}{\partial x} F(U) + B(U) \frac{\partial \alpha_1}{\partial x} = S(U) \tag{2}
$$

or, after some manipulation:

$$
\frac{\partial U}{\partial t} + FT(U) = S(U) \tag{3}
$$

where

$$
U = \begin{pmatrix} \alpha_1 \\ \alpha_1 \rho_1 \\ \alpha_1 \rho_1 \mathbf{u}_1 \\ \alpha_1 \rho_1 E_1 \\ \alpha_2 \\ \alpha_2 \rho_2 \\ \alpha_2 \rho_2 \mathbf{u}_2 \\ \alpha_2 \rho_2 E_2 \end{pmatrix}, \qquad FT(U) = \frac{\partial}{\partial x} F(U) + B(U) \frac{\partial \alpha_1}{\partial x},
$$

$$F(U) = \begin{pmatrix} 0 \\ \alpha_1 \rho_1 \mathbf{u}_1 \\ \alpha_1(\rho_1 \mathbf{u}_1 \otimes \mathbf{u}_1) + p_1 \\ \alpha_1(\rho_1 E_1 + p_1)\mathbf{u}_1 \\ 0 \\ \alpha_2 \rho_2 \mathbf{u}_2 \\ \alpha_2(\rho_2 \mathbf{u}_2 \otimes \mathbf{u}_2) + p_2 \\ \alpha_2(\rho_2 E_2 + p_2)\mathbf{u}_2) \end{pmatrix}$$

$$B(U) = \begin{pmatrix} \mathbf{u}_I \\ 0 \\ -p_I \\ -p_I \mathbf{u}_I \\ \mathbf{u}_I \\ 0 \\ -p_I \\ -p_I \mathbf{u}_I \end{pmatrix}, \quad S(U) = \begin{pmatrix} \mu(p_1 - p_2) \\ 0 \\ \lambda(\mathbf{u}_2 - \mathbf{u}_1) \\ \lambda \mathbf{u}_I \cdot (\mathbf{u}_2 - \mathbf{u}_1) - \mu p_I (p_1 - p_2) \\ -\mu(p_1 - p_2) \\ 0 \\ -\lambda(\mathbf{u}_2 - \mathbf{u}_1) \\ -\lambda \mathbf{u}_I \cdot (\mathbf{u}_2 - \mathbf{u}_1) + \mu p_I (p_1 - p_2) \end{pmatrix}.$$

Supposing the mechanical equilibrium, the equality of pressure and velocity can be obtained in the limit of a stiff mechanical relaxation as in [1, 8], *i.e.* the relaxation parameters, λ and μ are taken as infinite:

$$\mu = \frac{1}{\epsilon}, \qquad \lambda = \frac{1}{\epsilon}, \quad where \quad \epsilon \to 0^+. \tag{4}$$

As a consequence, the asymptotic development allows to find the solution such that the relaxation terms go to zero (for more details concerning asymptotic development, Refs. [1, 5, 17] are strongly recommended). Then, after some algebraic manipulations of system (1), the reduced model is thus obtained:

$$\begin{cases} \dfrac{\partial \alpha_1}{\partial t} + \mathbf{u} \cdot \nabla \alpha_1 = \dfrac{\rho_2 c_2^2 - \rho_1 c_1^2}{\frac{\rho_1 c_1^2}{\alpha - 1} + \frac{\rho_2 c_2^2}{\alpha_2}} \text{ div } \mathbf{u} \\ \dfrac{\partial \alpha_1 \rho_1}{\partial t} + \nabla(\alpha_1 \rho_1 \mathbf{u}) = 0 \\ \dfrac{\partial \alpha_2 \rho_2}{\partial t} + \nabla(\alpha_2 \rho_2 \mathbf{u}) = 0 \\ \dfrac{\partial \rho \mathbf{u}}{\partial t} + \nabla(\rho_k \mathbf{u} \otimes \mathbf{u} + p) = 0 \\ \dfrac{\partial E}{\partial t} + \nabla((E + p)\mathbf{u}) = 0 \end{cases} \tag{5}$$

where $\rho = \alpha_1 \rho_1 + \alpha_2 \rho_2$, $E = \alpha_1 \rho_1 e_1 + \alpha_2 \rho_2 e_2$, p and $\mathbf{u}$ are the mixture density, mixture total energy, the mixture pressure and the mixture velocity, respectively. Finally, c_k is the sound of speed of each phase.

We remember that $\alpha_1 + \alpha_2 = 1$, so only a single phase is considered in the unknowns of the system that, for the system 5 are: α_1, ρ_1, ρ_2, e_1, e_2, p and u. There are seven unknowns. Then, in order to close the system (5), an equation of state (EOS) for each pure phase is demanded in order to define all the thermodynamic properties. This model involves mechanical equilibrium between the phases at any time, as it is evident looking at the presence of only one pressure p and only one velocity vector, $\mathbf{u}$, in the system 5. Finally, the computations presented in this work rely on the five-equation model.

2.2. The numerical scheme

The DEM approach has been derived in [14] and in [5] for the five-equations model. We recall here the main lines of the scheme.

First, we remember that the DEM consists in applying at a discrete level, the same procedure used to obtain a compressible multiphase model, *i.e.*:

1. Suppose that each pure fluid is governed by the Euler equations.
2. Introduce, for each phase, the characteristic function X_k that satisfies the topological equation:

$$\frac{\partial X_k}{\partial t} + \sigma \cdot \nabla X_k = 0, \;\; with \;\; X_k = \begin{cases} 1 & \text{if } (\vec{x}\text{,t}) \text{ belongs to phase k} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

   where σ is the interface velocity between the two phases.
3. An averaging procedure, $\mathcal{E}(\cdot)$, as in Drew and Passmann [35], is applied to the Euler equations (see [14]).
4. A statistical average is performed in order to obtain an approximation of the mean quantities.

Obtaining the semi-discrete numerical approximation of the two-phase system (5) demands a two-steps procedure. First, the DEM method, previously described, is applied to a seven equations model, *i.e.* to the system (1). After obtaining its semi-discrete numerical approximation, a relaxation procedure is applied, always at a discrete level, in order to reach a mechanical equilibrium.

Now, let us suppose that at time t, the computational domain Ω is divided into the cells $\mathcal{C}_i =]x_{i-1/2}, x_{i+1/2}[$. At a time $t = t + s$ (with s small), we assume that the interface in $x_{i+1/2}$ moves at a velocity $\sigma_{i+1/2}$ and the interface in $x_{i-1/2}$ moves at a velocity $\sigma_{i-1/2}$. As a consequence, the cell $\mathcal{C}_i$ evolves in $\bar{\mathcal{C}}_i =]x_{i-1/2} + s\sigma_{i-1/2}, x_{i+1/2} + s\sigma_{i+1/2}[$ (see figure 1). The cell may be either smaller or larger than the original ones $\mathcal{C}_i$, depending on the signs of the velocities. Then, we denote with $F(U_L, U_R)$ the Godunov numerical flux between the states U_L and U_R, and with $F^{lag}(U_L, U_R)$ the flux across the contact discontinuity between the states U_L and U_R (see figure 1). The relation between the two fluxes is equal to :

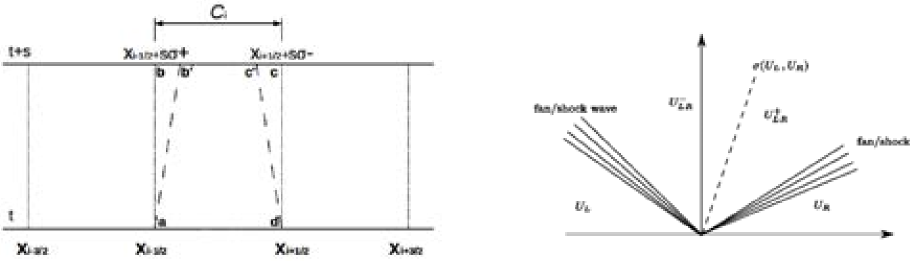


Figure 1: a) Subdivision of computational domain. b) The various states in the Riemann problem between states U_L and U_R.

$$F^{lag}(U_L, U_R) = F(U^+_{LR}) - \sigma(U_L, U_R)U^+_{LR} = F(U^-_{LR}) - \sigma(U_L, U_R)U^-_{LR}, \tag{7}$$

where the superscripts ± denote the state on the right and on the left of the contact discontinuity as in figure 1.

The semi-discrete scheme for the reduced five equations model in 1D is:

$$\begin{cases} \dfrac{\partial \alpha_1}{\partial t} = FT(U_1) + \dfrac{\alpha_1\alpha_2}{\alpha_2\rho_1 a_1^2 + \alpha_1\rho_2 a_2^2} \left\{ \dfrac{FT(U_8)}{\alpha_2\rho_2\chi_2} - \dfrac{u_2 FT(U_7)}{\alpha_2\rho_2\chi_2} + \dfrac{\frac{u_2}{2}^2 - e_2 - \rho_2\kappa_2}{\alpha_2\rho_2\chi_2} FT(U_6) + \right. \\ \qquad\qquad \left. + \dfrac{\rho_2^2\kappa_2 FT(U_1)}{\alpha_2\rho_2\chi_2} - \dfrac{FT(U_4)}{\alpha_1\rho_1\chi_1} + \dfrac{u_1 FT(U_3)}{\alpha_1\rho_1\chi_1} - \dfrac{\frac{u_1}{2}^2 - e_1 - \rho_1\kappa_1}{\alpha_1\rho_1\chi_1} FT(U_2) - \dfrac{\rho_1^2\kappa_1 FT(U_5)}{\alpha_1\rho_1\chi_1} \right\} \\ \dfrac{\partial \alpha_1\rho_1}{\partial t} = FT(U_2) \\ \dfrac{\partial \alpha_2\rho_2}{\partial t} = FT(U_6) \\ \dfrac{\partial \rho u}{\partial t} = FT(U_3) + FT(U_7) \\ \dfrac{\partial \rho E}{\partial t} = FT(U_4) + FT(U_8) \end{cases} \tag{8}$$

where χ_k and κ_k, are defined as follows:

$$\chi_k = \left(\frac{\partial e_k}{\partial P_k}\right)_{\rho_k} \quad ; \quad \kappa_k = \left(\frac{\partial e_k}{\partial \rho_k}\right)_{P_k} \tag{9}$$

where e_k is the phase internal energy.

As explained before, the vector $FT(U_j)$, with $j = 1, ..., 8$, is the sum of two contributions, *i.e.* the flux of hyperbolic system (conservative term) and the non-conservative terms, obtained for each equation of the system (1).

The correspondence of the semi-discrete system (8) with the model (5) has been demonstrated in [5]. Note that this method features initially two different thermodynamic states of phases, attaining, finally, a mechanical equilibrium. On the contrary, a direct discretization of the system (5) means directly the equality of initial pressure and velocity of the phases.

Following the adaptive multiresolution semi-intrusive scheme step is, now, to define the vector $FT(U_j)$ that for each component is composed by a conservative and a non-conservative terms:

$$\begin{aligned} FT(U_j) = {} & \frac{1}{\triangle x}\,\mathcal{E}\left(X(x_{i+1/2},t)F(U^*_{i+1/2}) - X(x_{i-1/2},t)F(U^*_{i-1/2})\right) + \\ & + \frac{1}{\triangle x}\left(\mathcal{E}([X]_{j=0})F^{lag}(U_i^-,U_{i-1}^+) - \mathcal{E}([X]_{j=N})F^{lag}(U_i^+,U_{i+1}^-)\right), \end{aligned} \tag{10}$$

where $U^*_{i+1/2}$ (or $U^*_{i+1/2}$) denotes the solution of Riemann problem between U_i^+ and U_{i+1}^- (respectively,U_{i-1}^+ and U_i^-). Quantities $[X]_{j=0}$ and $[X]_{j=N}$ are the jump of X at the beginning and at the end of computational cell, respectively.

Following the procedure demonstrated in [5, 14], the idea of DEM method is to avoid the introduction of approximated estimation of fluxes expectancy. This is estimated basing on the probability to find in two neighbor cells the same phase or two different phases (see the "flow patterns" in the table 1). As a consequence, we can define the flux indicator as in the following:

$$\beta^{(l,r)}_{i+1/2} = sign(\sigma(U_i^l, U_{i+1}^r)) = \begin{cases} 1 & if\ \sigma(U_i^l, U_{i+1}^r) \geq 0, \\ -1 & if\ \sigma(U_i^l, U_{i+1}^r) < 0, \end{cases}$$

where l and r indicate the phase at the left and the right of interface, respectively. Then, conservative and non-conservative terms of (10) can be developed supposing the four instances. Again, for sake of clarity, we briefly recall the main ideas of this strategy [5, 14].

Flow Patterns	Jump indicator	Flux Indicator
$\Sigma_1 - \Sigma_2$	$[X]_{1,1} = 0$	$\left(\beta^{(1,2)}_{i+1/2}\right)$
$\Sigma_1 - \Sigma_1$	$[X]_{1,2} = \begin{cases} -1\ if\ \sigma(1,2) > 0 \\ 0 \quad otherwise \end{cases}$	1
$\Sigma_2 - \Sigma_1$	$[X]_{2,1} = \begin{cases} 1\ if\ \sigma(2,1) > 0 \\ 0 \quad otherwise \end{cases}$	$\left(\beta^{(2,1)}_{i+1/2}\right)$
$\Sigma_2 - \Sigma_2$	$[X]_{2,2} = 0$	0

Table 1: The various flow configurations at cell boundary $i + 1/2$.

The terms of the vector $FT(U_j)$ (see (10)) can be defined as:

$$\begin{aligned} \mathcal{E}\left(X(x_{i+\frac{1}{2}},t)F(U^*_{i+\frac{1}{2}})\right) = {} & \mathcal{P}_{i+\frac{1}{2}}(\Sigma_1 - \Sigma_1)F(U_i^{(1)}, U_{i+1}^{(1)}) + \\ & + \mathcal{P}_{i+\frac{1}{2}}(\Sigma_1 - \Sigma_2)\left(\beta^{(1,2)}_{i+\frac{1}{2}}\right)F(U_i^{(1)}, U_{i+1}^{(2)}) + \mathcal{P}_{i+\frac{1}{2}}(\Sigma_2 - \Sigma_1)\left(\beta^{(2,1)}_{i+\frac{1}{2}}\right)F(U_i^{(2)}, U_{i+1}^{(1)}) \end{aligned}$$

$$\mathcal{E}\left(X(x_{i-\frac{1}{2}},t)F(U^*_{i-\frac{1}{2}})\right) = \mathcal{P}_{i-\frac{1}{2}}(\Sigma_1-\Sigma_1)F(U^{(1)}_{i-1},U^{(1)}_i)+ \\ +\mathcal{P}_{i-\frac{1}{2}}(\Sigma_1-\Sigma_2)\left(\beta^{(1,2)}_{i-\frac{1}{2}}\right)F(U^{(1)}_{i-1},U^{(2)}_i)+\mathcal{P}_{i-\frac{1}{2}}(\Sigma_2-\Sigma_1)\left(\beta^{(2,1)}_{i-\frac{1}{2}}\right)F(U^{(2)}_{i-1},U^{(1)}_i)$$

$$\mathcal{E}\left([X]_N F^{lag}(U^{N(w)}_i,U^-_{i+1})\right) = \mathcal{P}_{1+1/2}(\Sigma_1,\Sigma_2)\left(\beta^{(1,2)}_{i+1/2}\right)F^{lag}(U^{(1)}_i,U^{(2)}_{i+1})+ \\ -\mathcal{P}_{1+1/2}(\Sigma_2,\Sigma_1)\left(\beta^{(2,1)}_{i+1/2}\right)F^{lag}(U^{(2)}_i,U^{(1)}_{i+1})$$

$$\mathcal{E}\left([X]_0 F^{lag}(U^+_{i-1},U^0_i)\right) = -\mathcal{P}_{1-1/2}(\Sigma_1,\Sigma_2)\left(\beta^{(1,2)}_{i-1/2}\right)F^{lag}(U^{(1)}_{i-1},U^{(2)}_i)+ \\ +\mathcal{P}_{1-1/2}(\Sigma_2,\Sigma_1)\left(\beta^{(2,1)}_{i+1/2}\right)F^{lag}(U^{(2)}_{i-1},U^{(1)}_i)$$

It remains to evaluate the term of probability, $\mathcal{P}_{i\pm1/2}(\Sigma_p,\Sigma_q)$ (see [14]). For simplicity, we show the final formulation for $i+1/2$:

$$\begin{aligned}
\mathcal{P}_{i+1/2}(\Sigma_1,\Sigma_1) &= \min\left(\alpha^{(1)}_i,\ \alpha^{(1)}_{i+1}\right)\\
\mathcal{P}_{i+1/2}(\Sigma_1,\Sigma_2) &= \max\left(\alpha^{(1)}_i-\alpha^{(1)}_{i+1},0\right)\\
\mathcal{P}_{i+1/2}(\Sigma_2,\Sigma_1) &= \max\left(\alpha^{(2)}_i-\alpha^{(2)}_{i+1},0\right)\\
\mathcal{P}_{i+1/2}(\Sigma_1,\Sigma_2) &= \min\left(\alpha^{(2)}_i,\ \alpha^{(2)}_{i+1}\right).
\end{aligned}$$

where Σ_k indicates the phase, with $k=1,2$.

The system (8) can be written in vectorial form as follows:

$$\frac{W^{n+1}_i-W^n_i}{\Delta t}+\frac{\Delta F(W)_i}{\Delta x}=0 \tag{12}$$

where

$$W=\begin{pmatrix}\alpha_1\\ \alpha_1\rho_1\\ \alpha_2\rho_2\\ \rho\mathbf{u}\\ \rho E\end{pmatrix}$$ is the conservative variables vector of the reduced five equations model and

$$\Delta F(W)=\Delta x\begin{pmatrix}
FT(U_1)+\dfrac{\alpha_1\alpha_2}{\alpha_2\rho_1a_1^2+\alpha_1\rho_2a_2^2}\left\{\dfrac{FT(U_8)}{\alpha_2\rho_2\chi_2}-\dfrac{u_2FT(U_7)}{\alpha_2\rho_2\chi_2}+\dfrac{\frac{{u_2}^2}{2}-\varepsilon_2-\rho_2\kappa_2}{\alpha_2\rho_2\chi_2}FT(U_6)+\dfrac{\rho_2^2\kappa_2FT(U_1)}{\alpha_2\rho_2\chi_2}+\right.\\
\left.-\dfrac{FT(U_4)}{\alpha_1\rho_1\chi_1}+\dfrac{u_1FT(U_3)}{\alpha_1\rho_1\chi_1}-\dfrac{\frac{{u_1}^2}{2}-\varepsilon_1-\rho_1\kappa_1}{\alpha_1\rho_1\chi_1}FT(U_2)-\dfrac{\rho_1^2\kappa_1FT(U_5)}{\alpha_1\rho_1\chi_1}\right\}\\
FT(U_2)\\
FT(U_6)\\
FT(U_3)+FT(U_7)\\
FT(U_4)+FT(U_8)
\end{pmatrix}.$$

The numerical flux F(U) is obtained thanks to an approximate Riemann solver. It defines the contact speed $\sigma(U_L,U_R)$, allowing to define the Lagrangian flux F^{lag} (see equation (7)). The Riemann problems solution is sought for times that satisfy a CFL conditions of the type:

$$|\lambda_{max}|\,\frac{\Delta x}{\Delta t}\leqslant\frac{1}{2}.$$

In this paper, we have used the relaxation solver [36] for all computations (see [5] for more details).

2.2.1. Extension to second order

Now, we extend the approximation of the scheme (12) to a second order following an extension of a MUSCL approach. This approach for a multiphase flow had been proposed in [14] and in this study we apply exactly the same extension. Anyway we recall here the main lines.

The following scheme is an extension of a predictor-corrector scheme for a general conservation law $\partial U/\partial t + \partial F/\partial x{=}0$ (see [37]). We assume an uniform mesh Δx and we define four steps :

Step 1: From U_j^n, compute the limited slope δU and evaluate:

$$U_{i-\frac{1}{2}}^n = U_i^n - \frac{\Delta x}{2}\delta U_i^n \quad \text{and} \quad U_{i+\frac{1}{2}}^n = U_i^n + \frac{\Delta x}{2}\delta U_i^n$$

.

Step 2: Evaluate the solution over half a time step:

$$U_i^{n+\frac{1}{2}} = U_i^n - \frac{\Delta t}{2\Delta x}\left(F(U_{i+\frac{1}{2},l}^n, U_{i+\frac{1}{2},r}^n) - F(U_{i-\frac{1}{2},l}^n, U_{i-\frac{1}{2},r}^n)\right)$$

.

Step 3: From $U_j^{n+\frac{1}{2}}$, evaluate the limited slope $\delta U_j^{n+\frac{1}{2}}$ and compute:

$$U_{i-\frac{1}{2},r}^{n+\frac{1}{2}} = U_i^{n+\frac{1}{2}} - \frac{\Delta x}{2}\delta U_i^{n+\frac{1}{2}} \quad \text{and} \quad U_{i+\frac{1}{2},l}^{n+\frac{1}{2}} = U_i^n + \frac{\Delta x}{2}\delta U_i^{n+\frac{1}{2}}$$

.

Step 4: Compute the final solution :

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\left(F(U_{i+\frac{1}{2},l}^{n+\frac{1}{2}}, U_{i+\frac{1}{2},r}^{n+\frac{1}{2}}) - F(U_{i-\frac{1}{2},l}^{n+\frac{1}{2}}, U_{i-\frac{1}{2},r}^{n+\frac{1}{2}})\right)$$

.

Observe that the step 1 and 2 are identical to step 3 and 4, respectively. Let us focus now on the scheme adapted to a multiphase flow, in particular on the predictor step (steps 1 and 2).

The predictor-corrector scheme for the multiphase flows

The reconstruction of variables is done on the primitive variables V_i^n, where $V_k = (\alpha_k, \rho_k, u_k, P_k)^T$ for each phase k, because the volume fraction, α, should be between 0 and 1 and because the constraint $\rho_k \geqslant 0$ and $P_k \geqslant 0$. We extrapolate the primitive variables by using their limited slope $\delta_i V$ at most left (l) or right (r) points of the cell $]x_{i-1/2}, x_{i+1/2}[$:

$$V_{i-\frac{1}{2},r}^n = V_i^n - \frac{\Delta x}{2}\delta_i V \quad \text{and} \quad V_{i+\frac{1}{2},l}^n = V_i^n + \frac{\Delta x}{2}\delta_i V$$

.

As a consequence, denoting by $U_{i\pm1/2,r}^n$ (resp. $U_{i\pm1/2,l}^n$) the vector of conservative variables corresponding to $V_{i\pm1/2,r}^n$ (resp. $V_{i\pm1/2,l}^n$), we can write the final formulation of the predictor step, as follows:

$$\frac{W_i^{n+\frac{1}{2}} - W_i^n}{\Delta t} + \frac{\Delta F(W)_i}{\Delta x} = 0 \tag{13}$$

where the arguments are defined by the reconstructed left and right states at $x_{i\pm1/2}$. Since the components of the vector $\Delta F(W)$ should be defined at $x_{i\pm1/2}$, so the components of vector $FT(U_j)$ (see (10)) are defined as follows:

$$\begin{aligned}\mathcal{E}\left(XF\right)_{i-\frac{1}{2}} =& P_{1-\frac{1}{2}}(\Sigma_1, \Sigma_1) F(U_{i-\frac{1}{2},l}^{(1),n}, U_{i-\frac{1}{2},r}^{(1),n}) + \\ &+ P_{1-\frac{1}{2}}(\Sigma_1, \Sigma_2)\left(\beta_{i-\frac{1}{2}}^{(1,2)}\right) F(U_{i-\frac{1}{2},l}^{(1),n}, U_{i-\frac{1}{2},r}^{(2),n}) + \\ &+ P_{1-\frac{1}{2}}(\Sigma_2, \Sigma_1)\left(\beta_{i-\frac{1}{2}}^{(2,1)}\right) F(U_{i-\frac{1}{2},l}^{(2),n}, U_{i-\frac{1}{2},r}^{(1),n})\end{aligned} \tag{14}$$

$$\begin{aligned}\mathcal{E}\left(XF\right)_{i+\frac{1}{2}} =& P_{1+\frac{1}{2}}(\Sigma_1,\Sigma_1)F(U^{(1),n}_{i+\frac{1}{2},l},U^{(1),n}_{i+\frac{1}{2},r})+\\ &+P_{1+\frac{1}{2}}(\Sigma_1,\Sigma_2)\left(\beta^{(1,2)}_{i+\frac{1}{2}}\right)F(U^{(1),n}_{i+\frac{1}{2},l},U^{(2),n}_{i+\frac{1}{2},r})+\\ &+P_{1+\frac{1}{2}}(\Sigma_2,\Sigma_1)\left(\beta^{(2,1)}_{i+\frac{1}{2}}\right)F(U^{(2),n}_{i+\frac{1}{2},l},U^{(1),n}_{i+\frac{1}{2},r})\end{aligned} \tag{15}$$

$$\begin{aligned}&\Delta x\left(\mathcal{E}([X]_{j=0})F^{lag}(U^{-}_{i},U^{+}_{i-1})-\mathcal{E}([X]_{j=N})F^{lag}(U^{+}_{i},U^{-}_{i+1})\right)=\\ &\quad=P_{i+\frac{1}{2}}(\Sigma_1,\Sigma_2)\left(\beta^{(1,2),n}_{i+\frac{1}{2}}\right)F^{lag}(U^{(1),n}_{i+\frac{1}{2},l},U^{(2),n}_{i+\frac{1}{2},r})+\\ &\quad-P_{i+\frac{1}{2}}(\Sigma_2,\Sigma_1)\left(\beta^{(2,1)}_{i+\frac{1}{2}}\right)F^{lag}(U^{(2),n}_{i+\frac{1}{2},l},U^{(1),n}_{i+1/2,r})+\\ &\quad-P_{i-\frac{1}{2}}(\Sigma_1,\Sigma_2)\left(\beta^{(1,2)}_{i-\frac{1}{2}}\right)F^{lag}(U^{(1),n}_{i-\frac{1}{2},l},U^{(2),n}_{i-\frac{1}{2},r})+\\ &\quad+P_{i-\frac{1}{2}}(\Sigma_2,\Sigma_1)\left(\beta^{(2,1)}_{i-\frac{1}{2}}\right)F^{lag}(U^{(2)}_{i-\frac{1}{2},l},U^{(1)}_{i-\frac{1}{2},r})+\\ &\quad+\max\left(0,\Delta\alpha^1_i\right)F^{lag}(U^2_i,U^1_i)-\max\sum_{N-1}^{l=1}\left(0,\Delta\alpha^2_i\right)F^{lag}(U^1_i,U^2_i),\end{aligned} \tag{16}$$

where $\Delta\alpha^1_i = \alpha^1_{i+1/2,l} - \alpha^1_{i+1/2,r}$ and $\Delta\alpha^2_i = \alpha^2_{i+1/2,l} - \alpha^2_{i+1/2,r}$ are the limited slope of α^1 and α^2 in the cell C_i. The coefficient $\beta^{(1,2)}_{i\pm1/2}$ represents the sign on the contact speed evaluated at $x_{i\pm1/2}$. For more details, Refs. [14] is strongly recommended.

3. Adaptive-Semi-Intrusive formulation

In this section, some elements for the stochastic formulation of the DEM scheme, presented in the previous section, are reported. The adaptive-Semi-Intrusive (aSI) scheme is a novel intrusive numerical method to propagate uncertainties. In particular, the aSI scheme is based on the semi-intrusive approach [26], introducing the multiresolution (MR) basis. The aim of introducing the MR basis is twofold. First, for compressible CFD problems, the propagation of narrow discontinuity region is a common issue. The MR basis offers a natural compact representation of this kind of functions, as already demonstrated in the seminal papers of Harten [38–40]. Moreover, another interesting feature of the MR is the possibility to analyze locally the regularity of a function. This feature can be employed to drive a topological refinement of the mesh in a time-dependent way.

The MR framework, here employed, is ispired from the classical Harten framework [39, 40] where several building blocks are used as operators. First, the discretization operator $\mathcal{D}_k$ features the mapping between the continuous space of definition of the function in the analysis and discrete tessellation of resolution level k. The inverse operation is performed by a reconstruction operator $\mathcal{R}_k$. It is clear that the two operators should be consistent, *i.e.* $\mathcal{R}_k\mathcal{D}_k = I$. Moreover, operations between discrete levels are also demanded. The decimation operator D_k^{k-1} allows obtaining a coarser level from a finer one, while the prediction operator is designed to predict the value of a discrete element, for instance a cell-average, from a coarser resolution level. The main difference, with respect to the classical Harten framework, is the possibility to move directly (at each time step) from the coarser to the finest resolution level. This possibility well suits the scope of the UQ propagation analysis, as already demonstrated in [27, 41–43].

In the aSI scheme, the MR basis is injected in the SI method in order to represent function in the stochastic space. The aSI scheme allows to refine and derefine the overall physical-stochastic space with a good efficiency.

The key elements, featuring the coupling between the pure SI scheme and the MR framework, are the following. First, the SI scheme is based on representation of the quantities in terms of conditional expectancies. For the fluxes computations, a conservative reconstruction is then necessary. This step is fully demanded to the MR framework by the use of the reconstruction operator $\mathcal{R}_k$. Basically, the aSI scheme, driven by the MR analysis, locally identifies the region (of the combined physical-stochastic space), where the function is less regular. In these regions, a refinement of the stochastic space is performed and multiple calls of a time-update step of the SI are invoked.

Let us assume, for a generic resolution level k, a tessellation of the stochastic space as

$$\Xi^k = \bigcup_{j=1}^{N_\xi} \Xi_j^k, \quad \text{with} \quad \Xi_i^k \cap \Xi_j^k = 0 \quad \text{if} \quad i \neq j. \tag{17}$$

Let now suppose to consider the final step (13). If a conditional expectancy operator

$$\mathbb{E}(\bullet \,|\, \Xi_j) = \frac{1}{\mu(\Xi_j)} \int_{\Xi_j} \bullet(x, \xi, t)\, p(\xi, t)\, \mathrm{d}\xi \tag{18}$$

is applied on a generic cell Ξ_j^k of the tessellation, the final step (13), for the corrector, becomes

$$\mathbb{E}\left(W_i^{n+1} \,|\, \Xi_j^k\right) = \mathbb{E}\left(W_i^n \,|\, \Xi_j^k\right) + \frac{\Delta t}{\Delta x} \mathbb{E}\left(\Delta F(W)_i \,|\, \Xi_j^k\right). \tag{19}$$

The time-update, reported in Eq. (19), concerns the time-advancing strategy to increment the conditional expectancy of the solution, in a generic cell Ξ_j^k, by knowing its value at the previous time step and the expectancy of the fluxes at the interfaces. Let us imagine to formulate an initial value problem, *i.e.* a differential problems in which the initial condition is known. If a proper quadrature rule is chosen, in the combined physical-stochastic space, the value of the conditional expectancy of the initial solution can be obtained. The remaining step is to compute the computational expectancy of the fluxes. At this level, the interaction of the aSI formulation with the deterministic scheme is evident. In the particular case of DEM method (see the previous sections), solved by a predictor-corrector approach, it is possible to compute the value of the vector of conservative variables in a cell $\mathcal{C}_i$, knowing only the solutions at the cells $\{\mathcal{C}_{i-3}, \ldots, \mathcal{C}_{i+3}\}$. This derives from the need to compute a half time updated solution (for the predictor) in the cells $\{\mathcal{C}_{i-2}, \ldots, \mathcal{C}_{i+2}\}$, and then applying the corrector (on the updated values) on the cell $\{\mathcal{C}_{i-1}, \ldots, \mathcal{C}_{i+1}\}$. Remark that the computation of the slopes yields the enlargement of the stencil of one cell for side. The predictor step can be performed after that the local physical cell-average are computed. In principle, the scheme handles only conditional expectancy. By means of the reconstruction operator of the MR framework, the physical cell average values, for the stencil $\{\mathcal{C}_{i-3}, \ldots, \mathcal{C}_{i+3}\}$ are evaluated. The problem is equivalent to the deterministic one: the seven cell-average values are updated of half time step and the extrapolated values at the interfaces, of the cell of interest, can be computed. If this procedures is performed for all the Ng quadrature points of each physical interfaces, between spatial cells along the stochastic coordinate, the quadrature of the term ΔF can be easily obtained. The final step (19) can be finally applied. In the Algorithm 1, the set of operation, to compute the difference of the fluxes expectancies, is reported.

Algorithm 1: Computation of the fluxes expectancies in the aSI scheme for the DEM formulation with a predictor-corrector MUSCL approach.

```
for $\mathrm{ng} = 1, \ldots, \mathrm{Ng}$ do
    • Physical Vector assembling:
      Conservative reconstruction from MR reconstruction operator $\mathcal{R}_k$
      Conversion in primitive variables $V_i$
          $\mathbb{P}\mathbb{V}(\boldsymbol{\xi}_{\mathrm{ng}}) = \left\{V_{i-3}^n(\boldsymbol{\xi}_{\mathrm{ng}}), \ldots, V_{i+3}^n(\boldsymbol{\xi}_{\mathrm{ng}})\right\}$
    • Imposition of the boundary conditions
    • Slope computations (and limiting) $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-2}, \ldots, \mathcal{C}_{i+2}\}$:
          $\delta_\ell^n(\boldsymbol{\xi}_{\mathrm{ng}}) = \delta\left(V_{\ell-1}^n(\boldsymbol{\xi}_{\mathrm{ng}}), V_\ell^n(\boldsymbol{\xi}_{\mathrm{ng}}), V_{\ell+1}^n(\boldsymbol{\xi}_{\mathrm{ng}})\right)$
    • Extrapolation $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-3}, \ldots, \mathcal{C}_{i+3}\}$ (Step 1)
    • Semi-time step evolution $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-2}, \ldots, \mathcal{C}_{i+2}\}^a$ (Step 2):
    • Extrapolation $\forall \mathcal{C}_\ell \in \{\mathcal{C}_{i-1}, \ldots, \mathcal{C}_{i+1}\}$ (Step 3)
    • Delta flux computation: $\Delta F(W(\boldsymbol{\xi}_{\mathrm{ng})})_i$ (DEM solver)
end
Flux Quadrature:
    $\mathbb{E}\left(\Delta F(W)_i \,|\, \Xi_j^k\right) = \sum_{i=\mathrm{ng}}^{\mathrm{Ng}} w_{\mathrm{ng}} \Delta F(W(\boldsymbol{\xi}_{\mathrm{ng}}))_i$
```

Once the time-update step is formulated, this step can be considered as the result of the application of the discretization operator in MR. Performing, at each time step and for each physical coordinate, a MR driven refinement/derefinement (using the discretization operator $\mathcal{D}_k$ or relying only on prediction by P_{k-1}^k), the compact (with respect to the discrete dimensionality) representation of each conservative variable can be obtained. The final step of the UQ propagation process is the computation of statistics. They can be computed, even analytically, knowing the reconstruction operator in each cell along the stochastic coordinates. In this paper, the numerical test cases are carried out introducing a non-linear Essentially Non-Oscillatory (ENO) reconstruction based on cubic polynomials. The advantage of a such reconstruction technique, in the context of the MR approach proposed here, have been already shown in papers [27, 43].

4. Thermodynamic Closure

Defining thermodynamic properties is necessary in order to close the system describing compressible flows. Here, we consider as the equation of state (EOS), the Stiffened Gas (SG) to describe both liquid and gas phases:

$$\rho_k e_k = \frac{(P_k + \gamma P_{k,\infty})}{\gamma_k - 1} \tag{20}$$

where e_k is the phase internal energy, P_k is the phase pressure, γ_k and $P_{k,\infty}$ are two constants characterizing each fluid. The constants for these fluids are provided in table 2. The mixture SG-EOS can be easily obtained using the EOS of the single phases, by applying the definition of the total mixture energy equation:

$$\rho E = \alpha_1 \rho_1 e_1 + \alpha_2 \rho_2 e_2. \tag{21}$$

The internal energy of each phase, e_k, can be replaced by the Eq.(20), obtaining the mixture total energy as a function of the phase pressure. Under pressure equilibrium, we obtain the following expression for the

[a] The flux function can depends separately on the vector of the random parameters and from the unknown.

		Constant Fluid Conditions		
		γ	P_∞	$u_L = u_R$ $[m/s]$
All TC	Liq	4.4	6×10^9	0
	Gas	1.4	0	

n	Fluid	Test Case (TC) conditions						Uncertainty
		Left			Right			
		α	ρ $[\frac{kg}{m^3}]$	P [Pa]	α	ρ $[\frac{kg}{m^3}]$	P [Pa]	
TC1	G=Air	1-ϵ	0.3	1.0	1-ϵ	0.125	0.1	$\rho_G(\xi) = 0.3 + 1.6\xi$
	L=Water	ϵ	1000		ϵ	1000		
TC2	G=Air	0.5	50	1.0E+9	50	50	1.0E+5	$\alpha_G(\xi) = \alpha_G \pm 0.1\xi$
	L=Water	0.5	1000		0.5	1000		
TC3	G=Air	0.2	1	1.0E+9	0.8	1	1.0E+5	$P_{Left}(\xi) = P_{Left} \pm$ 0.5%
	L=Water	0.8	1000		ϵ	1000		

Table 2: Initial conditions for all test cases. $\epsilon = 10^{-8}$. For all test cases, in the right and left part, u_k=0.

pressure mixture:

$$P(\rho, e, \alpha_k) = \frac{\rho(E - \frac{\alpha_1\rho_1 q_1}{\rho} - \frac{\alpha_2\rho_2 q_2}{\rho}) - \left(\frac{\alpha_1\gamma_1 P_{\infty,1}}{\gamma_1 - 1} + \frac{\alpha_2\gamma_2 P_{\infty,2}}{\gamma_2 - 1}\right)}{\frac{\alpha_1}{\gamma_1 - 1} + \frac{\alpha_2}{\gamma_2 - 1}} \tag{22}$$

In this paper, the term q is supposed equal to zero for each phase.

5. Results

In this section, we show the results obtained for three test cases. Initial conditions and working fluids are specified for each test-case and summarized in table 2.

First, the implementation of the scheme is validated by running a stochastic test case well known in literature, for which the exact solution can be computed in the stochastic and physical spaces.

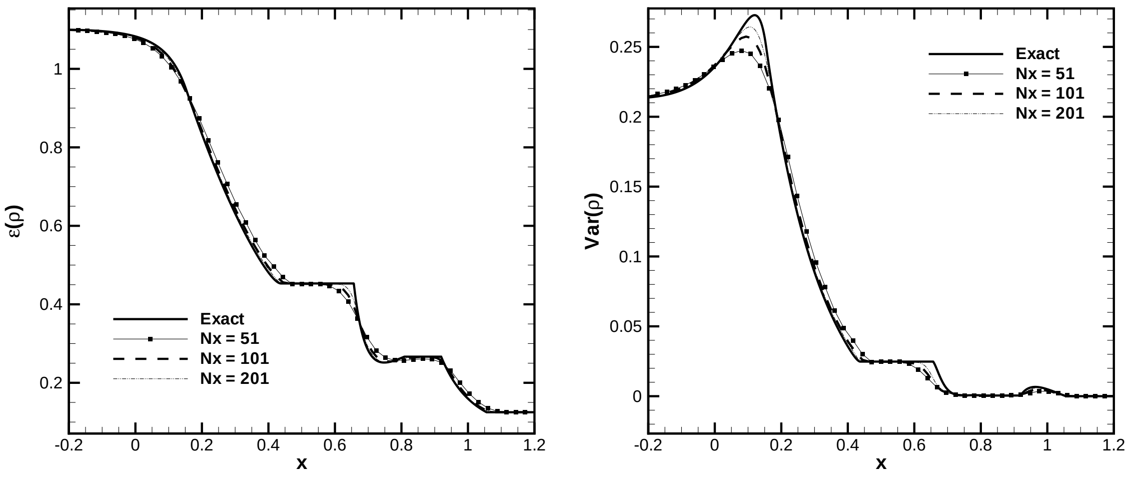
The other test-cases deals with a two-phase shock tube using a mixture of air and water as working fluid. Influence of uncertainty on the left gas volume fraction and on the left pressure, is investigated. Moreover stochastic and grid convergence are explored in different conditions.

5.1. TC1: validation of the scheme in a quasi-single phase fluid

The original test case [27] reproduces a single-phase (air) shock tube where the density on the left state is dependent on an uniformly distributed random parameter ξ. This test case is of interest since the exact solution in the stochastic space can be computed [27], thus, permitting to estimate scheme convergence. In particular, in this work, we consider a quasi single-phase shock tube, *i.e.* a mixture of air and water, where in each chamber of the tube a reduced liquid fraction is supposed (typically 10^{-8}).
This test-case has been modified in this sense for two different reasons:

- in order to verify that the coupling works well, *i.e.* that global stochastic/physical scheme is sufficiently robust to capture two-phase flow, eve, with a very reduced liquid fraction.
- Accuracy in the stochastic space can be assessed by making a comparison with respect to the exact solution (hypothesis that stochastic solution in a single-phase or quasi-single fluids are very similar).

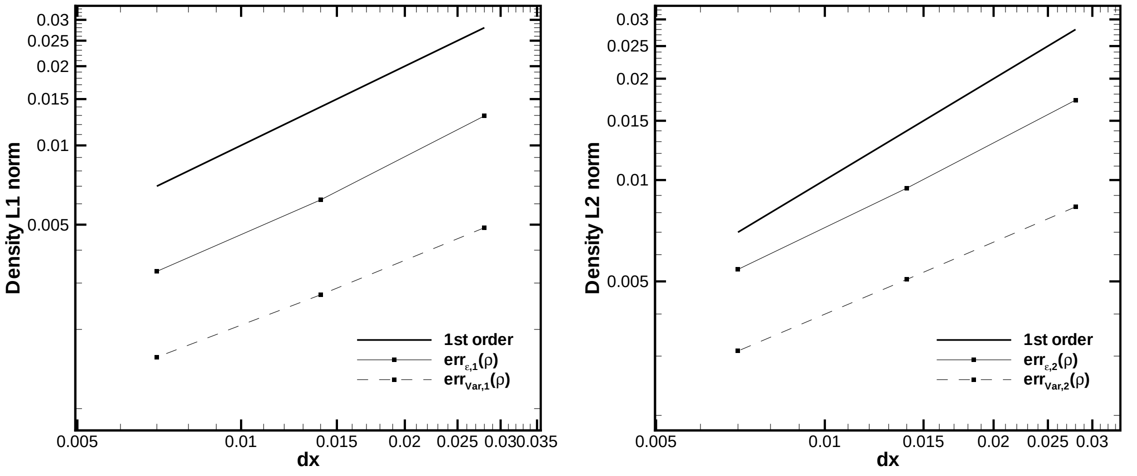
Initial conditions of this test case are specified in table 2. Left and right sides of the shock tube are filled out, principally, with air ($\alpha_{air} = 1 - 10^{-8}$) and with a very low percentage of water ($\alpha_{water} = 10^{-8}$). Density on the left state is dependent on an uniformly distributed random parameter $\xi \sim \mathcal{U}[0, 1]$: $\rho_L(\xi) = 0.3 + 1.6\xi$ kg/m^3. Values of the pressures are $p_L = 1$ and $p_R = 0.1$, while the right value of the density is $\rho_R = 0.125$.



(a)

(b)

Figure 2: Evolution of the density expectancy (a) and density variance (b) for the cell averaged solution of an uncertain shock tube problem at the final time $t = 0.31$ and for different physical meshes.



(a)

(b)

Figure 3: Spatial convergence for the stochastic shock tube problem with density uncertain initial condition [27]. L^1 (a) and L^2 (a) norms are shown for the statistic (mean and variance) of the solution.

Simulations are performed over a physical domain $\Omega = [-\frac{1}{5}, \frac{6}{5}]$ until a final time $t = 0.31$ with the position of the diaphragm equal to $x_d = 0.42$. The time space is divided in 6200 equal time steps of length $\Delta t = 5\times10^{-5}$.

Simulations are carried out over equally spaced meshes of 51, 101, 201 points employing the aSI scheme based on the MUSCL method (see 2.2.1 section) with a Superbee limiter.

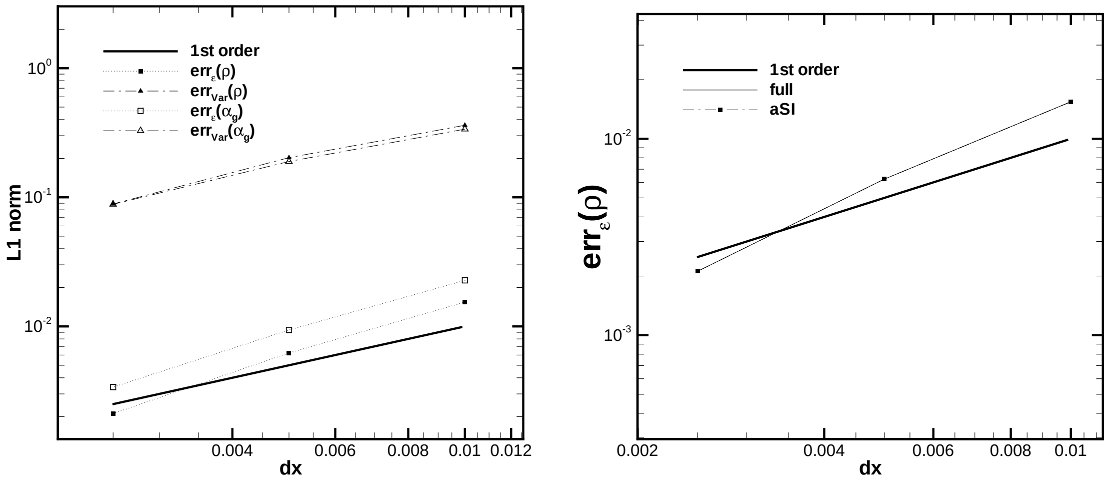
In figure 2, the evolution of the density expectancy and density variance are reported. Note that the exact solution is reported over a mesh of 2001 equally spaced points in the physical space. As it is evident, by increasing the number of points in the physical space, stochastic solution converge to the exact one, for both mean and variance.

In figure 3, the spatial convergence is reported for both the mean and the variance in L^1 and L^2 for the density ρ. The aSI method is obtained with a level of 128 ($m = 7$) stochastic cells with $\varepsilon = 10^{-4}$, while the reference solution is the exact solution obtained in [27].

5.2. TC2: two phase flow with uncertainty on gas volume fraction

In this case, the shock tube is filled out with water and air at the same volume fraction ($\alpha_k = 0.5$) on the right and on the left of diaphragm, located at x=0.5m. Initial conditions of this test case are described in the table 2. The deterministic solution has been validated in [17].
The gas volume fraction on the left state is dependent on an uniformly distributed random parameter $\xi \sim \mathcal{U}[0, 1]$: $\alpha_G(\xi) = \alpha_G \pm 0.1\xi$ and its propagation in the shock tube is observed. Simulations are performed over a physical domain $\Omega = [0, 1]$ until a final time $t = 193.744$ μs. The time space is divided in 1900 equal time steps of length $\Delta t = 1 \times 10^{-7}$. The simulations are carried out over equally spaced meshes of 101, 201, 401 and 801 points employing the aSI scheme based on the MUSCL method with a Van Leer limiter.

In figure 4(a), the spatial convergence is reported for both the mean and the variance in L^1 for the density ρ. It has been obtained with the aSI method with a level of 128 ($m = 7$) stochastic cells. Results obtained by the aSI method have been compared with the ones obtained by a full SI scheme, in terms of L1 norm (figure 4(b)) and of density mean and variance curves (figure 5), showing a perfect overlapping of the curves. For this reason, since we observed for all computations the same behavior of both the methods, hence, the figures and the observations will be presented only the aSI scheme results.

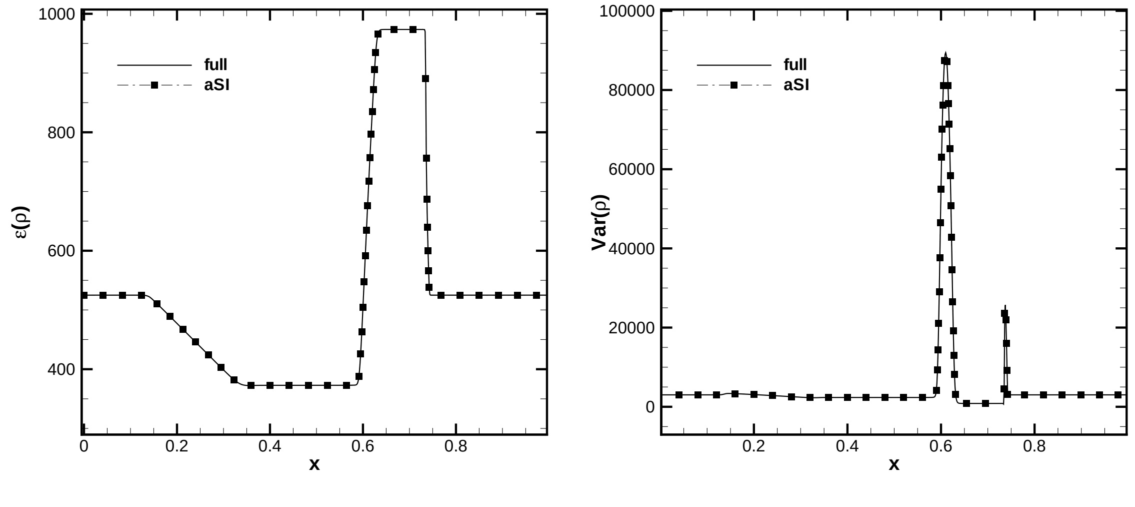


(a)

(b)

Figure 4: (a) Spatial convergence for the stochastic shock tube problem with an uncertain volume fraction as initial condition. L1 norms are shown for the density expectancy and variance of the solution.(b) Comparison between aSI and full SI scheme obtained with a level of 128 (m=7) is shown on the stochastic spatial convergence of the density expectancy.

The figures 6 and 7 show the deterministic spatial convergence in terms of mean and variance carried out over equally spaced meshes of 101, 201, 401 and 801 points employing the aSI scheme with a level of 512 ($m = 9$) stochastic cells.The most significant differences can be observed on the liquid and gas densities for both the statistics (mean and variance) of the solution (see figures 6(c)-6(d) and 7(c)-7(c)). The coarser mesh shows a behavior very different compared with the finest mesh for $0.6 < x < 0.75$, corresponding to the contact discontinuity and the shock wave.



(a)

(b)

Figure 5: Comparison between aSI and full SI scheme obtained with a level of 128 (m=7) on a deterministic space grid of 801 points, for (a) density expectancy and (b) density variance.

The scheme allows to evaluate clearly the propagation of uncertainty and with this test case we study the influence of inlet left gas volume fraction, α_g. Observing the variance profiles of thermodynamic variables, all of them present a pick in correspondence to the shock (see figure 7), but the phase density profiles and, of course, the gas volume fraction profile show a significant variation in correspondence to $x = 0.6$ m (see figure 7(a)-7(d)). These differences are more evident in figure 8, where the mean and standard deviation curves are compared on all the deterministic space. It is evident that the pressure is not influenced by the uncertainties. On the contrary, the phase densities profiles change especially in correspondence to $0.6 < x < 0.75$ (see figures 8(c) and 8(d)), while the inlet uncertainty influence the gas volume fraction behavior before and after the shock (see figure 8(a)).

5.2.1. TC3: two-phase flow with pressure uncertainty

The last test case has been proposed in [1] and it reproduces a two-phase shock tube with initial conditions summarized in table 2. Simulations are performed over a physical domain $\Omega = [0, 1]$ until a final time t=200 μs. The time space is divided in 20000 equal time steps of length $\Delta t = 1 \times 10^{-8}$. The simulations are carried out over equally spaced meshes of 101, 201, 401 and 801 points employing the aSI scheme based on the MUSCL method with a VanLeer limiter.
An uncertainty of 5% is supposed for the initial left pressure and its propagation in the shock tube is observed.

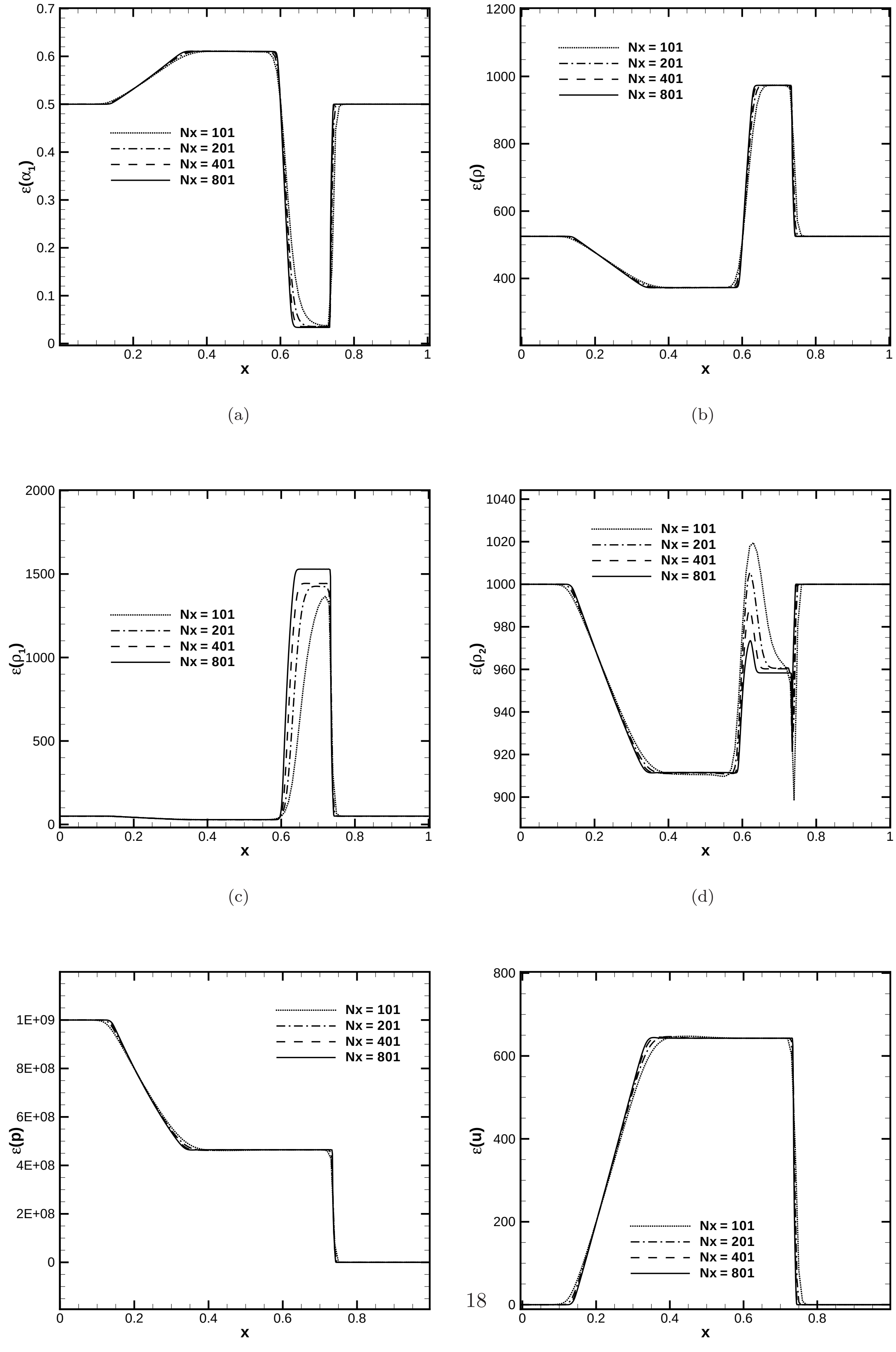
In figure (9), the spatial convergence is reported for both the mean and the variance in L^1 for the mixture density ρ and the gas volume fraction α_g. It has been obtained with the aSI method with a level of 512 ($m = 9$) stochastic cells.

Figures 10 and 7 show the deterministic spatial convergence in terms of mean and variance carried out over equally spaced meshes of 101, 201, 401 and 801 points employing the aSI scheme with a level of 512 ($m = 9$) stochastic cells. The most significant differences between the coarsest and the finest meshes in term of gas density and gas volume fraction in correspondence to $0.8 < x < 0.9$ m (see figures 10(a) and 10(c)). In this test-case, we study the influence of inlet left pressure variation. On the contrary of previous case, the profiles of all thermodynamic variables do not show a significant variation of curves, except in correspondence of shock (see figures 11 and 12).

6. Conclusions

This paper deals with a scheme for simulating stochastic two-phase compressible flows. This scheme relies on a DEM formulation, but reformulated for including an adaptive semi-intrusive scheme (aSI), thus efficiently capturing the propagation of uncertainties. Several test-cases have been investigated. In particular, shock tube configuration has been considered in order to explore the stochastic and grid convergence in different conditions. This scheme displays good convergence properties in each test case for both stochastic and physical spaces. Convergence curves are shown in the physical and stochastic spaces, respectively. Moreover, the variability (in terms of mean and standard deviation) of several properties, such as density, pressure and velocity is computed by considering different kinds of uncertainty, *i.e.* on the initial volume fraction, density or pressure.

Thanks to the robustness of the scheme and to its ability to solve the interface problems, this scheme will be extended to a multi-dimensions investigation in the stochastic and physical spaces.



(a)

(b)

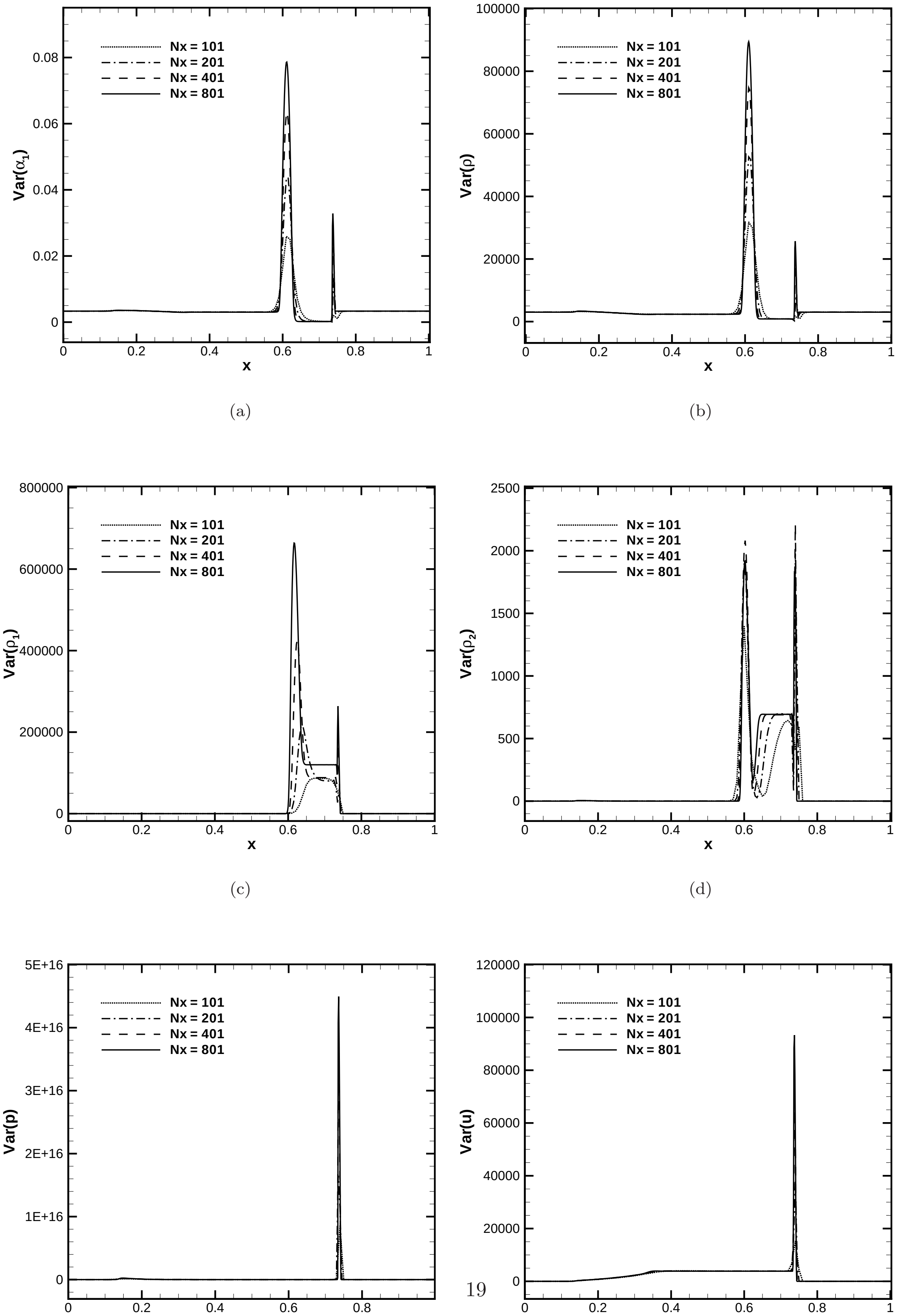
(c)

(d)

(e)

(f)

Figure 6: Deterministic spatial convergence of the expectancy for the (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity. aSI scheme obtained with a level of 512 (m=9).



(a)

(b)

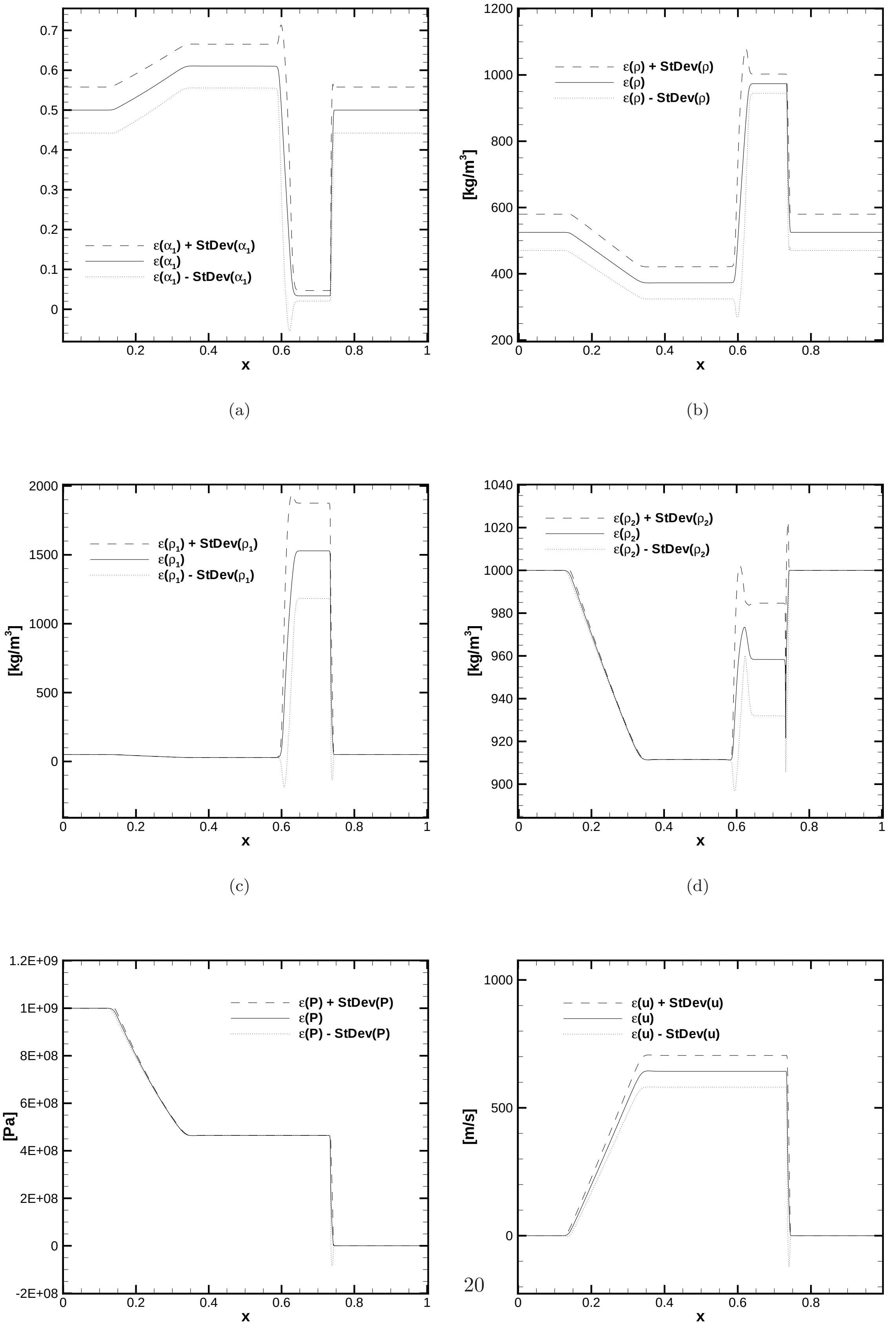
(c)

(d)

(e)

(f)

Figure 7: Deterministic spatial convergence of the variance for (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity. aSI scheme obtained with a level of 512 (m=9).



(a)

(b)

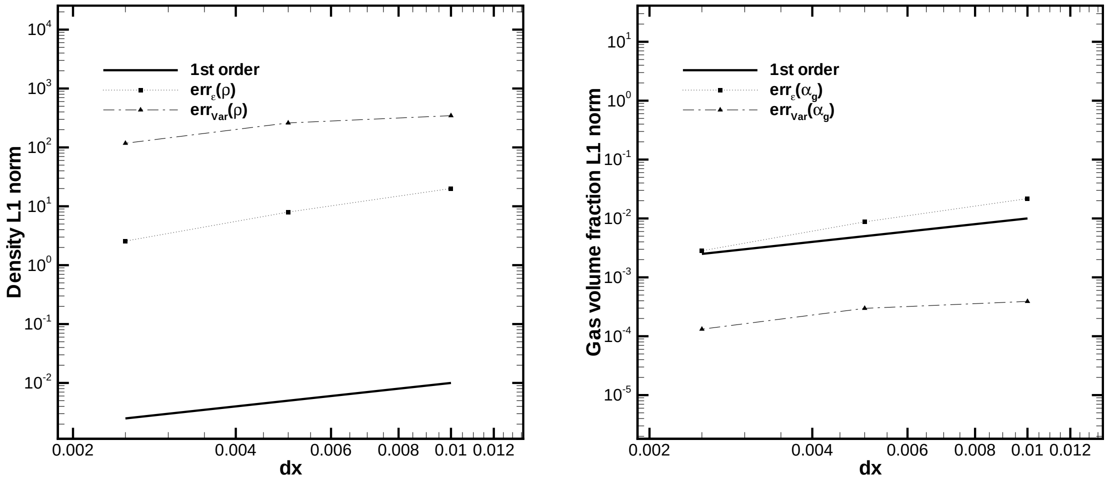
(c)

(d)

(e)

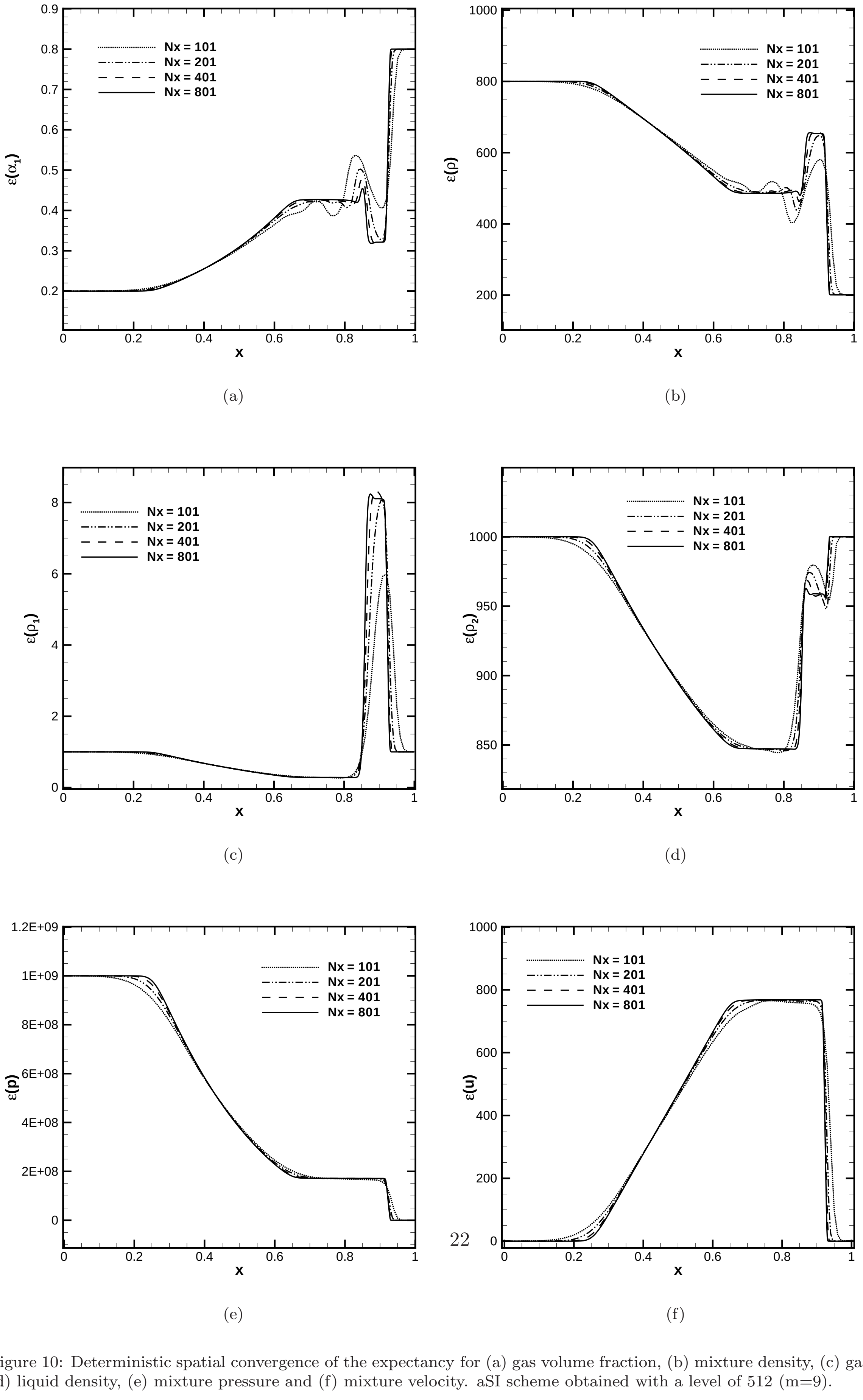
(f)

Figure 8: Confidence intervals ($\mu \pm \sigma$) for (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity. aSI scheme obtained with a level of 512 (m=9).



(a) (b)

Figure 9: Spatial convergence for the stochastic shock tube problem with an uncertain pressure as initial condition. L1 norms are shown for the density (a) and volume fraction (b) expectancy and variance of the solution. aSI scheme obtained with a level of 512 (m=9).



(a)

(b)

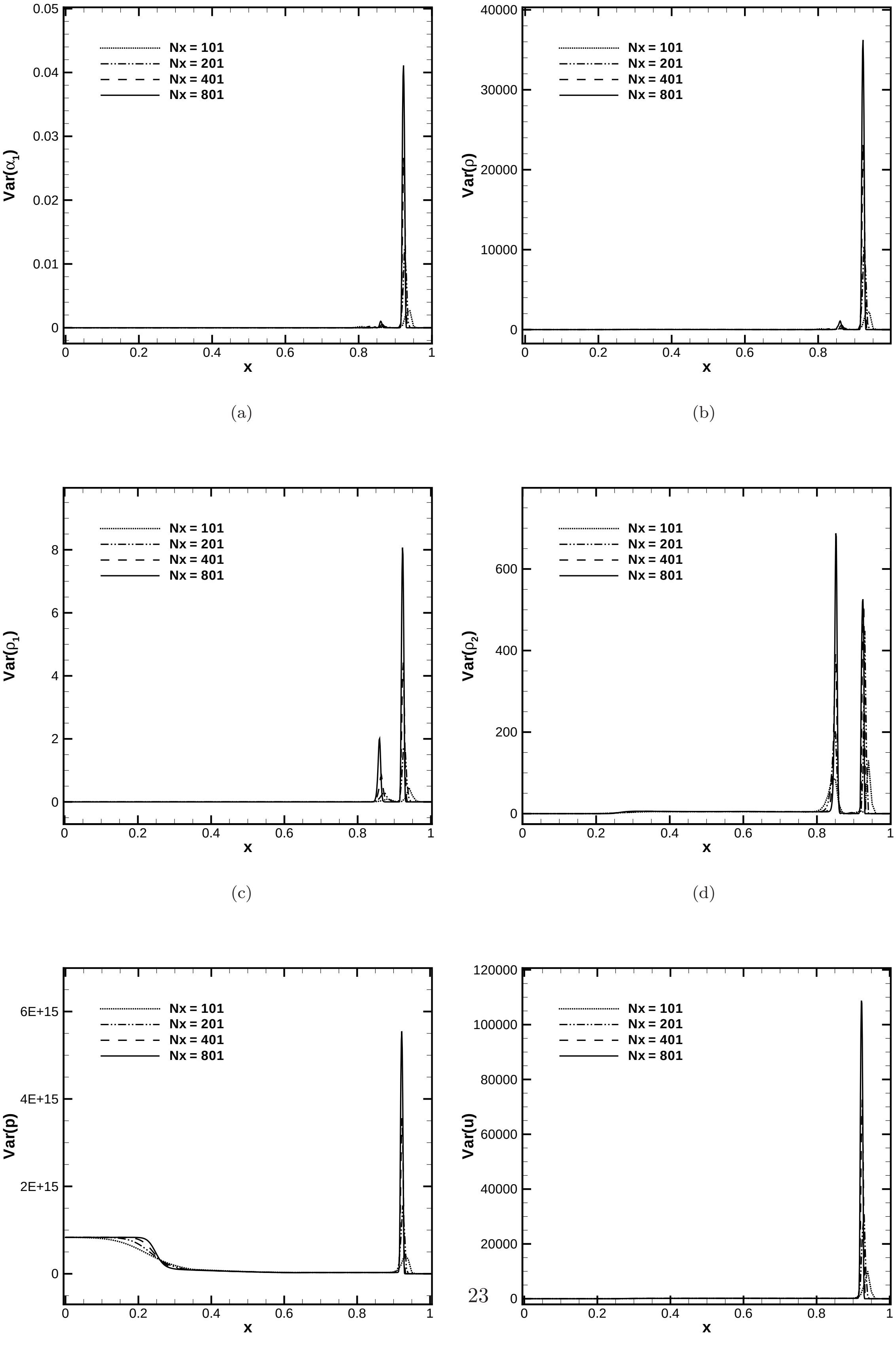
(c)

(d)

(e)

(f)

Figure 10: Deterministic spatial convergence of the expectancy for (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity. aSI scheme obtained with a level of 512 (m=9).



(a)

(b)

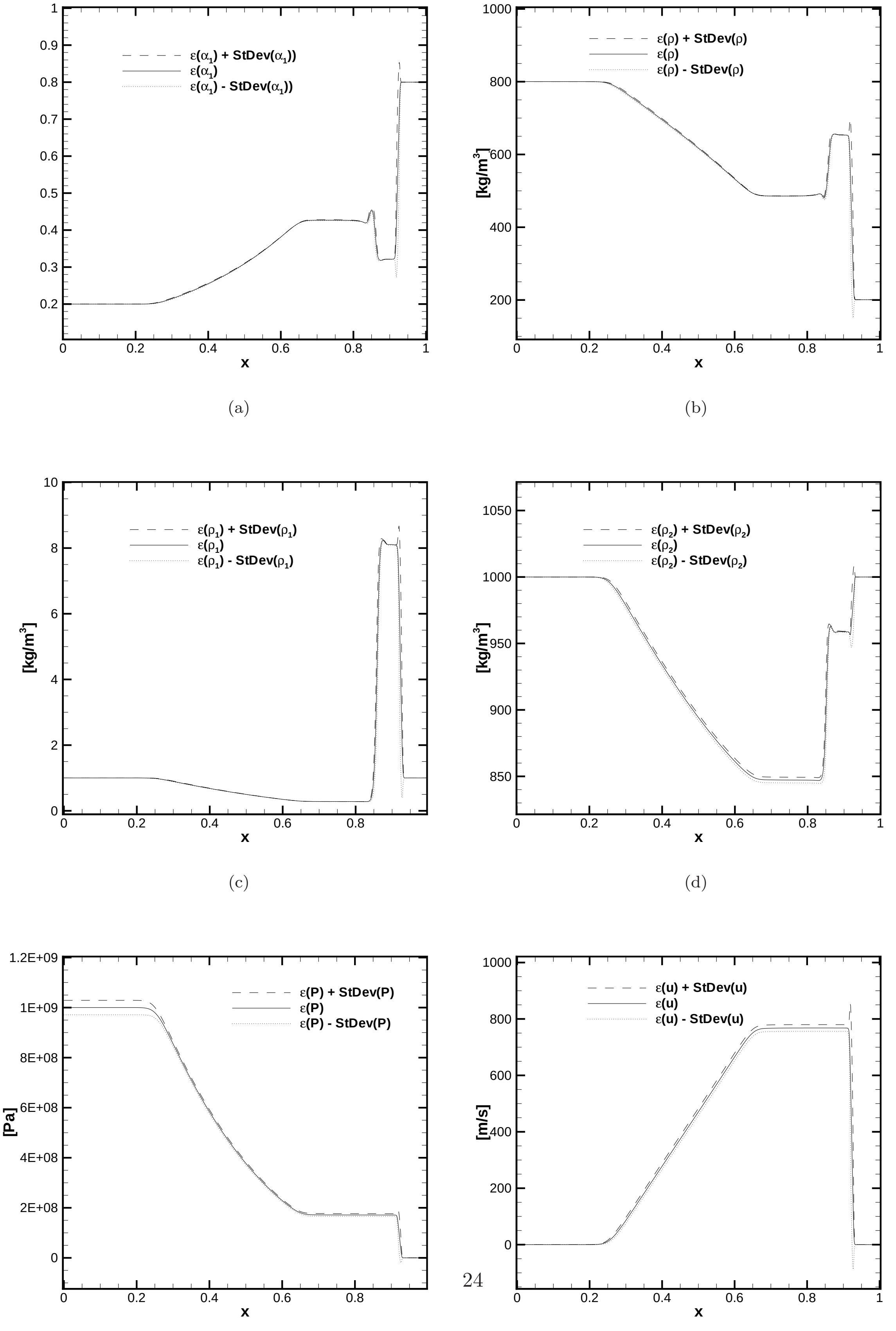
(c)

(d)

(e)

(f)

Figure 11: Deterministic spatial convergence of the variance for (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity variance, respectively. aSI scheme obtained with a level of 512 (m=9).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 12: Confidence intervals ($\mu \pm \sigma$) for (a) gas volume fraction, (b) mixture density, (c) gas density, (d) liquid density, (e) mixture pressure and (f) mixture velocity. aSI scheme obtained with a level of 512 (m=9).

[1] A. Murrone and H. Guillard. A five equation reduced model for compressible two-phase flow problems. *Journal of Computational Physics*, 202:664–698, 2005.

[2] R. Abgrall. How to prevent pressure oscillations in multicomponent flow calculations: a quasi conservative approach. *Journal of Computational Physics*, 125:150–160, 1996.

[3] J. Massoni, R. Saurel, B. Nkonga, and R. Abgrall. Proposition de méthodes et modèles eulériens pour les problèmes à interfaces entre fluides compressibles en présence de transfert de chaleur: Some models and eulerian methods for interface problems between compressible fluids with heat transfer. *International Journal of Heat and Mass Transfert*, 45:1287–1307, 2002.

[4] A. Ambroso, C. Chalons, and P.-A. Raviart. A godunov-type method for the seven-equation model of compressible two-phase flow. *Computers and Fluids*, 54:67–91, 2012.

[5] R. Abgrall and V. Perrier. Asymptotic expansion of multiscale numerical scheme for compressible multiscale flow. *Society for Industrial and Applied Mathematics*, 5:84–115, 2006.

[6] E. Goncalves. Numerical study of expansion tube problems: Toward the simulation of cavitation. *Computers and Fluids*, 72:1–19, 2013.

[7] R. Saurel, F. Petitpas, and R. A. Berry. Simple and efficient relaxation methods for interfaces separating compressible fluids, cavitating flows and shocks in multiphase mixtures. *Journal of Computational Physics*, 228:16781712, 2009.

[8] A. K. Kapila, R. Menikoff, J. B. Bdzil, S. F. Son, and D. S. Stewart. Two-phase modeling of deflagration-to-detonation transition in granular materials: Reduced equations. *Physics of Fluids*, 13:3002–3024, 2001.

[9] M.R. Baer and J.W. Nunziato. A two-phase mixture theory for deflagration to detonation transition (ddt) in reactive granular materials. *International Journal of Multhiphase Flow*, 12:861–889, 1986.

[10] R. Saurel and R. Abgrall. A multiphase godunov method for compressible multifluid and multiphase flows. *Journal of Computational Physics*, 150:425–467, 1999.

[11] R. Saurel and O. Lemetayer. A multiphase model for compressible flows with interfaces, shocks,detonation waves and cavitation. *Journal of Fluid Mechanics*, 431:239–271, 2001.

[12] Y. Sun and C. Beckermann. Diffuse interface modeling of two-phase based on averaging:mass and momentum equations. *Physica D: Nonlinear Phenomena*, 198:281–308, 2004.

[13] O. Le Métayer, J. Massoni, and R. Saurel. Modelling evaporation fronts with reactive riemann solvers. *Journal of Computational Physics*, 205:567610, 2005.

[14] R. Abgrall and R. Saurel. Discrete equations for physical and numerical compressible multiphase mixtures. *Journal of Computational Physics*, 186:361–396, 2003.

[15] G. Allaire, S. Clerc, and S. Kokh. A five-equation model for the simulation of interfaces between compressible fluids. *Journal of Computational Physics*, 181:577–616, 2002.

[16] G.Perigaud and R. Saurel. A compressible flow model with capillary effects. *Journal of Computational Physics*, 209:139–178, 2005.

[17] R. Abgrall and M.G. Rodio. Asymptotic expansion of a multiscale numerical scheme for compressible viscous multiphase flows. *Computers and Fluids*. Submitted.

[18] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics*, 230(10):3668–3694, February 2011.

[19] Ivo Babuška, Fabio Nobile, and Raul Tempone. A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. *SIAM Review*, 52(2):317, 2010.

[20] Jasmine Foo and George Em Karniadakis. Multi-element probabilistic collocation method in high dimensions. *Journal of Computational Physics*, 229:1536–1557, March 2010.

[21] Olivier Le Maître. Uncertainty propagation using WienerHaar expansions. *Journal of Computational Physics*, 197(1):28–57, June 2004.

[22] Olivier Le Maître. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *Journal of Computational Physics*, 197(2):502–531, July 2004.

[23] J Tryoen, Olivier Le Maître, M Ndjinga, and A Ern. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *Journal of Computational Physics*, 229:6485–6511, 2010.

[24] Julie Tryoen. *Methodes de Galerkin stochastiques adaptatives pour la propagation d'incertitudes parametriques dans les systemes hyperboliques*. PhD thesis, Univeriste Paris-Est.

[25] P Pettersson and Gianluca Iaccarino. Numerical analysis of the Burgers equation in the presence of uncertainty. *Journal of Computational Physics*, 228:8394–8412, 2009.

[26] R. Abgrall and P.M. Congedo. A semi-intrusive deterministic approach to uncertainty quantification in non-linear fluid flow problems. *Journal of Computational Physics*, 235:828–845, 2013.

[27] R. Abgrall, P.M. Congedo, G. Geraci, and G. Iaccarino. An adaptive multiresolution semi-intrusive scheme for uq in compressible fluid problems. *International Journal for Numerical Methods in Fluids*. Submitted.

[28] A. Gel, R. Garg, C. Tong, M. Shahnam, and C. Guenther. Applying uncertainty quantification to multiphase flow computational fluid dynamics. *Powder Technology*, 242:27–39, 2013.

[29] K. K. So, T. Chantrasmi, X.Y. Hu, J. Witteveen, C. Stemmer, G. Iaccarino, and et al. Uncertainty analysis for shock-bubble interaction. In *Proceedings of the 2010 summer program, center for turbulence research. USA: Stanford University.*, 2010.

[30] M. Presho, A. Malqvist, and V. Ginting. Density estimation of two-phase flow with multiscale and randomly perturbed data. *Advances in Water Resources*, 33:1130–1141, 2010.

[31] R. Conejeros and B. Lenoach. Effect of uncertainty on 2-phase flow into a horizontal completion. *Journal of Petroleum Science and Engineering*, 58:309–324, 2007.

[32] H. Li and D. Zhang. Efficient and accurate quantification of uncertainty for multiphase flow with the probabilistic collocation method. *SPE J.*, 14:665–679, 2009.

[33] P. Pettersson, G. Iaccarino, and J. Nordstrom. An intrusive hybrid method for discontinuous two-phase flow under uncertainty. *Computers and Fluids*, 86:228–239, 2013.

[34] Remi Abgrall, Maria Giovanna Rodio, and Pietro Marco Congedo. Towards an efficient algorithm for the simulation of viscous two-phase flows with real gas effects. Research Report RR-8173, INRIA, December 2012.

[35] D.A. Drew and S.L. Passman. *Theory of Multicomponent Fluids*, volume 135. Applied Mathematical Sciences, Springer, New York, 1998.

[36] F. Coquel and B. Perthame. Relaxation of energy and approximate riemann solvers for general pressure laws in fluid dynamics. *Society for Industrial and Applied Mathematics*, 35:2223–2249, 1998.

[37] E.F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, Berlin, 1997.

[38] Ami Harten. Adaptive multiresolution schemes for shock computations. *Journal of Computational Physics*, 115(2):319–338, August 1994.

[39] Ami Harten. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Communications on Pure and Applied Mathematics*, 48(12):1305–1342, 1995.

[40] Ami Harten. Multiresolution Representation of Data : A General Framework. *SIAM Journal on Numerical Analysis*, 33(3):1205–1256, 1996.

[41] Rémi Abgrall, Pietro Marco Congedo, and Gianluca Geraci. A One-Time Truncate and Encode Multiresolution Stochastic Framework. *Journal of Computational Physics*, 2012. In Press.

[42] Rémi Abgrall, Pietro Marco Congedo, and Gianluca Geraci. Toward a Unified Multiresolution Scheme in the Combined Physical / Stochastic Space for Stochastic Differential Equations. *Mathematics and Computers in Simulation*, 2012. Submitted.

[43] Rémi Abgrall, Pietro Marco Congedo, Gianluca Geraci, and Gianluca Iaccarino. Non-linear Multiresolution framework for Uncertainty Quantification in Computational Fluid Dynamics. *Journal of Computational and Applied Mathematics*, 2013. Submitted.

Paper *P6*

Received 29 January 2012
Revised 20 August 2012
27 December 2012
Accepted 4 April 2013

TSI metamodels-based multi-objective robust optimization

Pietro Marco Congedo, Gianluca Geraci and Rémi Abgrall
INRIA Bordeaux Sud-Ouest, Talence, France, and
Valentino Pediroda and Lucia Parussini
Mechanical and Naval Engineering Department, Università di Trieste, Trieste, Italy

Abstract

Purpose – This paper aims to deal with an efficient strategy for robust optimization when a large number of uncertainties are taken into account.

Design/methodology/approach – ANOVA analysis is used in order to perform a variance-based decomposition and to reduce stochastic dimension based on an appropriate criterion. A massive use of metamodels allows reconstructing response surfaces for sensitivity indexes in the design variables plan. To validate the proposed approach, a simplified configuration, an inverse problem on a 1D nozzle flow, is solved and the performances compared to an exact Monte Carlo reference solution. Then, the same approach is applied to the robust optimization of a turbine cascade for thermodynamically complex flows.

Findings – First, when the stochastic dimension is reduced, the error on the variance between the reduced and the complete problem was found to be roughly estimated by the quantity $(1 - \bar{T}_{TSI}) \times 100$, where $\bar{T}_{TSI}$ is the summation of TSI concerning the variables respecting the TSI criterion. Second, the proposed strategy allowed obtaining a converged Pareto front with a strong reduction of computational cost by preserving the same accuracy.

Originality/value – Several articles exist in literature concerning robust optimization but very few dealing with a global approach for solving optimization problem affected by a large number of uncertainties. Here, a practical and efficient approach is proposed that could be applied also to realistic problems in engineering field.

Keywords ANOVA, Kriging, Metamodel, Robust optimization, Uncertainty quantification

Paper type Research paper

1. Introduction

Dealing with moderate and high dimensional stochastic spaces is actually one of the most important problem in uncertainty quantification (UQ) community. Several methods are proposed in Agarwal and Aluru (2009), Baglietto *et al.* (2010), Blatman and Sudret (2011), Caflisch *et al.* (1997), Cao *et al.* (2003), Foo *et al.* (2008), Ma and Zabaras (2010) and Wang and Sloan (2003), but their accuracy on realistic problems with highly non-linear effects is still not proven. This problem is even more challenging when coupled to the robust design optimization, where conventional optimization procedures

Rémi Abgrall has been financed in part by the EU ERC Advanced Grant "ADDECCO" N. 226316. Gianluca Geraci has been fully financed by the EU ERC Advanced Grant "ADDECCO" N. 226316.

Engineering Computations: International Journal for Computer-Aided Engineering and Software
Vol. 30 No. 8, 2013
pp. 1032-1053
© Emerald Group Publishing Limited
0264-4401
DOI 10.1108/EC-01-2012-0012

aims at taking in account uncertainty in the design procedures (for a detailed review see Verstraete and Périaux (2010) and Schuëller and Jensen (2008)). Finding a design which is relatively invariant with respect to parameter variability, production tolerance, and other uncertain conditions is referred to as robust design optimization. Robust design method, also called the Taguchi method, was pioneered in Taguchi (1989) even if it suffers important limitations from an optimization efficiency point of view. Several methods incorporate the uncertainties into the optimization problem. For example, the metamodel approach uses the data to build a metamodel of the robustness measures by using a set of given design points. In this context, the response surface methodology, neural networks, Kriging models, have been proposed as metamodeling techniques (Jensen and Catalan, 2007; Namura *et al.*, 2011; Xu and Albin, 2003). In general, metamodeling techniques are not well suited for large-scale robust optimization problems when the number of design variables is large. Another class of methods, usually referred to as the deterministic approach to robust optimization, calculate explicitly the desired robustness measures. In this class, several methods prove their efficiency for high-dimensional uncertain parameter (Sankaran *et al.*, 2010).

The large amount of computational effort required for considering a large number of uncertainties is well known in literature as curse of dimensionality (Bellman and Richard, 1961). Two different methodologies have been proposed to tackle this issue in the UQ framework. First, a strategy to reduce the number of points required for the numerical integration, named sparse grid (Gao and Hesthaven, 2010; Ma and Zabaras, 2010), has been introduced. This technique can lead to a strong reduction of the quadrature points for moderate dimensional problem, provided that the function has some regularity. It is based essentially on some results of the interpolation theory. Certainly the sparse grid allows to avoid the exponential growth of the number of points with the stochastic dimension, so in this sense prevents from the curse of dimensionality, but for high dimensional stochastic spaces the number of simulations required could be equally prohibitive. More recently, the attention has shifted to both the number of points required and the number of stochastic dimensions. There are a few studies (Congedo *et al.*, 2011c; Foo *et al.*, 2008; Gao and Hesthaven, 2010), exploring the possibility to identify the most important uncertainties and as a consequence reducing the number of dimension of the stochastic space. If the number of uncertainties could be reduced, a better statistics estimation could be achieved with a lower computational cost. This reduction strategy can be used into a robust optimization framework, thus decreasing the final cost for obtaining the optimal individual. Generally, ANOVA-based approaches are used in order to decompose the variance according to the different contributions, permitting to create a ranking of the most predominant uncertainties (Congedo *et al.*, 2011b).

In this work, the focus is to propose an optimization technique, partially presented in Congedo *et al.* (2011c), applied to stochastic partial differential equations (PDE), for which evaluating fitness function is very expensive. An innovative algorithm is proposed based on the following steps:

- computing the ranking of most predominant uncertainties (via ANOVA) for some samples in the design space;
- building a response surface of the sensitivity indexes (associated to each uncertainty) in the design space; and
- solving the reduced stochastic problem for each design during the optimization.

This strategy is applied first to a simple test-case where the computation of the reference solution is feasible: the optimization of a 1D compressible nozzle flow. Then, it is applied to the optimization of the working conditions of a thermodynamically complex flow in a turbine cascade.

In Section 2, the problem is defined and the main numerical ingredients, i.e. ANOVA decomposition, Sobol sensitivity indexes and description of the criterion used for judging whether one uncertainty should be discarded in the reduced stochastic problem, are presented. The algorithm proposed in this paper, with a discussion over the computation cost, is described in Section 3. Then, a simple test-case, i.e. the optimization of a compressible nozzle flow, is presented in Section 4. Section 5 deals with the application of the proposed strategy to a realistic configuration: the optimization of a turbine cascade. Finally, conclusions and perspectives are depicted in Section 6.

2. Problem formulation

Let us consider to have a system of equations and an output of interest, i.e. a fitness function $f = f(y, \xi)$, where the vector y is the ensemble of N design parameters $y \in \Theta \subset \Re^N$. Moreover, let us suppose that the output of the system can be dependent by d uncertainties parameters ξ_i assumed so that $\xi = \{\xi_1, \ldots, \xi_d\} \in \Xi \subset \Re^d$ with $p(\xi_i)$ the probability density function associated to the ith stochastic dimension. If independent distributed random variables $\xi_i \in \Xi_i$ are considered, the space Ξ can be obtained by tensorization of monodimensional spaces, i.e. $\Xi_i \subset \Re$, $\Xi = \Xi_1 \times \cdots \times \Xi_d$ and it follows that $p(\xi) = \prod_i p(\xi_i)$.

Let us formulate the robust optimization problem as follows:

$$\min_y g(f(y, \xi)), \tag{1}$$

where:

$$g(f(y, \xi)) = \int_{\Xi} f^k p(\xi) d\xi, \tag{2}$$

i.e. minimizing (for example) some statistical moments (of order k), g, of the fitness function f with respect to ξ.

Computation of the statistical moments can be prohibitive if d is too high, because of the well-known problem of curse of dimensionality (Bellman and Richard, 1961). One possibility is to compute a reduced set of predominant uncertainties, ξ_r, for calculating g with a lower computational cost as follows:

$$g(f(y, \xi)) \approx \int_{\Xi_r} f^k p(\xi_r) d\xi_r, \tag{3}$$

In this work, an efficient model reduction in the stochastic space and an efficient optimization algorithm are proposed. In this section, we illustrate how a model reduction (solving equations (2) and (3)) can be performed, in terms of ANOVA decomposition (Section 2.1), total sensitivity index (TSI) computation (permitting to identify the most important uncertainties in terms of contribution to the variance) (Section 2.2) and a reduction criterion (Section 2.3). Then, the complete algorithm for solving equations (1)-(3) is described in Section 3.

2.1 ANOVA analysis

Let us assume, with no loss of generality, that the fitness function is dependent only on the stochastic vector $\xi : f = f(\xi)$.

Assuming $f(\xi) \in L^2(\xi, p(\xi))$, a Sobol unique functional decomposition exists:

$$f(\xi) = \sum\nolimits_{u \subseteq (1,\ldots,d)} f_u(\xi_u) \tag{4}$$

here u is a set of integers with cardinality $v = |u|$ and $\xi_u = \{\xi_{u_1}, \ldots, \xi_{u_v}\}$. Each function f_u is computed by using the following relation (Crestaux *et al.*, 2009):

$$f_u(\xi_u) = \int_{\Xi_u} f(\xi)p(\xi_u)d\xi_u - \sum\nolimits_{w \subset u} f_w(\xi_w) \tag{5}$$

where $\Xi_u \subset \Xi$ is obtained directly by tensorization of the monodimensional spaces Ξ_i associate to the multi-index u, i.e. $\Xi_u = \Xi_{u_1} \times \cdots \times \Xi_{u_v}$.

The mean of the function $f(\xi)$ can be computed using the definition:

$$f_0 = \int_{\Xi} f(\xi)p(\xi)d\xi. \tag{6}$$

This functional decomposition is called ANOVA if each of the 2^d elements of the decomposition, except f_0, verifies for every ξ_i:

$$\int_{\Xi_i} f_u(\xi_u)p(\xi_i)d\xi_i = 0, \;\; \forall i \in u \tag{7}$$

From equation (7), it follows the orthogonality:

$$\int_{\Xi} f_u(\xi_u)f_w(\xi_w)p(\xi)d\xi = 0, \;\; u \neq w \tag{8}$$

The Sobol functional decomposition (equation (4)) allows to identify the contribution of each single variable and all the coupled interactions. The great advantage of the ANOVA decomposition is the possibility to link each term of the representation (equation (4)) to the corresponding contribution to the variance of the original function. Employing the decomposition defined in equation (4), it is possible to decompose the variance of $f = f(\xi)$ (obviously proportional to $f^2(\xi)$) as follows:

$$\sigma^2(f) = \sum\nolimits_{\substack{u \subseteq \{1,\ldots,d\} \\ u \neq 0}} \sigma_u^2(f_u) \tag{9}$$

where:

$$\sigma_u^2(f_u) = \int_{\Xi_u} f_u^2(\xi_u)p(\xi_u)d\xi_u \tag{10}$$

and $\Xi_u = \Xi_{u_1} \times \cdots \times \Xi_{u_v}$.

Now, the contribution of each term in equation (9) can be evaluated by means of the so-called Sobol sensitivity indexes (SI) defined as:

$$S_u = \frac{\sigma_u^2}{\sigma^2}. \tag{11}$$

They measure the sensitivity of the variance due to the v-order ($v = |u|$) interaction between the variables in ξ_u. Remark that the summation of the $2^d - 1$ Sobol indexes is equal to one. Another index, called TSIs, can be computed summing all the SI terms where the uncertainty $j \in u$ appears:

$$TSI_j = \sum_{j \in u} S_u. \tag{12}$$

This index allows to estimate the overall importance of a single stochastic variable. It can be used (as done in this work) as an estimator for judging whether one uncertainty could be discarded. The total amount of the TSI, defined as follows:

$$T_{TSI} = \sum_{j=1}^{d} TSI_j \tag{13}$$

can be used for evaluating the non-linearity of the system. It takes always values greater than 1 (it could be equal to 1 only in case of no interactions among the uncertainties, so TSI should be equal to SI for each uncertainty).

To make things clearer, let us apply this analysis to a specific case, where $f(\xi)$ is dependent only on two stochastic variables. The Sobol decomposition (equation (4)) becomes:

$$f(\xi) = f_0 + f_1 + f_2 + f_{12}$$

where:

$$f_1(\xi_1) = \int_{\Xi_2} f(\xi)p(\xi_2)d\xi_2 - f_0, \quad f_2(\xi_2) = \int_{\Xi_1} f(\xi)p(\xi_1)d\xi_1 - f_0,$$

$$f_{1,2} = f(\xi) - f_0 - f_1 - f_2$$

For this example, the property (equation (7)) reduces to:

$$\int_{\Xi_1} f_1(\xi_1)p(\xi_1)d\xi_1 = \int_{\Xi_2} f_2(\xi_2)p(\xi_2)d\xi_2 = \int_{\Xi} f_{12}(\xi_1, \xi_2)p(\xi)d\xi = 0$$

and the orthogonality is expressed as follows:

$$\int_{\Xi} f_1(\xi_1)f_2(\xi_2)p(\xi)d\xi = \int_{\Xi} f_1(\xi_1)f_{12}(\xi_1, \xi_2)p(\xi)d\xi = \int_{\Xi} f_2(\xi_2)f_{12}(\xi_1, \xi_2)p(\xi)d\xi = 0$$

In this case, variance can be decomposed as follows:

$$\sigma^2(f) = \sigma_1^2(f_1) + \sigma_2^2(f_2) + \sigma_{12}^2(f_{12})$$

As a consequence, the first-order contribution of each uncertainty and the contribution of the interaction can be easily computed.

2.2 TSI computation from polynomial chaos expansion

Sobol indexes can be computed by using a whatever sampling stochastic method (Monte Carlo, quasi-Monte Carlo), as shown in the paper of Sobol (2001), but can be done more efficiently when a polynomial chaos expansion (PCE) is used (Crestaux *et al.*, 2009). Let us remember the PCE:

$$f(\xi) = \bar{f}(\xi) + O_T = \sum_{k=0}^{P} \beta_k \phi_k(\xi) + O_T, \tag{14}$$

where the number of terms is related to the maximal degree of the polynomial reconstruction n_0 and the dimension of the system:

$$d : P + 1 = \frac{(n_0 + d)!}{n_0! d!}.$$

In this work, the coefficients of the PCE are computed by a quadrature employing the points generated by a full tensorization of monodimensional quadrature rules. In particular, employing uniform distribution for the stochastic variables, a Legendre quadrature is chosen as monodimensional quadrature rule following the so-called Wiener-Askey scheme (Askey and Wilson, 1985). After the evaluation of f in each point of the full tensorization, the coefficients can be computed exploiting the orthogonality of the basis, as follows:

$$\beta_k = \frac{\langle f(\xi), \Psi_k(\xi) \rangle}{\langle \Psi_k(\xi), \Psi_k(\xi) \rangle}, \quad k = 0, \ldots, P, \tag{15}$$

where:

$$\langle f(\xi), g(\xi) \rangle = \int_{\Xi} f(\xi) g(\xi) d\xi$$

indicates the inner product.

For further details on the polynomial chaos techniques see Le Maître (2005). Each element f_u of the functional decomposition of $f(\xi)$ is approximated by the relative term f_u:

$$f_u(\xi_u) \approx \bar{f}_u(\xi_u) = \sum_{k \in K_u} \beta_k \Psi_k(\xi_u), \tag{16}$$

where the set of indexes K_u is given by:

$$K_u = \{k \in \{1, \ldots, P\} | \Psi_k(\xi_u)\} \tag{17}$$

where each term $\Psi_k(\xi_u)$ represents the polynomial term of the expansion (equation (14)) dependent only from the set of variables ξ_u.

Because of the orthogonality, the variance $(\bar{\sigma}^2(f) = \sigma^2(\bar{f}) \approx \sigma^2(f))$ and the conditional variance $(\bar{\sigma}_u^2(f_u) = \sigma_u^2(\bar{f}_u) \approx \sigma_u^2(f_u))$ can be computed as follows:

$$\begin{aligned}\sigma^2(f_u) &= \sum_{k=1}^{P} \beta_k^2 \langle \Psi_k, \Psi_k \rangle \\ \sigma_u^2(f_u) &= \sum_{k \in K_u} \beta_k^2 \langle \Psi_k, \Psi_k \rangle\end{aligned}. \tag{18}$$

Sobol sensitivity indexes can be computed directly from equation (18):

$$S_u \approx \bar{S}_u = \frac{\sigma_u^2(f_u)}{\bar{\sigma}^2(f_u)} = \frac{\sum_{k \in K_u} \beta_k^2 \langle \Psi_k, \Psi_k \rangle}{\sum_{k=1}^{P} \beta_k^2 \langle \Psi_k, \Psi_k \rangle} \tag{19}$$

where the TSI is defined by equation (12).

2.3 Sensitivity index criterion for PDE

Once TSI_j is computed (for each uncertainty j), the problem is to specify the threshold that TSI_j must not exceed, before the uncertainty j could be discarded, i.e. to choose a criterion for the reduction of the stochastic dimension. In a recent work, Gao and Hesthaven (2010) proposed a criterion based on TSI in order to identify the most important parameters in the resolution of stochastic ordinary differential equations. They proposed to freeze, i.e. replace with their mean values, all the stochastic variables whose TSI_j is inferior to 2 percent. This permits to obtain a good prediction of the statistical moment at a lower computational cost.

The aim is to reduce $f = f(\xi)$ in the problem $f^R = f^R(\xi_r)$ where $\dim(\xi_r) \leq \dim(\xi)$, $n = \dim(\xi) - \dim(\xi_r)$, $\xi_r \in \Xi_r = \Xi_{r_1} \times \cdots \times \Xi_{r_{d-n}} \subset \Re^{d-n}$ and n is equal to the number of uncertainties that are discarded. In this case, the relative total TSI, $(\bar{T}_{TSI})$, that is the total amount of the TSI for the reduced problem normalized by the total amount of TSI for the complete problem, T_{TSI}, is equal to[1]:

$$\bar{T}_{TSI} = \frac{\sum_{j=1}^{d-n} TSI_j}{T_{TSI}} < 1. \tag{20}$$

Here, the approach proposed in Gao and Hesthaven (2010) is applied to several PDE, doing systematically the following steps:

- Solve the complete stochastic problem using a quasi-Monte Carlo method and compute reference mean and variance.
- Apply the ANOVA analysis computing the ranking of most predominant uncertainties in terms of TSI_j.
- Solve the reduced stochastic problem, obtained by discarding progressively the less influent uncertainties basing on TSI_j (2 percent criterion) and compute mean and variance.
- Compute the relative error between the reduced and complete stochastic problem in terms of mean and variance.

The efficacy of the 2 percent criterion is widely investigated on elliptic, parabolic and hyperbolic PDE. This campaign, not reported here for brevity and whose details are reported in Abgrall *et al.* (2012) leads to these conclusions:

- the error on the mean is always inferior to 1 percent except for the parabolic case; and
- the error on the variance can be roughly estimated as equal to $(1 - \bar{T}_{TSI}) \times 100$.

Finally, the estimation proposed in Gao and Hesthaven (2010), remains valid also for different kinds of PDE. Hence, it is used in the algorithm described in the next section.

3. Optimization algorithm

In this section, the algorithm for multi-objective robust design optimization is described. Then, let us illustrate how to solve equations (1)-(3). The algorithm is constituted by two main steps, that are schematically shown in Figure 1.

During the first step (reported in Figure 1(a)), the focus is on building a response surface for each TSI_j, $T\tilde{S}I_j$, in the design space. This step is constituted by the following actions:

(1) An initial set of N designs (design variables y_l with $l = 1, \ldots, N$) in the design space, i.e. a design of experiment (called hereafter DOE), is generated.

(2) Equation (2) is solved for each design y_l using the PC expansion or other techniques (quasi-Monte Carlo, collocation, etc.).

(3) TSI_j is computed for each uncertainty j and for each y_l.

(4) A response surface in the design space for each uncertainty j, $T\tilde{S}I_j(y)$, is built by using a Kriging method based on a DACE approach (Novak and Ritter, 1996) and the set of $TSI_j(y_l)$, that are computed at the previous step. The advantage of DACE approach is the possibility of implementing an adaptive response surface in order to minimize the statistical error between the real function and the extrapolated one.

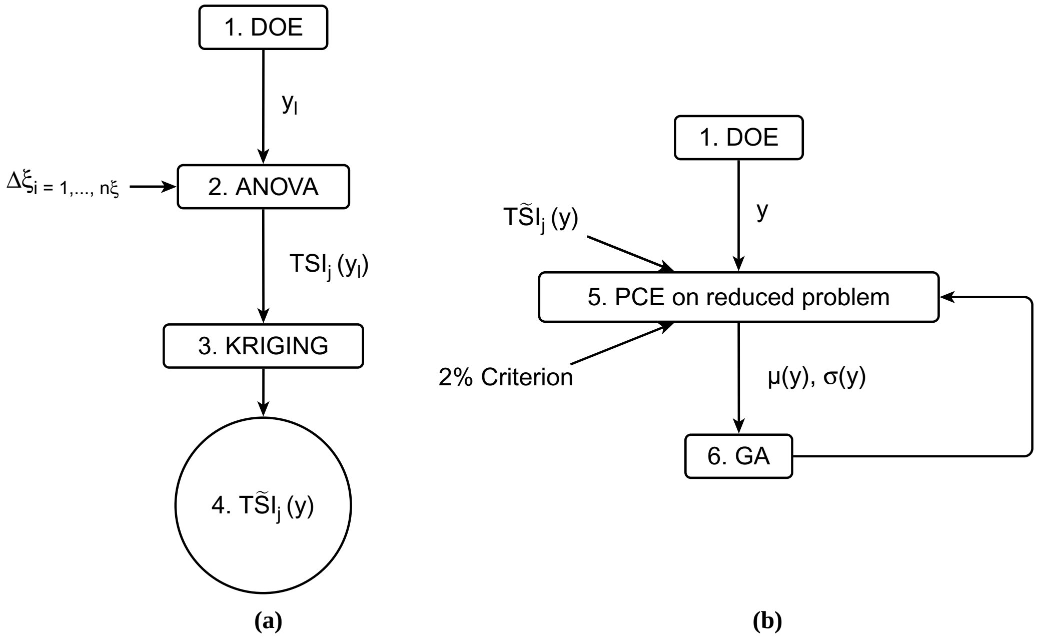


Figure 1.
Compact scheme for the Kriging procedure (a) and the overall optimization strategy (b)

During the second step of the algorithm (reported in Figure 1(b)), the focus is on solving equation (1). The optimizer is the NSGA-II algorithm (Deb *et al.*, 2002). The main tuning parameters of the algorithm are the population size, the number of generations, the crossover and mutation probabilities and the distribution indexes for crossover and mutation operators. Typical values for the last four parameters are, respectively, 0.9, 1, 20 and 20. Remark that the global strategy proposed in this work can be applied for a whatever kind of optimizer.

In order to initialize the genetic algorithm, the same initial set of N samples, y_l, considered during the first step, are taken into account.

Now, at each iteration of the proposed algorithm (i.e. for each evaluation of the fitness function), the following operations are performed:

(5) Supposing a given design y_n, the response surface is used to compute the approximated value of TSI_j, i.e. $TSI_j(y_n) \cong T\tilde{S}I_j(y_n)$, for each uncertainty. Moreover, the 2 percent criterion on TSI_j is applied to build the reduced set of uncertainties to consider, i.e. ξ_r. Finally, the stochastic problem, expressed in equation (3), is solved by means of PC expansion.

(6) Fitness functions, in terms of mean and variance, are computed and used by the optimizer.

When convergence is reached, the fitness functions of the optimal designs are re-computed by considering the whole set of uncertainties.

3.1 Estimated savings in terms of computational cost

The computational cost of solving the optimization problem expressed in equations (1) and (2) can be roughly estimated as follows:

$$Cost = N_{des} \times N_s \times C_{\det},$$

where N_{des} is the number of designs generated during the optimization, N_s is the number of stochastic samples used for solving equation (2), and $C_{\det}$ is the cost of a deterministic simulation. Moreover, N_s requires $(p+1)^d$ deterministic computations, with a polynomial expansion of order p (sufficient for the convergence of the statistics estimation (equation (2))) and d the total number of uncertainties. Remark that, since $g(f)$ depend on y, p and then N_s could be different between two different designs. As a consequence, a more correct estimation is the following:

$$Cost = C_{\det} \times \sum_{i=1}^{N_{des}} (p(y)+1)^d.$$

The use of the proposed approach could reduce N_s. In particular, using the same notation introduced in Section 2, if r is the number of uncertainties in the reduced problem, the savings in terms of computational cost can be estimated as follows:

$$1 - \frac{\sum_{i=1}^{N_{des}} (p_r(y)+1)^{r(y)}}{\sum_{i=1}^{N_{des}} (p(y)+1)^d}, \tag{21}$$

where $p_r(y)$ is the order of the polynomial expansion (sufficient for the convergence of the statistics estimation (equation (3))) for the reduced problem. Remark that

$p_r(y) \leq p(y)$. Remark also that the cost of the first step described in this section, is equal to the cost for initializing the complete robust optimization (taking into account the whole set of uncertainties), i.e. applying ANOVA analysis for the computation of TSI_j is not an adding cost with respect to the computation of the statistical moments. As a consequence, savings in terms of computational cost of the proposed algorithm exist only when the second step is applied.

If we suppose a minimal order of p equal to 2 for ensuring statistics convergence, savings should be at least greater than $1 - \sum_{i=1}^{N_{des}}(3)^{r(y)}/\sum_{i=1}^{N_{des}}(3)^{d}$. To illustrate some practical example, if one uncertainty can be systematically discarded when d = 5 (then r = 4), the savings is at least greater than 67 percent.

4. Verification of the proposed approach on a nozzle design

The algorithm proposed in Section 3 is validated on a simplified configuration, namely, transonic flow through a quasi-1D nozzle. For this configuration, a fast exact deterministic solution is available, which allows to perform a robust optimization with a complete stochastic analysis of a flow problem affected by a large number of uncertain parameters at a moderate computational cost. As a consequence, the fully complete optimization (i.e. fitness functions are always evaluated on the whole set of uncertainties) and the proposed algorithm, can be compared in terms of accuracy and computational cost.

The reduced cost of the exact solver enables also the computation of the reference solution for the complete and reduced strategy with a quasi-Monte Carlo approach with Sobol sequences.

Let us consider the steady flow of a perfect gas (with heat coefficient ratio γ) in the following convergent-divergent nozzle geometry:

$$y = (A_e/A_t - 1 - \alpha - \beta)x^3 + \beta x^2 + \alpha x + 1, \tag{22}$$

where A_e and A_t are the cross-sectional exit area and the throat area, respectively. Moreover, α and β are two geometric parameters. The nozzle pressure ratio (exit pressure over reservoir pressure) is denoted by p_e/p_0. The values it can take are constrained so that a shock is always created in the nozzle divergent.

Stochastic solutions for the flow field inside the nozzle are computed by assuming that the following parameters can be considered as uniformly distributed random variables: A_e/A_t, α and β (representative of geometric tolerances), the gas specific heat ratio γ (it represents the thermodynamic properties of the gas), p_e/p_0 (defining the operating conditions of the nozzle), with variation ranges given in Table I.

The only design parameter is the geometric parameter α (that can assume a value α_0 between 0.05 and 0.1). The optimization problem is formulated as follows:

Variable	Min.	Max.
γ	1.3	1.5
p_{es}/p_0	0.8181855	0.8347145
A_e/A_t	1.75	1.96
α	± 3 percent	
β	0.4	0.6

Table I. Min./max. values for the uncertainties in the nozzle flow problem

$$\min_{\alpha} \left|\sigma^2(x_s) - \sigma^2_{tar}\right|. \tag{23}$$

where x_s is the shock position in the divergent part, σ^2_{tar} (equal to 1.045×10^{-04}) is the target variance, and the uncertainties are the ones given in Table I, where α is assumed to vary between $0.97 \cdot \alpha_0$ and $1.03 \cdot \alpha_0$.

The fully complete robust optimization and the proposed algorithm have produced a very similar optimal design (the exact one is obtained with α equal to 0.055), with a relative error of 0.008 percent. A very promising reduction of 99.1 percent in terms of computational cost is obtained, because the complete stochastic problem (equation (2)) with five uncertainties is systematically reduced to only one uncertainty-problem (equation (3)). Remark that in this case, one predominant uncertainties, p_e/p_0, exist, then this is the best case scenario in order to show the proposed algorithm advantages. For more realistic cases, where highly non-linear effects could exist, the gain could be strongly reduced. For this reason, a complex configuration is considered in the next section.

5. Results on complex flows in a turbine cascade

5.1 Base configuration

Our final test-case deals with the simulation of complex flows in a turbine cascade of a ORCs cycle. ORCs are Rankine cycles that use properly chosen low-boiling molecularly heavy organic compounds to drive the turbine in place of steam. This makes them suitable for the exploitation of low grade heat sources like biomass combustion, geothermal reservoirs and heat recovery from industrial processes (we refer, e.g. to Angelino and Paliano (1998) and Hung *et al.* (1997) for a complete description of the properties and applications of ORCs and of ORC working fluids). A thermodynamically complex flow in a ORCs turbine cascade is characterized by a significant uncertainty on the physical parameters and on the operating conditions at the turbine inlet (Congedo *et al.*, 2011a). Indeed, the ORCs are mainly used in biomass and geothermal applications where the renewable heat sources display a non-negligible level of variability. Besides, the thermophysical properties of the working fluids are themselves characterized by a strong uncertainty (Colonna *et al.*, 2006; Guardone *et al.*, 2004). When designing a turbine specifically adapted to ORCs cycles, a meaningful numerical prediction of the performance must take into account these uncertainties on the thermophysical properties but also on the inlet boundary conditions. In Congedo *et al.* (2011b), some efficient procedures to perform shape optimization in a 2D complex flow with multiple-source uncertainties (thermodynamic model, operating conditions and geometry) have been presented.

In the present work, the turbine blade under consideration is the two dimensional VKI LS-59 cascade, a configuration which has been widely studied (Congedo *et al.*, 2011a; Kiock *et al.*, 1986). An unstructured computational fluid-dynamics (CFD) solver (that solves the Euler equations) is used to ensure the reliability of the computed results for dense gas flows through a turbine cascade (Congedo *et al.*, 2011a). The two-dimensional flow domain is discretized by a structured C-grid comprised of 192 × 16 cells. The boundary conditions are imposed as follows: at the inlet and outlet boundaries, non-reflecting boundaries are applied using the method of characteristics; a slip condition is imposed at the wall, which uses multi-dimensional linear extrapolation from interior points to calculate the wall pressure; periodicity conditions are prescribed at the inter-blade passage boundaries. The Euler equations are completed with the Peng-Robinson (PRSV) equation of state for taking into account complex effects. It is defined as follows:

$$p = \frac{RT}{v-b} - \frac{a}{v^2 + 2bv - b^2}. \tag{24}$$

where p, T and v denote, respectively, the fluid pressure, the fluid temperature and its specific volume, a and b are substance specific parameters related to the fluid critical-point properties. This model is completed with a model describing the caloric behavior of the fluid, approximated through a power law for the isochoric specific heat in the ideal gas limit, defined as follows $c_{v\infty}(T) = c_{v\infty}(T_c) \times (T/T_c)^n$. Globally, PRSV depends on the following parameters, the fluid acentric factor ω (a and b depend on this parameter), the isobaric specific heat in the ideal gas state at the critical temperature T_c, i.e. $c_{v\infty}(T_c)$, and a fluid-dependent parameter n.

The siloxane dodecamethylcyclohexasiloxane ($C_{12}H_{36}Si_6O_6$), commercially known as D_6, is the fluid considered in this study. The physical properties of D_6 are reported in Table II, while PRSV coefficients for this fluid are reported in Table III.

Performance of the turbine cascade, that can be computed as a result of the CFD simulation, can be evaluated by using several output criteria. Here, the power output per unit depth (PO) expressed as $\Delta h \cdot \dot{m}/w_{mol}$ (W) is taken into account, where Δh is the enthalpy variation through turbine stage, $\dot{m}$ is the mass flow rate and w_{mol} is the molecular weight.

5.2 Sources of uncertainties

Three main sources of uncertainties are considered in this study (globally eight uncertainties):

1. the uncertainties on the turbine inlet conditions, i.e. inlet total temperature, T_{in}/T_c, inlet total pressure, p_{in}/p_c, angle of incidence β and the stagger angle θ;
2. the uncertainties on the thermodynamic model, i.e. ω, $c_{v\infty}$ and n; and
3. uncertainties on turbine geometrical parameters, i.e. the blade thickness ϕ.

Table II. Thermodynamic data for D6, where M is the percentage molecular weight and T_b is the boiling temperature at 1 atm

M (g/mole)	T_c (K)	P_c (kPa)	T_b (K)
444.9	645.8	961	518.1

Source: Properties are taken from Guardone *et al.* (2004)

Table III. Thermodynamic constants for D6, PRSV equation of state, mean and min./max. values using an uniform probability density function

	n	$c_{v\infty}$	ω
Mean	0.5729	105.86	0.7361
Range	0.5385-0.6073	99.50-112.20	0.7214-0.7508

Source: Data taken from Cinnella *et al.* (2011)

Since no experimental estimation are available, we assume systematically uniform probability density functions.

Basing on Colonna *et al.* (2008), the 3.0 percent of uncertainty for the temperature and pressure at the inlet conditions are taken into account. The PRSV thermodynamic model is considered as a good trade-off between the accuracy of thermodynamic properties and the functional complexity since it depends on a limited number of parameters, hence a reduced number of uncertainty sources (Cinnella *et al.*, 2011). The following uncertainties are retained for this model (see Table III and Cinnella *et al.* (2011) and Colonna *et al.* (2008)), listed with their associated error bars: the acentric factor ω (2 percent), the isobaric specific heat in the ideal gas state $c_{v\infty}$ (6 percent) and a fluid-dependent parameter n (6 percent). For the other parameters, it is assumed an uncertainty of 3 percent for the angle of incidence β and the stagger angle θ, and an uncertainty of 2 percent for the thickness ϕ (varying from 0.98 to 1.02 with the mean equal to 1.0).

5.3 Problem definition

Optimization problem is defined as follows:

$$\max_{\frac{T_{in}}{T_c},\frac{p_{in}}{p_c},\beta,\vartheta} |\mu(PO)| \quad and \quad \min_{\frac{T_{in}}{T_c},\frac{p_{in}}{p_c},\beta,\vartheta} |\sigma^2(PO)|,$$

i.e. to find the optimal values for T_{in}/T_c, p_{in}/p_c, β and θ (four design variables) in order to maximize the mean of power output, $\mu(PO)$, and to minimize its standard deviation, $\sigma(PO)$ (two objective-optimization problem). Ranges for each design variable are defined in Table IV. Remark that the lower limit for the temperature is given by the saturation curve limit (SCL). Seeing that CFD code can compute only one-phase flows, it has to be verified that the uncertainty region does not cross the maximal saturation curve (that can be computed as the upper limit of the 100 percent confidence intervals when uncertainties on thermodynamic model are taken into account).

Finally, the optimization problem consists in finding the optimal values for four design variables where the output to maximize is dependent from eight uncertainties.

5.4 ANOVA decomposition over the geometric plan and construction of Kriging response surface

The algorithm described in Figure 1(a) is applied on the problem defined in Section 5.3.

First, an initial DOE of 50 samples (y_l with $1=1,\ldots,50$) in the design space constituted by the four design variables, i.e. T_{in}/T_c, p_{in}/p_c, β and θ, is generated. Then, equation (2) is solved for each design y_l, considering a stochastic space constituted by the eight uncertainties defined in the previous section. A quasi-Monte Carlo plan (based on Sobol sequences) of 200 individuals in the stochastic space is considered, and the PCE is used to compute $TSI_j(y_l)$ for each uncertainty. The convergence of TSI indexes for each uncertainty and design is verified by increasing the number of individuals

Table IV. Ranges of design variables in the optimization plan

p_{in}/p_c	T_{in}/T_c	β	ϑ
0.7-0.98	SCL-1.15	25°-35°	29°-39°

to 500. Remark that, during this stage, the convergence of TSI is more important than the convergence of the variance since only the correct assessment of the response surface is necessary to perform the second step of the algorithm. Finally, using the set of $TSI_j(y_l)$, a response surface in the design space for each uncertainty j, $T\tilde{S}I_j(y)$, is built using a Kriging method.

In Figures 2 and 3, $T\tilde{S}I_j(y)$ contours are reported for each uncertainty j, in the plan p-T, where a point in the plan p-T is associated to the design y, characterized by a couple (p_{in}, T_{in}) of inlet thermodynamic conditions. As shown in Figure 2(a) and (b), $T\tilde{S}I_j(y)$ associated to the uncertainty on p_{in} varies from 8 to 44 percent while varies from 39 to 83 percent for the uncertainty on T_{in}. Concerning the uncertainties on two geometrical parameters, θ and ϕ (Figure 2(c) and (d)), $T\tilde{S}I_j(y)$ varies from 7 to 25 percent and from 0.7 to 2.9 percent, respectively. $T\tilde{S}I_j(y)$ associated to the uncertainties on the thermodynamic model, i.e. ω, $c_{v\infty}$ and n (Figure 3), and on the geometrical parameter ϕ, are lower than 0.29 percent, then they could be discarded using the 2 percent criterion (Section 2). The influence of the uncertainty on the

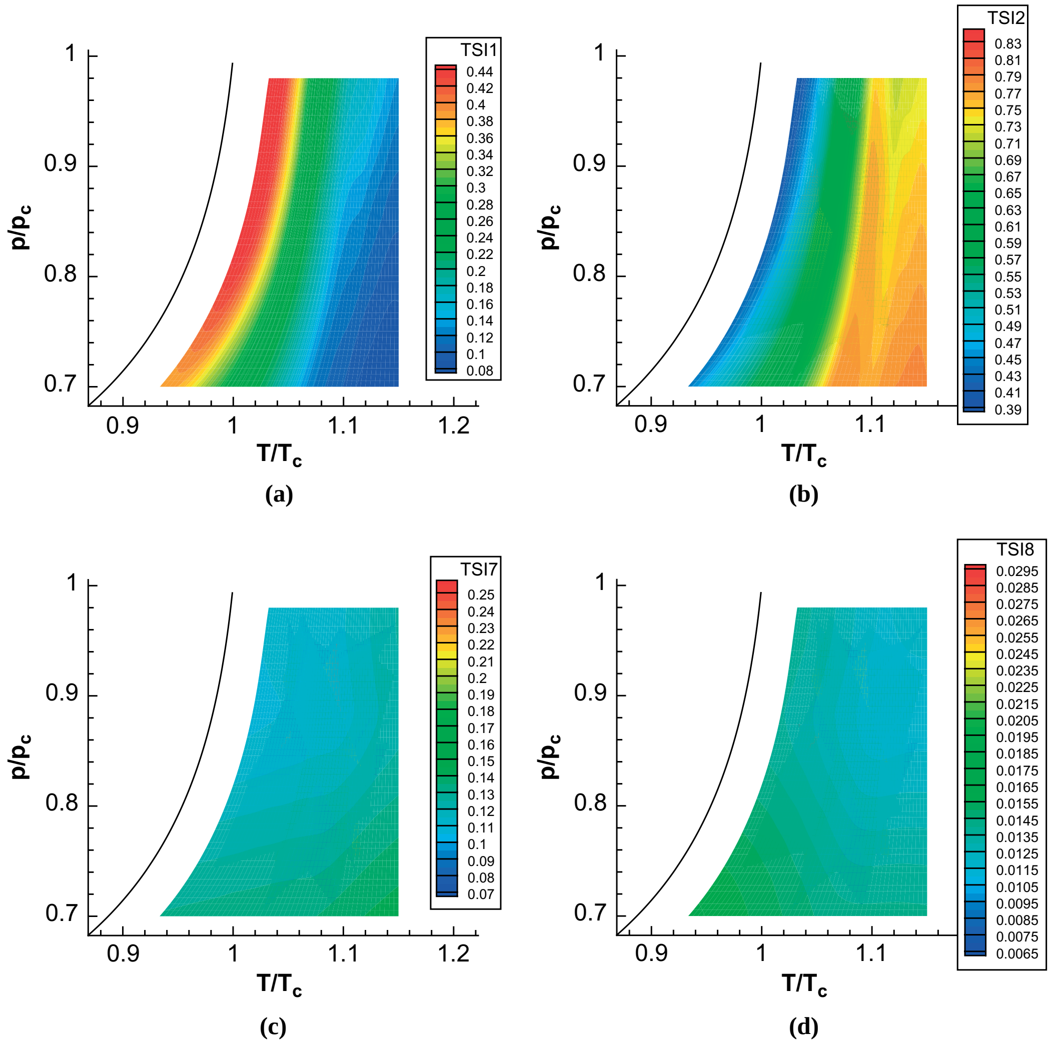


Figure 2.
$T\tilde{S}I_j(y)$ contours in the plan p-T for p_{in} (a), T_{in} (b), θ (c), ϕ (d)

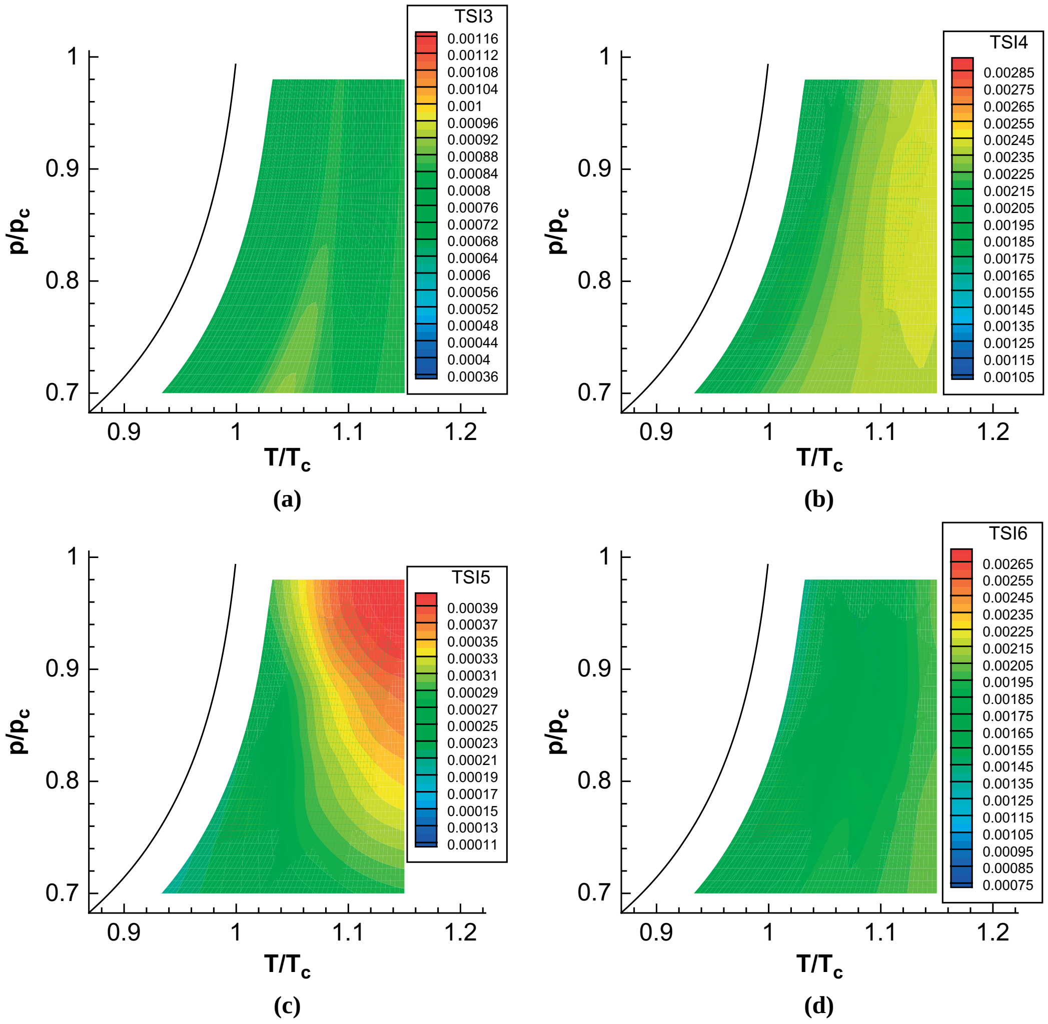


Figure 3.
$T\tilde{S}I_j(y)$ contours in the plan p-T for $c_{v\infty}$ (a), n (b), ω (c), β (d)

thermodynamic model is limited with respect to that on the inlet thermodynamic conditions. Note this hierarchy is likely to depend on the choice of equation of state: the PRSV model has been found in Congedo *et al.* (2011b) to be less sensitive than other models to uncertainties on its parameters and the present conclusion is consistent with these previous findings.

The response surface $T\tilde{S}I_j(y)$ and the 2 percent criterion are used in the second step of the algorithm as explained in the following paragraph.

5.5 Optimization

At this stage, the second step of the algorithm described in Section 3 (Figure 1(b)) is applied.

The genetic algorithm is initialized with the same DOE of 50 samples (y_l with $l = 1, \ldots, 50$), generated during the first step. At each iteration n, the response surface $T\tilde{S}I_j(y_n)$ and the 2 percent criterion are used to define equation (3), i.e. ξ_r, for the design y_n. The stochastic problem, expressed in equation (3), is solved by means of PC expansion, and $\mu(PO)$ and $\sigma(PO)$ are computed. For reaching convergence, 20 designs y evolved during 40 generations.

The converged Pareto front is shown in Figure 4. Various configurations are obtained with a large variation of the PO, going from 0.91 to 1.46.

Four individuals are extracted from the Pareto front in order to evaluate differences in the solution: one individual at the lowest variance (denoted hereafter LV), one at the largest mean (denoted HM), and two others, denoted BT1 and BT2, representing potential trade-off between mean and standard deviation.

In Figure 5, the mean dimensionless pressure (normalized with respect to the critical pressure) is shown in the computational domain for LV, HM, BT1 and BT2. Remark that high inlet turbine pressure are associated to high mean of PO, displaying a strong dependence of turbine performances from thermodynamic inlet conditions. In a similar way, standard deviation of the dimensionless pressure is reported in Figure 6. Standard deviation is higher around the compression shock location near the trailing edge. Moreover, the standard deviation of PO seems related to the peak of maximal standard deviation of the pressure, i.e. when the maximal standard deviation is lower, standard deviation of PO is lower too.

5.6 A posteriori *validation and computational cost reduction*

Finally, statistics of the optimal designs, LV, HM, BT1 and BT2, are computed by considering the whole set of uncertainties, i.e. performing a complete stochastic computation (equation (2)) without uncertainty reduction. The interest is twofold, i.e. to verify that:

(1) the reduced problem statistics of the optimal individuals are well computed with respect to the complete stochastic problem; and

(2) LV, HM, BT1 and BT2 designs still belong to the Pareto front when statistics are computed with a greater accuracy.

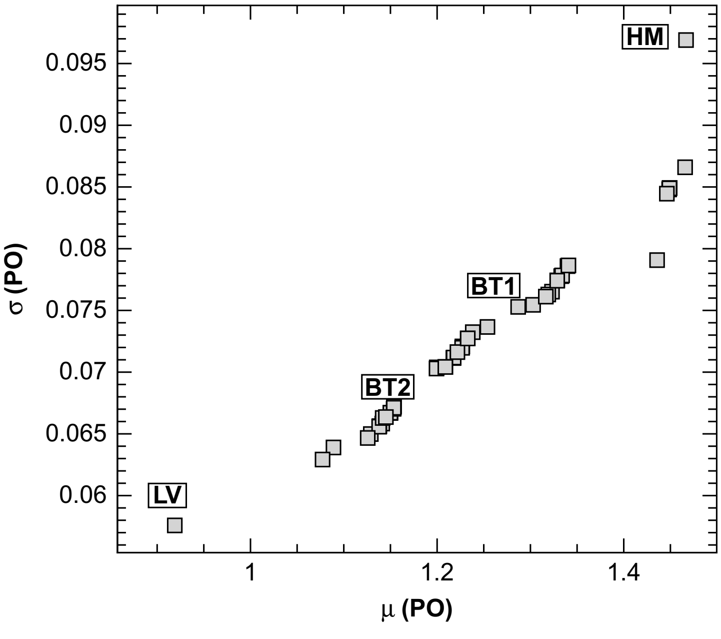


Figure 4.
Pareto front in the plan $\mu(PO)[W] - \sigma(PO)[W]$

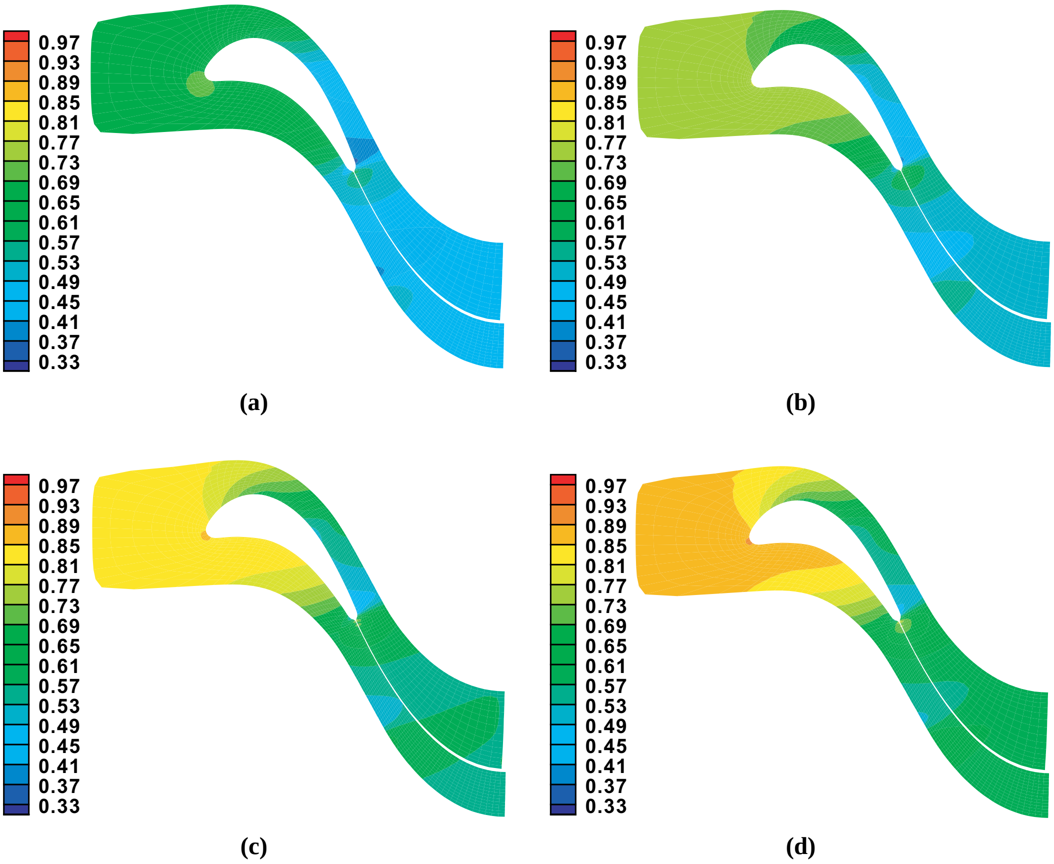


Figure 5. Mean of p/p_c for LV (a), BT2 (b), BT1 (c), HM (d)

In Figure 7, the Pareto front constituted by LV, HM, BT1 and BT2 designs are reported, where statistics are computed by means of the reduced (grey square) and the complete stochastic problem (circle). As shown in Figure 7, these four designs still belong to the Pareto front even if statistics are evaluated by taking into account all the uncertainties. This represents a validation of the proposed algorithm. Moreover, the relative error of the mean and standard deviation are lower than 0.5 percent, that confirms the efficacy of the reduction strategy. In Figure 8, the coefficient of variation (ratio of the standard deviation to the mean) for the dimensionless pressure (normalized with respect to the critical pressure) is computed for the LV design by means of the complete and reduced stochastic problem in order to display the similarity of the two solutions (a relative error always lower than 0.2 percent is found).

Now, let us focus on the savings in terms of computational cost by comparing the proposed algorithm and the complete stochastic problem (using the whole set of uncertainties). Considering a whatever design y obtained during the optimization, the number of uncertainties of the reduced problem, i.e. r (Section 3.1), varies from 3 to 4, where 8 is the global number of uncertainties, as it can be clearly seen in Figures 2 and 3. Remark that the polynomial expansion order can be different between two designs, then only equation (21) (Section 3.1) can be used to compute an estimation of the savings in terms of computational cost. The reduction is around 87 percent with

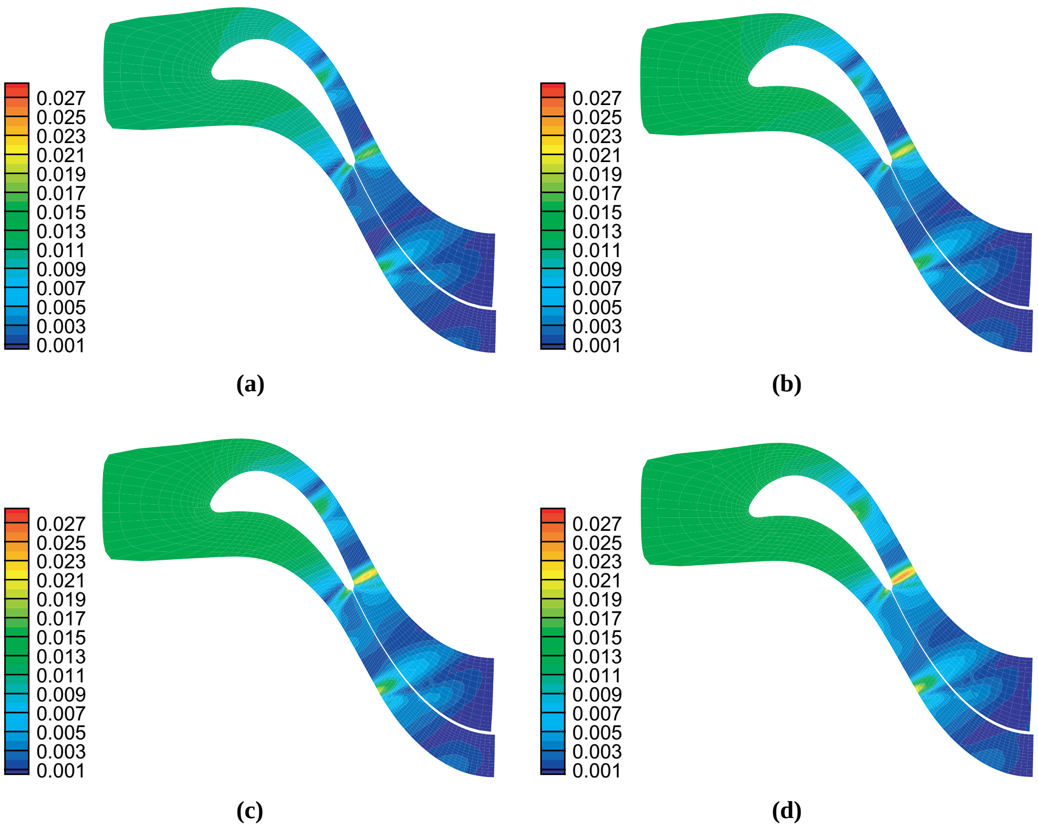
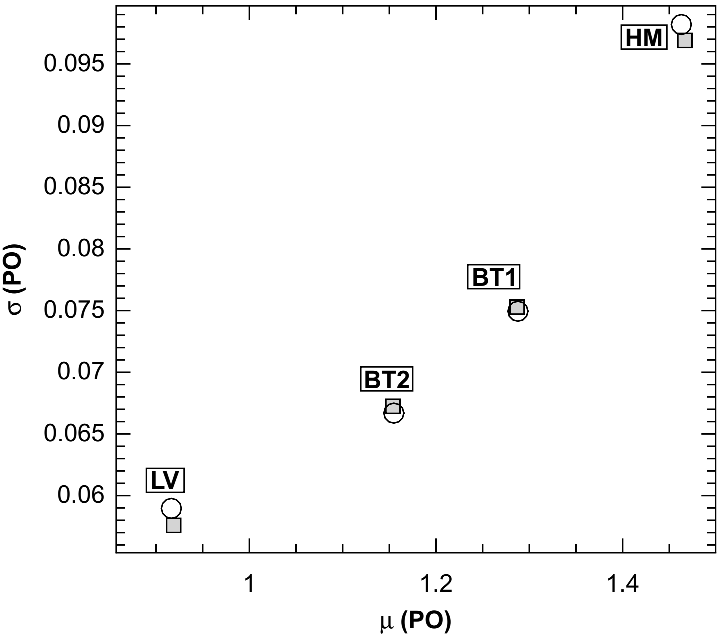


Figure 6.
Standard deviation of p/p_c for LV (a), BT2 (b), BT1 (c), HM (d)



Notes: Grey square – reduced; circle – complete

Figure 7.
Pareto front in the plan $\mu(PO)[W] - \sigma(PO)[W]$

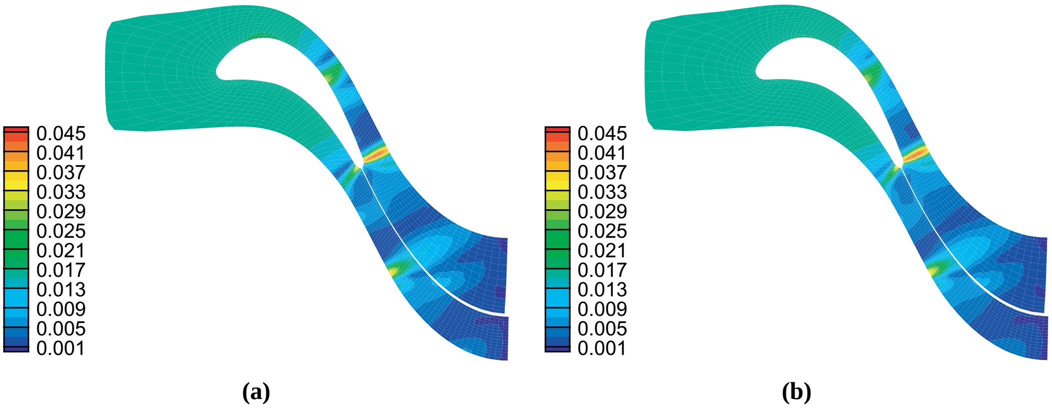


Figure 8.
Coefficient of variation of p/p_c for LV design, reduced (a) and complete (b)

respect to the complete robust optimization, where the global number of deterministic simulations is nearly equal to 650,000 (5 millions) for the proposed approach (the complete robust optimization). Then, the proposed approach is very promising also in realistic configurations, permitting a strong reduction of the computational cost by preserving nearly the same order of accuracy.

6. Conclusions

In this work, a stochastic optimization method is developed in order to efficiently perform optimization in the presence of uncertainties. The idea is to reduce the number of dimensions in the stochastic problem associated to a given design. ANOVA analysis is used to perform a variance-based decomposition and to compute the TSIs for each uncertainty and an initial set of designs. Then, a response surface is generated for each TSI in the design space, that is used during the optimization loop. In this way, the uncertainties with a TSI lower than 2 percent (TSI criterion) can be discarded in the reduced stochastic problem associated to a whatever design. Through an experimental campaign on PDE, the error on the variance is roughly estimated by the quantity $(1-\bar{T}_{TSI})\times 100$, where $\bar{T}_{TSI}$ is the summation of TSI concerning the variables respecting the TSI criterion. During the optimization, the stochastic problem associated to a given design is reduced, thus decreasing the cost of the statistics estimation. This method is general and can be used with a deterministic black box solver.

The optimization method is successfully tested on two problems in fluid mechanics: a 1D compressible nozzle flow, and a thermodynamically complex flow in a turbine cascade.

With this technique, a computational gain of the order of 10-90 is obtained, for problems with some predominant uncertainties with respect to a full PCE. This approach can be further improved by employing an algorithm for the reduction of the number of points required for the quadrature, i.e. a sparse grid technique or an adaptive algorithm. For very high-dimensional problems, this strategy can be easily applied, provided that convergence on sensitivity indexes is attained. In this case, computational cost for computing sensitivity indexes is similar to that one for computing the various statistical moments. As a consequence, this strategy does not require an adding cost with respect to more classical techniques, thus it is expected to work well also for very high-dimensional problems.

In future work, the aim is to make the optimization process more robust by improving the construction of the TSI response surface. The plan is also to extend the algorithm to include high-order decomposition.

Note

1. Remark that each TSI_j is computed only on the complete problem. The total amount of TSI of the reduced problem is a measure of the interactions that a reduced function can capture with respect to the complete one.

References

Abgrall, R., Congedo, P.M. and Geraci, G. (2012), "Numerical investigation on the total sensitivity index influence in the solution of stochastic partial differential equations", INRIA Research Report, RR-7911.

Agarwal, N. and Aluru, N.R. (2009), "A domain adaptive stochastic collocation approach for analysis of MEMS under uncertainties", *Journal of Computational Physics*, Vol. 228 No. 20, pp. 7662-7688.

Angelino, G. and Paliano, P. (1998), "Multicomponent working fluids for organics Rankine cycles", *Energy*, Vol. 23, pp. 449-463.

Askey, R. and Wilson, J. (1985), "Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials", *Memoirs of the American Mathematical Society*, Vol. 54 No. 319.

Baglietto, M., Cervellera, C., Sanguineti, M. and Zoppoli, R. (2010), "Management of water resource systems in the presence of uncertainties by nonlinear approximation techniques and deterministic sampling", *Computational Optimization and Application*, Vol. 47, pp. 349-376.

Bellman, R.E. and Richard, B. (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ.

Blatman, G. and Sudret, B. (2011), "Adaptive sparse polynomial chaos expansion based on least angle regression", *Journal of Computational Physics*, Vol. 230 No. 6, pp. 2345-2367.

Caflisch, R.E., Morokoff, W. and Owen, A.B. (1997), "Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension", *Journal of Computational Finance*, Vol. 1, pp. 26-47.

Cao, Y., Hussaini, M. and Zang, T. (2003), "An efficient Monte Carlo method for optimal control problems with uncertainty", *Computational Optimization and Application*, Vol. 26, pp. 219-230.

Cinnella, P., Congedo, P.M., Pediroda, V. and Parussini, L. (2011), "Sensitivity analysis of dense gas flow simulations to thermodynamic uncertainties", *Physics of Fluids*, Vol. 23 No. 116101, pp. 1-20.

Colonna, P., Harinck, J., Rebay, S. and Guardone, A. (2008), "Real-gas effects in organic Rankine cycle turbine nozzles", *AIAA J. Prop. Power*, Vol. 24 No. 2, pp. 282-294.

Colonna, P., Nannan, N., Guardone, A. and Lemmon, E. (2006), "Multiparameter equations of state for selected siloxanes", *Fluid Phase Equilibria*, Vol. 244 No. 2, pp. 193-211.

Congedo, P., Corre, C. and Cinnella, P. (2011a), "Numerical investigation of dense-gas effects in turbomachinery", *Computers & Fluids*, Vol. 49, pp. 290-301.

Congedo, P., Corre, C. and Martinez, J.M. (2011b), "Shape optimization of an airfoil in a BZT flow with multiple-source uncertainties", *Computer Methods in Applied Mechanics and Engineering*, Vol. 200 Nos 1-4, pp. 216-232.

Congedo, P., Geraci, G., Abgrall, R., Pediroda, V. and Parussini, L. (2011c), "Efficient ANOVA decomposition and metamodels-based multi-objective robust optimization", in Poloni, C., Quagliarella, D., Periaux, J., Gauger, N. and Giannakoglu, K. (Eds), *Evolutionary and Deterministic Methods for Design, Optimization and Control, Capua, 14-16 September*, June, pp. 554-569.

Crestaux, T., Le Maître, O. and Martinez, J.M. (2009), "Polynomial chaos expansion for sensitivity analysis", *Reliability Engineering & System Safety*, Vol. 94 No. 7, pp. 1161-1172.

Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2002), "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Transactions on Evolutionary Computation*, Vol. 6 No. 2.

Foo, J., Wan, X. and Karniadakis, G.E. (2008), "The multi-element probabilistic collocation method (ME-PCM): error analysis and applications", *Journal of Computational Physics*, Vol. 227 No. 22, pp. 9572-9595.

Gao, Z. and Hesthaven, J. (2010), "Efficient solution of ordinary differential equations with high-dimensional parametrized uncertainty", *Communications in Computational Physics*, Vol. 228, pp. 1-33.

Guardone, A., Vigevano, L. and Argrow, B. (2004), "Assessment of thermodynamic models for dense gas dynamics", *Physics of Fluids*, Vol. 16, pp. 3878-3887.

Hung, T., Shai, T. and Wang, S. (1997), "A review of organic Rankine cycles (ORCs) for the recovery of low-grade waste heat", *Energy*, Vol. 22 No. 7, pp. 661-667.

Jensen, H.A. and Catalan, M.A. (2007), "On the effects of non-linear elements in the reliability-based optimal design of stochastic dynamical systems", *Int. J. Non-Linear Mech.*, Vol. 42 No. 5, pp. 802-816.

Kiock, R., Lehthaus, F., Baines, N.C. and Sieverding, C.H. (1986), "The transonic flow through a plane turbine cascade as measured in four European wind tunnels", *ASME J. Eng. Gas Turb. Power*, Vol. 108 No. 2, pp. 277-284.

Le Maître, O. (2005), "Methodes spectrales pour la propagation d'incertitudes parametriques dans le modele numerique", PhD thesis.

Ma, X. and Zabaras, N. (2010), "An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations", *Journal of Computational Physics*, Vol. 229 No. 10, pp. 3884-3915.

Namura, N., Shimoyama, K., Jeong, S. and Obayashi, S. (2011), "Kriging/RBF-hybrid response surface method for highly nonlinear functions", *IEEE Congress on Evolutionary Computation*, pp. 2534-2541.

Novak, E. and Ritter, K. (1996), "High dimensional integration of smooth functions over cubes", *Numerische Mathematik*, Vol. 75, pp. 79-97.

Sankaran, S., Audet, C. and Marsden, A.L. (2010), "A method for stochastic constrained optimization using derivative-free surrogate pattern search and collocation", *Journal of Computational Physics*, Vol. 229, pp. 4664-4682.

Schuëller, G.I. and Jensen, H.A. (2008), "Computational methods in optimization considering uncertainties – an overview", *Comput. Methods Appl. Mech. Eng.*, Vol. 198, pp. 2-13.

Sobol, I. (2001), "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates", *Mathematics and Computers in Simulation*, Vol. 55 Nos 1-3, pp. 271-280.

Taguchi, G. (1989), *Introduction to Quality Engineering*, American Supplier Institute, Bingham Farms, MI.

Verstraete, T. and Périaux, J. (Eds) (2010), *Introduction to Optimization and Multidisciplinary Design in Aeronautics and Turbomachinery*, Vol. 1, VKI LS 2010-07, Hardcover.

Wang, X. and Sloan, I.H. (2003), "Why are high-dimensional finance problems often of low effective dimension?", *SIAM Journal on Scientific Computing*, Vol. 27, pp. 159-183.

Xu, D. and Albin, S. (2003), "Robust optimization of experimentally derived objective functions", *IIE Trans.*, Vol. 35 No. 9, pp. 793-802.

Further reading

Griebel, M. (2006), "Sparse grids and related approximation schemes for higher dimensional problems", in Pardo, L., Pinkus, A., Suli, E. and Todd, M. (Eds), *Foundations of Computational Mathematics* (*FoCM05), Santander*, Cambridge University Press, Cambridge, pp. 106-161.

Guardone, A., Zamfirescu, C. and Colonna, P. (2010), "Maximum intensity of rarefaction shock waves for dense gases", *Journal of Fluid Mechanics*, Vol. 642, pp. 127-146.

Sacks, J., Welch, W., Michell, T. and Wynn, H. (1989), "Design and analysis of computer experiments", *Statistical Science*, Vol. 4, pp. 409-453.

About the authors

Pietro Marco Congedo is a Research Scientist at INRIA – Bordeaux Sud Ouest, France. He graduated with honours in materials engineering at University of Lecce (Italy). After his Master in fluid mechanics at Arts et Métiers (Paris, France), he received his PhD in energy systems at University of Lecce in 2007. His research interests are in the numerical simulation and optimisation of real gas flows, thermodynamics of complex flows, multi-phase flows, uncertainty quantification and robust optimisation. Pietro Marco Congedo is the corresponding author and can be contacted at: pietro.congedo@inria.fr

Gianluca Geraci is a PhD candidate in applied mathematics at Institut de Mathématiques de Bordeaux, France. He received his Master degree with full marks and honours at Politecnico di Milano (Italy) in aeronautical engineering in 2010. His research interests are in CFD for compressible flows and uncertainty quantification methods.

Rémi Abgrall is a Professor at University of Bordeaux since 1996, and currently in secondment at INRIA. He heads an INRIA research group, Bacchus. His main research topic is on the numerical approximation of compressible fluid dynamics problems, compressible multiphase and interfaces, and more recently uncertainty quantification. He is Editor of the *International Journal on Numerical Methods in Fluids* and Associate Editor of several journals including the *Journal of Computational Physics*, *Mathematics of Computation*, *Journal of Scientific Computing*, *Computers and Fluids*. He is the recipient of an advance research grant of the European Research Council.

Valentino Pediroda is an Assistant Professor at the Department of Mechanical Engineering of the University of Trieste, Italy. After graduating in mechanical engineering at University of Trieste, he received his PhD in numerical methods at University of Udine. His researches are focused on multi-objective optimisation, response surfaces and design under uncertainties.

Lucia Parussini is an Assistant Professor at the Department of Mechanical Engineering of the University of Trieste, Italy. She graduated in mechanical engineering at the University of Trieste. She received her PhD in numerical methods at the University of Udine. Her research field is numerical methods for fluid dynamic problems affected by uncertainties.

Paper *P7*

Elsevier Editorial System(tm) for Reliability Engineering & System Safety
Manuscript Draft

Manuscript Number:

Title: Decomposition and Computation of high-order statistics

Article Type: Research Paper

Keywords: high-order statistics, skewness, kurtosis, UncertaintyQuantification.

Corresponding Author: Dr. Pietro Marco Congedo,

Corresponding Author's Institution: INRIA Bordeaux Sud-Ouest

First Author: Pietro Marco Congedo

Order of Authors: Pietro Marco Congedo; Gianluca Geraci; Rémi Abgrall; Gianluca Iaccarino

Abstract: ANOVA analysis is a very common numerical technique for computing a hierarchy of most important input parameters for a given output when variations are computed in terms of variance. This second central moment can not be retained as an universal criterion for ranking some variables, since a non-gaussian output could require higher order (more than second) statistics for a complete description and analysis.
In this work, we illustrate how third and fourth-order statistic moments, i.e. skewness and kurtosis, respectively, can be decomposed. It is shown that this decomposition is correlated to a Polynomial Chaos (PC) expansion, permitting to easily compute each term. Then, new sensitivity indexes are proposed basing on the computation of skewness and kurtosis.
PC-based numerical technique is used in order to compute the convergence of the sensitivity indexes according to the polynomial order by using the exact solution as the reference one.
Then, a functional decomposition based on variance, skewness and kurtosis is applied on several test-functions, displaying how sensitivity indexes vary according to the order of the statistical moment.
Then, the problem of how reducing the complexity of a stochastic problem is considered. In particular, two strategies are considered, one focused on the reduction of the number of dimensions, the other on the reduction of the order of interaction. The impact on the statistics of the reduced function is then assessed.

Highlights (for review)

Highlights

1) A correlation is found between the functional decomposition and polynomial chaos (PC).

2) PC-based technique is used to compute convergence of high-order sensitivity indexes.

3) Sensitivity indices bases on skewness and kurtosis decomposition are introduced.

4) Importance of ranking uncertainties in terms of high-order moments is demonstrated.

5) Two different model reduction strategies are proposed and investigated.

Decomposition and Computation of high-order statistics

R. Abgrall[a], P.M. Congedo[a,*], G. Geraci[a], G. Iaccarino[b]

[a] INRIA Bordeaux–Sud-Ouest, Equipes Bacchus, 200, Avenue de la Vieille Tour, 33400 Talence, FRANCE
[b] Mechanical Engineering Dept, Stanford University, Bld 500, CA 94305-3035, USA

Abstract

ANOVA analysis is a very common numerical technique for computing a hierarchy of most important input parameters for a given output when variations are computed in terms of variance. This second central moment can not be retained as an universal criterion for ranking some variables, since a non-gaussian output could require higher order (more than second) statistics for a complete description and analysis.

In this work, we illustrate how third and fourth-order statistic moments, *i.e.* skewness and kurtosis, respectively, can be decomposed. It is shown that this decomposition is correlated to a Polynomial Chaos (PC) expansion, permitting to easily compute each term. Then, new sensitivity indexes are proposed basing on the computation of skewness and kurtosis. PC-based numerical technique is used in order to compute the convergence of the sensitivity indexes according to the polynomial order by using the exact solution as the reference one. Then, a functional decomposition based on variance, skewness and kurtosis is applied on several test-functions, displaying how sensitivity indexes vary according to the order of the statistical moment. Then, the problem of how reducing the complexity of a stochastic problem is considered. In particular, two strategies are considered, one focused on the reduction of the number of dimensions, the other on the reduction of the order of interaction. The impact on the statistics of the reduced function is then assessed.

Keywords: high-order statistics, skewness, kurtosis, Uncertainty Quantification.

1. Introduction

Optimization and design in the presence of uncertain operating conditions, material properties and manufacturing tolerances poses a tremendous challenge to the scientific computing community. In many industry-relevant situations the performance metrics depend in a complex, non-linear fashion on those factors and the construction of an accurate representation of this relationship is difficult. Probabilistic uncertainty quantification (UQ) approaches represent the inputs as random variables and seek to construct a statistical characterization of few quantities of interest. Several methodologies are proposed to tackle this issue, most of all focused on stochastic spectral methods [1, 2, 3, 4, 5], that can provide considerable speed-up in computational time when compared to Monte Carlo (MC) simulation. In realistic situations however, the presence of a large number of uncertain inputs leads to an exponential increase of the cost thus making these methodologies unfeasible [6]. This situation becomes even more challenging when robust design optimization is tackled [7, 8].

Several UQ methods have been developed with the objective of reducing the number of solution required to obtain a statistical characterization of the quantity of interest, such as Sparse grid techniques or adaptive mesh generation. These techniques can lead to a dramatical reduction of the quadrature points for moderate dimensional problem, provided that the function has some regularity. Classical sparse grids [9] are constructed from tensor products of one-dimensional quadrature formulas. Some Galerkin-based methods deals with multi-resolution wavelet expansions [10, 11], domain decomposition in the random space [12], adaptive h-refinement [3] for dealing with arbitrary probability distributions.

[*] Corresponding author
Telephone Number: +33(0)5 24 57 40 58, *Email:* pietro.congedo@inria.fr

Among the collocation-based stochastic spectral methods, in [13] they proposed the use of sparse grid quadrature for stochastic collocation. Older studies show the errors and efficiency of sparse grid integration and interpolation [14, 15], Smolyak constructions based on one-dimensional nested Clenshaw-Curtis rules [14, 16] and the integration error of sparse grids based on one-dimensional Kronrod-Patterson rules [17].

An alternative solution for reducing the cost of the UQ method is based on approaches attempting to identify the relative importance of the input uncertainties on the output. If some dimensions could be identified as negligible, they could be discarded in a reduced stochastic problem. If the number of uncertainties could be reduced, a better statistic estimation could be achieved with a lower computational cost.

Concerning the computation of the most influent parameters, it is important to determine the uncertain inputs which have the largest impact on the variability of the model output. In literature, Global sensitivity analysis (GSA) aims at quantifying how uncertainties in the input parameters of a model contribute to the uncertainties in its output (see for example [18]), where global sensitivity analysis techniques are applied to probabilistic safety assessment models). Sometimes, GSA classifies the inputs according to their importance on the output variations and it gives a hierarchy of most important ones.

Traditionally, GSA is performed using methods based on the decomposition of the output variance [19], *i.e.* ANOVA. The ANOVA approach involves splitting a multi-dimensional function into its contributions from different groups of subdimensions. In 2001, Sobol used this formulation to define global sensitivity indices [19], displaying the relative variance contributions of different ANOVA terms. In [20], they introduced two High-Dimensional Model Reduction (HDMR) techniques to capture input-output relationships of physical systems with many input variables. These techniques are based on ANOVA-type decompositions.

Since it requires a large number of function evaluations, several techniques have been developed to compute the different so-called sensitivity indices at low cost [21]. In [22, 23, 24], generalized Polynomial Chaos Expansions (gPC) are used to build surrogate models for computing the Sobol's indices analytically as a post-processing of the PC coefficients. In [6], they combine multi-element polynomial chaos with analysis of variance (ANOVA) functional decomposition to enhance the convergence rate of polynomial chaos in high dimensions and in problems with low stochastic regularity. In [25], the use of adaptive ANOVA decomposition is investigated as an effective dimension-reduction technique in modeling incompressible and compressible flows with high-dimension of random space. In Sudret [26], sparse Polynomial Chaos (PC) expansions are introduced in order to compute sensitivity indices. An adaptive algorithm allows to build a PC-based metamodel that only contains the significant terms whereas the PC coefficients are computed by least-square regression.

Other approaches are developed if the assumption of independence of the input parameters is not valid. New indices have been proposed to address the dependence [27, 28], but this attempts are limited to a linear correlation. In [29], they introduce a global sensitivity indicator which looks at the influence of input uncertainty on the entire output distribution without reference to a specific moment of the output (moment independence) and which can be defined also in the presence of correlations among the parameters. In [30], a gPC methodology to address global sensitivity analysis for this kind of problems is introduced. A moment-independent sensitivity index that suits problems with dependent parameters is reviewed. Recently, in [31], a numerical procedure is set-up for moment-independent sensitivity methods.

The ANOVA-based analysis create a hierarchy of most important input parameters for a given output when variations are computed in terms of variance. A strong limitation of this approach is the fact that it is based on the variance since the second central moment can not be considered like a general indicator for a complete description of output variations. For example, any Gaussian signal is completely characterized by its mean and variance. Consequently the 3rd order moment of a Gaussian signal is zero. Unfortunately, many signals encountered in practice have non-zero high-order statistics, but second-order statistics contain no phase information. As a consequence of this, phase signals cannot be correctly identified by a 2nd-order technique. Remark also that many measurement noises are Gaussian, and so in principle the high-order statistics are less affected by Gaussian background noise than the 2nd order measures. For well describing the complexity of engineering systems, computation of Higher-Order (HO) statistics are of primary importance, for example the third order, the *skewness* (measure of the non-symmetry of the distribution, *i.e.* any symmetric distribution will have a third central moment of zero), and the fourth order, the *kurtosis* (measure of whether the distribution is tall or short, compared to the normal distribution of the same variance). Now, let us imagine to compute the more influential parameters for a given output. The hierarchy of important parameters based on 2nd-order statistical moment (like in ANOVA analysis) is not the same if a different statistic order is considered.

Depending on the problem, a n-order decomposition could be of interest. It seems of primary importance to collect the set of hierarchies obtained from n-order statistical moment decomposition, for a correct ranking of all the uncertainties.

For computing HO statistics, the most diffused methods are related to Monte Carlo and quasi-Monte Carlo approaches. Very few papers exist showing the application of polynomial-chaos techniques to the computation of HO statistics [32, 33].

First objective of this paper is to provide a general method in order to compute the decomposition of high-order statistics, then to formulate an approach similar to ANOVA but for *skewness and kurtosis*. The idea is to compute the most influential parameters not only for the variance but also for higher orders permitting to improve the sensitivity analysis. This is a fundamental step in order to formulate also innovative optimization methods for obtaining very robust designs by taking into account a complete description of the output statistics. Second objective is to illustrate the correlation between the high-order functional decomposition and the PC-based techniques, thus displaying how to compute each term from a numerical point of view. Finally, two reduction strategies are considered for reducing i) the number of dimensions in the stochastic space and ii) the order of interactions. These strategies are tested on some test-cases and their performances evaluated with respect to the complete non-reduced model.

The paper is organized as follows. Section §2 illustrates some definitions for high-order statistics. In section §3, functional decomposition for variance, skewness and kurtosis are presented. In section §4, the correlation between the functional decomposition and a Polynomial Chaos framework is depicted. Section §5 extend some sensitivity indices to high-order decomposition. Then, Section §6 presents several results showing how the Polynomial Chaos expansion can be used practically to compute high-order statistics, and the importance of considering skewness and kurtosis sensitivity indices when ranking a set of uncertainties. In section §7, conclusions and perspectives are drawn.

2. High-order statistics definition

Let us consider a real function $f = f(\boldsymbol{\xi})$ with $\boldsymbol{\xi}$ a vector of random inputs $\boldsymbol{\xi} \in \Xi^d = \Xi_1 \times \cdots \times \Xi_n$ ($\Xi \subset \mathbb{R}^d$) and $\boldsymbol{\xi} \in \Xi^d \longmapsto f(\boldsymbol{\xi}) \in L^2(\Xi^d, p(\boldsymbol{\xi}))$, where $p(\boldsymbol{\xi}) = \prod_{i=1}^{d} p(\xi_i)$ is the probability density function of $\boldsymbol{\xi}$.

The central statistical moment of f of order n can be defined as follows

$$\mu^n(f) = \int_{\Xi^d} (f(\boldsymbol{\xi}) - E(f))^n p(\boldsymbol{\xi}) \mathrm{d}\xi, \tag{1}$$

where $E(f)$ indicates the expected value of f

$$E(f) = \int_{\Xi^d} f(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \mathrm{d}\xi. \tag{2}$$

In the following, we indicate with σ^2, s, and k, the variance (second-order moment), the skewness (third-order), and the kurtosis (fourth-order), respectively. Skewness and kurtosis (see Appendix A for more details) can be also computed as follows

$$\begin{aligned} s &= E(f^3) - 3E(f^2)E(f) + 2E(f)^3 \\ s &= E(f^3) - 3\sigma^2 E(f) - E(f)^3, \end{aligned} \tag{3}$$

$$\begin{aligned} k &= E(f^4) - 4E(f^3)E(f) + 6E(f^2)E(f)^2 - 3E(f)^4 \\ k &= E(f^4) - 4sE(f) - 6\sigma^2 E(f)^2 - E(f)^4. \end{aligned} \tag{4}$$

These expressions are used for the functional decomposition described in the following sections.

3. Functional decomposition

Let us apply the definition of the Sobol functional decomposition [19] to the function f as follows

$$f(\boldsymbol{\xi}) = \sum_{i=0}^{N} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}), \tag{5}$$

where the multi-index $\boldsymbol{m}$, of cardinality $card(\boldsymbol{m}) = d$, can contain only elements equal to 0 or 1. Clearly, the total number of admissible multi-indices $\boldsymbol{m_i}$ is $N + 1 = 2^d$; this number represent the total number of contributes up to the dth-order of the stochastic variables $\boldsymbol{\xi}$. The scalar product between the stochastic vector $\boldsymbol{\xi}$ and $\boldsymbol{m_i}$ is employed to identify the functional dependences of $f_{\boldsymbol{m_i}}$. In this framework, the multi-index $\boldsymbol{m}_0 = (0, \ldots, 0)$, is associated to the mean term $f_{\boldsymbol{m}_0} = \int_{\Xi^d} f(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \mathrm{d}\xi$. As a consequence, $f_{\boldsymbol{m}_0}$ is equal to the expectancy of f, *i.e.* $E(f)$. Let us assume, in the following, to order the N multi-indices $\boldsymbol{m_i}$ in the following way:

$$\begin{aligned}
\boldsymbol{m}_1 &= (1, 0, \ldots, 0)\\
\boldsymbol{m}_2 &= (0, 1, \ldots, 0)\\
&\vdots\\
\boldsymbol{m}_d &= (0, \ldots, 1)\\
\boldsymbol{m}_{d+1} &= (1, 1, 0, \ldots, 0)\\
\boldsymbol{m}_{d+2} &= (1, 0, 1, 0, \ldots, 0)\\
&\vdots\\
\boldsymbol{m}_N &= (1, \ldots, 1).
\end{aligned} \tag{6}$$

Except the term $\boldsymbol{m}_0$, that should be the first in the series, the remaining N multi-indices $\boldsymbol{m_i}$ should be classified with respect to a prescribed criterion. However, this criterion does not affect in any way the successive ANOVA functional decomposition.

The decomposition (5) is of ANOVA-type in the sense of Sobol [19] if all the members in Eq. (5) are orthogonal, *i.e.* as follows

$$\int_{\Xi^d} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) f_{\boldsymbol{m}_j}(\boldsymbol{\xi} \cdot \boldsymbol{m}_j) p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} = 0 \quad \text{with} \quad \boldsymbol{m_i} \neq \boldsymbol{m}_j, \tag{7}$$

and for all the terms $f_{\boldsymbol{m_i}}$, except f_0, it holds

$$\int_{\Xi^d} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) p(\xi_j) \mathrm{d}\xi_j = 0 \quad \text{with} \quad \xi_j \in (\boldsymbol{\xi} \cdot \boldsymbol{m_i})\,. \tag{8}$$

Each term $f_{\boldsymbol{m_i}}$ of (5) can be expressed as follows

$$f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) = \int_{\Xi^{d-\mathrm{card}(\hat{\boldsymbol{m}}_i)}} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) p(\bar{\boldsymbol{\xi}}_i) \mathrm{d}\bar{\boldsymbol{\xi}}_i \quad - \sum_{\substack{\boldsymbol{m}_j \neq \boldsymbol{m_i} \\ \mathrm{card}(\hat{\boldsymbol{m}}_j) < \mathrm{card}(\boldsymbol{m_i})}} f_{\boldsymbol{m}_j}(\boldsymbol{\xi} \cdot \boldsymbol{m}_j), \tag{9}$$

where the multi-indexes $\hat{\boldsymbol{m}}_{\boldsymbol{i}}$, have a cardinality equal to the number of non-null elements in $\boldsymbol{m_i}$ and $\bar{\boldsymbol{\xi}}_i$ contains all the variables not contained in $(\boldsymbol{\xi} \cdot \boldsymbol{m_i})$, *i.e.* $(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) \cup \bar{\boldsymbol{\xi}}_i = \boldsymbol{\xi}$.

Hereinafter in order to substantially reduce the complexity of the notation, the integrals are written with respect to their probability measure (relative to the multi-index $\boldsymbol{m_i}$):

$$\mathrm{d}\mu_i = p(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) \tag{10}$$

The functional decomposition (5) is usually employed [19] to compute the contribution of each term to the overall variance, as shown in the next section.

3.1. Variance decomposition

ANOVA analysis is based on the variance decomposition in its conditional contributions. Variance can be written in terms of conditional expectancy of f and f^2 as (see Appendix A for more details):

$$\sigma^2 = E(f^2) - E(f)^2. \tag{11}$$

As a consequence, the problem is to compute $E(f^2)$, seeing that $E(f)$ is known and equal to $f_{\boldsymbol{m}_0}$. Starting from Eq. (5), it is easy to compute

$$f^2(\boldsymbol{\xi}) = \sum_{i=0}^{N} f^2_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}) + 2\sum_{i=0}^{N}\sum_{j=i+1}^{N} f_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}) f_{\boldsymbol{m}_j}(\boldsymbol{\xi}\cdot\boldsymbol{m}_j). \tag{12}$$

If the equation (12) is integrated in the stochastic space and the orthogonality property (7) is applied, variance can be decomposed as

$$\sigma^2 = \sum_{i=1}^{N}\int_{\Xi^d} f^2_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \sum_{i=1}^{N}\int_{\hat{\Xi}_i} f^2_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})p(\boldsymbol{\xi}\cdot\boldsymbol{m_i})\mathrm{d}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}), \tag{13}$$

where the symbol $\hat{\Xi}_i$ is employed to indicate $\Xi^{\mathrm{card}(\hat{\boldsymbol{m}}_i)}$ for brevity.

Variance can be expressed as the summation of all the conditional contributions

$$\sigma^2 = \sum_{i=1}^{N}\sigma^2_{\boldsymbol{m_i}}. \tag{14}$$

So, a comparison with the equation (13) shows that

$$\sigma^2_{\boldsymbol{m_i}} = \int_{\hat{\Xi}_i} f^2_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})p(\boldsymbol{\xi}\cdot\boldsymbol{m_i})\mathrm{d}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}). \tag{15}$$

Then, the same type of analysis is applied to skewness and kurtosis.

3.2. Skewness decomposition in conditional terms

In this section, the same procedure already presented in the previous section for the variance, is extended to the computation of the skewness. In this case, the major drawbacks is the presence of an higher number of terms to compute with respect to the variance case. In the case of the variance, due to the properties of the ANOVA terms, all the mixed contributions are zero due to orthogonality. This is not the case of the mixed contribution for the skewness. However some terms can be identified as orthogonal, as well as the case of the variance reducing the overall number of terms to compute.

The first step in order to obtain the skewness in terms of the ANOVA components of the function $f(\boldsymbol{\xi})$ is to raise the ANOVA functional decomposition of the function $f(\boldsymbol{\xi})$ to the third power by employing the multinomial theorem as follows

$$\begin{aligned} s = (f(\boldsymbol{\xi}) - f_0)^3 &= \left(\sum_{i=1}^{N} f_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})\right)^3 \\ &= \sum_{i=1}^{N}\int_{\hat{\Xi}_i} f^3_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})\mathrm{d}\mu_i + 3\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\int_{\hat{\Xi}_{ij}} f^2_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}) f_{\boldsymbol{m}_j}(\boldsymbol{\xi}\cdot\boldsymbol{m_j})\mathrm{d}\mu_{ij} \\ &+ 6\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k=j+1}^{N}\int_{\hat{\Xi}_{ijk}} f_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i}) f_{\boldsymbol{m}_j}(\boldsymbol{\xi}\cdot\boldsymbol{m_j}) f_{\boldsymbol{m}_k}(\boldsymbol{\xi}\cdot\boldsymbol{m}_k)\mathrm{d}\mu_{ijk}, \end{aligned} \tag{16}$$

where $\hat{\Xi}_{ij} = \Xi^{\mathrm{card}(\hat{\boldsymbol{m}}_{ij})}$ and $\hat{\Xi}_{ijk} = \Xi^{\mathrm{card}(\hat{\boldsymbol{m}}_{ijk})}$. In the following, the notation is simplified by omitting the explicit dependence of the function $f_{\boldsymbol{m_i}}$ with respect to its coordinates, *i.e.* $f_{\boldsymbol{m_i}} = f_{\boldsymbol{m_i}}(\boldsymbol{\xi}\cdot\boldsymbol{m_i})$.

Here, a special notation is introduced in order to compute multi-indexes as $\boldsymbol{m}_{ab\cdots z}$ as follows

$$\boldsymbol{m_{ab\cdots z}} = \boldsymbol{m_a} \boxplus \boldsymbol{m_b} \boxplus \cdots \boxplus \boldsymbol{m_z} = \left(\frac{m_{a_1} + m_{b_1} + \cdots m_{z_1}}{\left\| m_{a_1} + m_{b_1} + \cdots + m_{z_1} \right\|_{\neq 0}}, \ldots, \frac{m_{a_d} + m_{b_d} + \cdots + m_{z_d}}{\left\| m_{a_d} + m_{b_d} + \cdots + m_{z_d} \right\|_{\neq 0}} \right), \tag{17}$$

where the norm $\left\| \cdot \right\|_{\neq 0}$ is defined as

$$\left\| \alpha \right\|_{\neq 0} = \begin{cases} |\alpha| & \text{if} \quad \alpha \neq 0 \\ 1 & \text{if} \quad \alpha = 0. \end{cases} \tag{18}$$

The expression presented in (16) includes some terms always equal to zero due to the orthogonality of the ANOVA functional components. In particular, a more compact final expression can be obtained as:

$$s = \sum_{p=1}^{N} \int_{\hat{\Xi}_p} f^3_{\boldsymbol{m_p}} \mathrm{d}\mu_p + 3 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_q} \subset \boldsymbol{m_p}} \int_{\hat{\Xi}_{pq}} f^2_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{\substack{r=q+1 \\ \boldsymbol{m_{pq}} = \boldsymbol{m_r}}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pq}. \tag{19}$$

In the Appendix B, it is illustrated how obtaining equation (19) starting from (16). One of the most important contribution of this kind of approach is the possibility to identify the conditional terms related to each single variable or group of variables as expressed for the variance by means of relation (15). In the case of skewness, the conditional terms have a more complex expression (except the first order terms, *i.e.* the terms related to the single variables). This complexity arises from the presence of mixed contribution. For obtaining an additional form of the kind

$$s = \sum_{i=1}^{N} s_{\boldsymbol{m_i}}, \tag{20}$$

it is mandatory to identify all the set of indexes whose interactions become part of an assigned multi-index $\boldsymbol{m_i}$.

Considering that to each multi-index $\boldsymbol{m_i}$ is associated a set of $2^{|\boldsymbol{m_i}|} - 1$ sub-interactions and denoting this set as $\mathcal{P}_i$ (for instance if $\boldsymbol{m_i} = (1, 1)$ then the set $\mathcal{P}_i = \{(1, 0), (0, 1), (1, 1)\}$) holds, from the equation (19) it is possible to identify each contribution as follows

$$s_{\boldsymbol{m_i}} = \int_{\hat{\Xi}_i} f^3_{\boldsymbol{m_i}} \mathrm{d}\mu_i + 3 \int_{\hat{\Xi}_i} f^2_{\boldsymbol{m_i}} \sum_{\boldsymbol{m_q} \in \mathcal{P}_{i,\neq}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i + 6 \sum_{\boldsymbol{m_p} \in \mathcal{P}_{i,\neq}} \sum_{\substack{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_{i,\neq} \\ \boldsymbol{m_{pq}} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i. \tag{21}$$

Note that the equation (21) is explicitly obtained in the Appendix B.

3.2.1. Kurtosis decomposition in conditional term

In this section, how decomposing the kurtosis is described. The functional decomposition based on the functional Sobol form (Eq. (5)), after the application of the multinomial theorem, is equal to

$$
\begin{aligned}
k &= (f(\boldsymbol{\xi}) - f_0)^4 = \left(\sum_{i=1}^{N} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) \right)^4 \\
&= \sum_{i=1}^{N} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}}^4 (\boldsymbol{\xi} \cdot \boldsymbol{m_i}) p(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) + 4 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \int_{\hat{\Xi}_{ij}} f_{\boldsymbol{m_i}}^3 (\boldsymbol{\xi} \cdot \boldsymbol{m_i}) f_{\boldsymbol{m_j}}(\boldsymbol{\xi} \cdot \boldsymbol{m_j}) p(\boldsymbol{\xi} \cdot \boldsymbol{m_{ij}}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_{ij}}) \\
&+ 6 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \int_{\hat{\Xi}_{ij}} f_{\boldsymbol{m_i}}^2 (\boldsymbol{\xi} \cdot \boldsymbol{m_i}) f_{\boldsymbol{m_j}}^2 (\boldsymbol{\xi} \cdot \boldsymbol{m_j}) p(\boldsymbol{\xi} \cdot \boldsymbol{m_{ij}}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_{ij}}) \qquad (22) \\
&+ 12 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\substack{k=j+1 \\ k \neq i}}^{N} \int_{\hat{\Xi}_{ijk}} f_{\boldsymbol{m_i}}^2 (\boldsymbol{\xi} \cdot \boldsymbol{m_i}) f_{\boldsymbol{m_j}}(\boldsymbol{\xi} \cdot \boldsymbol{m_j}) f_{\boldsymbol{m_k}}(\boldsymbol{\xi} \cdot \boldsymbol{m_k}) p(\boldsymbol{\xi} \cdot \boldsymbol{m_{ijk}}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_{ijk}}) \\
&+ 24 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=j+1}^{N} \sum_{h=k+1}^{N} \int_{\hat{\Xi}_{ijkh}} f_{\boldsymbol{m_i}}(\boldsymbol{\xi} \cdot \boldsymbol{m_i}) f_{\boldsymbol{m_j}}(\boldsymbol{\xi} \cdot \boldsymbol{m_j}) f_{\boldsymbol{m_k}}(\boldsymbol{\xi} \cdot \boldsymbol{m_k}) f_{\boldsymbol{m_h}}(\boldsymbol{\xi} \cdot \boldsymbol{m_h}) p(\boldsymbol{\xi} \cdot \boldsymbol{m_{ijkh}}) \mathrm{d}(\boldsymbol{\xi} \cdot \boldsymbol{m_{ijkh}}).
\end{aligned}
$$

As already made for the skewness, the previous expression includes some terms always equal to zero thanks to the orthogonality properties of the ANOVA contributions. The final expression for the kurtosis (for more details see the Appendix C) is equal to

$$
\begin{aligned}
k = &\sum_{p=1}^{N} \int_{\hat{\Xi}_p} f_{\boldsymbol{m_p}}^4 \mathrm{d}\mu_p + 4 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_q} \subset \boldsymbol{m_p}} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^3 f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}}^2 \mathrm{d}\mu_{pq} \\
&+ 12 \sum_{p=1}^{N} \sum_{\substack{q=1 \\ q \neq p}}^{N} \sum_{\substack{r=q+1 \\ \boldsymbol{m_{qr}} \setminus \cap_{qr} \subseteq \boldsymbol{m_p}}}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} + 24 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \sum_{\substack{t=r+1 \\ \boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}}}^{N} \int_{\hat{\Xi}_{pqrt}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_{pqrt}.
\end{aligned}
\qquad (23)
$$

Note that the operator of subtraction by set is employed with the standard notation \.

Let us provide the relations to identify the conditional contribution related to a variable or a set of variables. In particular, if a specific multi-index $\boldsymbol{m_i}$ is provided, then the conditional expression for the kurtosis $k_{\boldsymbol{m_i}}$ is equal to (see the Appendix C for more details)

$$
\begin{aligned}
k_{\boldsymbol{m_i}} = &\int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}}^4 \mathrm{d}\mu_i + 4 \int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}}^3 \sum_{\boldsymbol{m_q} \in \mathcal{P}_{i,\neq}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i + 6 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_i \\ \boldsymbol{m_{pq}} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}}^2 \mathrm{d}\mu_i \\
&+ 12 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_r} \in \mathcal{P}_i, r > q \\ \boldsymbol{m_p} \boxplus \cap_{qr} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_i \\
&+ 24 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\boldsymbol{m_q} \in \mathcal{P}_i, q > p} \sum_{\boldsymbol{m_r} \in \mathcal{P}_i, r > q} \sum_{\substack{t > r, \boldsymbol{m_r} \in \mathcal{P}_i \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt} \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{rt}} \boxplus \cap_{pq}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_i
\end{aligned}
\qquad (24)
$$

4. Correlation with Polynomial Chaos Framework

This section is devoted to illustrate how variance, skewness and kurtosis from the functional decomposition are correlated with the polynomial chaos framework. If a polynomial chaos formulation is used, an approximation $\tilde{f}$ of the function f is provided

$$
f(\boldsymbol{\xi}) \approx \tilde{f}(\boldsymbol{\xi}) = \sum_{k=0}^{P} \beta_k \Psi_k(\boldsymbol{\xi}), \qquad (25)
$$

where P is computed according to the order of the polynomial expansion n_0 and the stochastic dimension of the problem d

$$P + 1 = \frac{(n_0 + d)!}{n_0!d!}. \tag{26}$$

Each polynomial $\Psi_k(\boldsymbol{\xi})$ of total degree n_o is a multivariate polynomial form which involve tensorization of 1D polynomial form by using a multi-index $\boldsymbol{\alpha}^k \in \mathbb{N}^d$, with $\sum_{i=1}^{d} \boldsymbol{\alpha}_i^k \leq n_0$:

$$\Psi_k(\boldsymbol{\xi} \cdot \boldsymbol{m}^{\star,k}) = \prod_{i=1}^{d} \psi_{\alpha_i^k}(\xi_i) \tag{27}$$

where the multi index $\boldsymbol{m}^{\star,k} = \boldsymbol{m}^{\star,k}(\alpha^k) \in \mathbb{N}^d$ is a function of $\boldsymbol{\alpha}^k$: $\boldsymbol{m}^{\star,k} = (m_1^{\star,k}, \ldots, m_d^{\star,k})$, with $m_i^{\star,k} = \alpha_i^k / \left\|\alpha_i^k\right\|_{\neq 0}$.

Remark that, for each polynomial basis, $\psi_0(\xi_i) = 1$ and then $\Psi_0(\boldsymbol{\xi}) = 1$. Then, the first coefficient β_0 is equal to the expected value of the function, *i.e.* $E(f)$. The polynomial basis is chosen according to the Wiener-Askey scheme in order to select orthogonal polynomial terms with respect to the probability density function $p(\boldsymbol{\xi})$ of the input. Thanks to the orthogonality, the following relation holds

$$\int_{\Xi} \Psi_i(\boldsymbol{\xi})\Psi_k(\boldsymbol{\xi})p(\boldsymbol{\xi})\mathrm{d}\xi = \delta_{ij}\langle\Psi_i(\boldsymbol{\xi}), \Psi_i(\boldsymbol{\xi})\rangle \tag{28}$$

where $\langle\cdot,\cdot\rangle$ indicates the inner product and δ_{ij} is the Kronecker delta function.

The orthogonality can be advantageously used to compute the coefficients of the expansion in a non-intrusive PC framework as follows

$$\beta_k = \frac{\langle f(\boldsymbol{\xi}), \Psi_k(\boldsymbol{\xi})\rangle}{\langle\Psi_k(\boldsymbol{\xi}), \Psi_k(\boldsymbol{\xi})\rangle}, \quad \forall k. \tag{29}$$

4.1. *Variance decomposition*

First, we compute the term $E(f^2)$ as follows

$$\int_{\Xi^d} f(\boldsymbol{\xi})^2 p(\boldsymbol{\xi})\mathrm{d}\xi = \int_{\Xi^d} \left(\sum_{k=0}^{P} \beta_k \Psi_k(\boldsymbol{\xi})\right)^2 p(\boldsymbol{\xi})\mathrm{d}\xi. \tag{30}$$

This term can be computed easily due to the orthogonality :

$$\int_{\Xi^d} \left(\sum_{k=0}^{P} \beta_k \Psi_k(\boldsymbol{\xi})\right)^2 p(\boldsymbol{\xi})\mathrm{d}\xi = \sum_{k=0}^{P} \beta_k^2 \langle\Psi_k^2(\boldsymbol{\xi})\rangle. \tag{31}$$

As a consequence, variance can be easily computed as

$$\sigma^2 = E(f^2) - E(f)^2 = \sum_{k=1}^{P} \beta_k^2 \langle\Psi_k^2(\boldsymbol{\xi})\rangle. \tag{32}$$

Finally, an explicit correlation between the last expression and the Eq. (13) is found. As done for the the functional decomposition of the variance (see §3.1), let us compute each conditional term of the variance. Remembering the equation (14), each conditional term can be computed as

$$\sigma_{\boldsymbol{m_i}}^2 = \sum_{k \in K_{\boldsymbol{m_i}}} \beta_k^2 \langle\Psi_k^2(\boldsymbol{\xi})\rangle, \tag{33}$$

where $K_{\boldsymbol{m_i}}$ represent the set of indices associated to the variable included in the vector $(\boldsymbol{\xi} \cdot \boldsymbol{m_i})$:

$$K_{\boldsymbol{m_i}} = \left\{k \in \{1, \ldots, P\} \,|\, \boldsymbol{m}^{\star,k} = \boldsymbol{m}^{\star,k}(\alpha^k) = \boldsymbol{m_i}\right\} \tag{34}$$

4.2. *Skewness decomposition*

In this section, following what already reported for the variance, the decomposition of the skewness is performed with respect to a PC expansion. The PC expansion can be raised to the third power to obtain (by applying again the multinomial theorem)

$$
\begin{aligned}
s &= \int_{\Xi} (f(\boldsymbol{\xi}) - \beta_0)^3\, p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\Xi} \left(\sum_{p=1}^{P} \beta_p \Psi_p(\boldsymbol{\xi}) \right)^3 p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \\
&= \sum_{p=1}^{P} \beta_p^3 \langle \Psi_p^3(\boldsymbol{\xi}) \rangle + 3 \sum_{p=1}^{P} \beta_p^2 \sum_{\substack{q=1 \\ q\neq p}}^{P} \beta_q \langle \Psi_p^2(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi}) \rangle + 6 \sum_{p=1}^{P} \sum_{q=p+1}^{P} \sum_{r=q+1}^{P} \beta_p\beta_q\beta_r \langle \Psi_p(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi}) \rangle .
\end{aligned}
\tag{35}
$$

As already shown for the ANOVA functional decomposition, also in this case there are several terms always equal to zero thanks to the orthogonality properties of the PC basis. The final form, explicitly obtained in the Appendix D is equal to

$$
s = \sum_{p=1}^{P} \beta_p^3 \langle \Psi_p^3(\boldsymbol{\xi}) \rangle + 3 \sum_{p=1}^{P} \beta_p^2 \sum_{\substack{q=1 \\ q\neq p}}^{P} \beta_q \langle \Psi_p^2(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi}) \rangle \Delta_q^p + 6 \sum_{p=1}^{P} \sum_{q=p+1}^{P} \sum_{r=q+1}^{P} \beta_p\beta_q\beta_r \langle \Psi_p(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi}) \rangle \Delta_{pqr}, \tag{36}
$$

where two functions are introduced for the selection. The first one Δ_q^p is defined as follows

$$
\Delta_q^p = \begin{cases} 0 & \text{if} \quad \alpha_j^p = 0 \quad \text{and} \quad \boldsymbol{m_q}_j = 1 \\ 1 & \text{otherwise} \end{cases} \tag{37}
$$

while the function Δ_{pqr} is defined as

$$
\Delta_{pqr} = \begin{cases} 0 & \text{if} \quad \boldsymbol{m_p}_j + \boldsymbol{m_q}_j + \boldsymbol{m_r}_j = 1, 2 \\ 1 & \text{otherwise.} \end{cases} \tag{38}
$$

If a fixed multi-index $\boldsymbol{m_i}$ is of interest, the previous expression reduces to

$$
s = \sum_{p\in K_{\boldsymbol{m_i}}} \beta_p^3 \langle \Psi_p^3(\boldsymbol{\xi}) \rangle + 3 \sum_{p\in K_{\boldsymbol{m_p}}} \beta_p^2 \sum_{\substack{q\in K_{\boldsymbol{m_q}} \\ \boldsymbol{m_{pq}}=\boldsymbol{m_i}}} \beta_q \langle \Psi_p^2(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi}) \rangle \Delta_q^p + 6 \sum_{p\in K_{\boldsymbol{m_p}}} \sum_{\substack{q\in K_{\boldsymbol{m_q}} \\ q\geq p+1}} \sum_{\substack{r\in K_{\boldsymbol{m_r}} \\ \boldsymbol{m_{pqr}}=\boldsymbol{m_i}}} \beta_p\beta_q\beta_r \langle \Psi_p(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi}) \rangle \Delta_{pqr}. \tag{39}
$$

Note that the two functions Δ_q^p and Δ_{pqr}, should be computed before computing the integral associated to each term for an efficient implementation. Only if this value is one then the integral need to be truly computed. However, it is clear that a brute force approach in which all the terms are computed still works even if it is not efficient from a computational point of view.

4.3. *Kurtosis decomposition*

The conditional terms for the kurtosis are presented in this section. After the application of the multinomial theorem, the expression for the kurtosis is obtained from the PC series expansion as follows

$$
\begin{aligned}
k &= \int_{\Xi} (f(\boldsymbol{\xi}) - \beta_0)^p\, (\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\Xi} \left(\sum_{p=1}^{P} \beta_p \Psi_p(\boldsymbol{\xi}) \right)^4 p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \\
&= \sum_{k=1}^{P} \beta_k^4 \langle \Psi_k^4(\boldsymbol{\xi}) \rangle + 4 \sum_{i=1}^{P} \beta_i^3 \sum_{\substack{j=1 \\ j\neq i}}^{P} \beta_j \langle \Psi_i^3, \Psi_j \rangle + 6 \sum_{i=1}^{P} \beta_i^2 \sum_{j=i+1}^{P} \beta_j^2 \langle \Psi_i^2, \Psi_j^2 \rangle \\
&+12 \sum_{i=1}^{P} \beta_i^2 \sum_{\substack{j=1 \\ j\neq i}}^{P} \beta_j \sum_{\substack{k=j+1 \\ k\neq i}}^{P} \beta_k \langle \Psi_i^2, \Psi_j\Psi_k \rangle + 24 \sum_{i=1}^{P} \sum_{j=i+1}^{P} \sum_{k=j+1}^{P} \sum_{h=k+1}^{P} \beta_i\beta_j\beta_k\beta_h \langle \Psi_i\Psi_j, \Psi_k\Psi_h \rangle .
\end{aligned}
\tag{40}
$$

Also in this case, many integrals have not to be computed if the orthogonal contributions are clearly identified. Hereafter the final expression, obtained in Appendix E, is equal to

$$\begin{aligned} k &= \int_{\Xi} (f(\boldsymbol{\xi}) - \beta_0)^p \, (\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\Xi} \left(\sum_{p=1}^{P} \beta_p \Psi_p(\boldsymbol{\xi}) \right)^4 p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \\ &= \sum_{p=1}^{P} \beta_p^4 \langle \Psi_p^4(\boldsymbol{\xi}) \rangle + 4 \sum_{p=1}^{P} \beta_p^3 \sum_{\substack{q=1 \\ q \neq p}}^{P} \beta_q \langle \Psi_p^3, \Psi_q \rangle \Delta_q^p + 6 \sum_{p=1}^{P} \beta_p^2 \sum_{q=p+1}^{P} \beta_q^2 \langle \Psi_p^2, \Psi_q^2 \rangle \\ &+12 \sum_{p=1}^{P} \beta_p^2 \sum_{\substack{q=1 \\ q \neq p}}^{P} \beta_q \sum_{\substack{r=q+1 \\ r \neq p}}^{P} \beta_r \langle \Psi_p^2, \Psi_q \Psi_r \rangle \Delta_{qr}^p + 24 \sum_{p=1}^{P} \sum_{q=p+1}^{P} \sum_{r=q+1}^{P} \sum_{t=r+1}^{P} \beta_p \beta_q \beta_r \beta_t \langle \Psi_p \Psi_q, \Psi_r \Psi_t \rangle \Delta_{pqrt}, \end{aligned} \tag{41}$$

where the function Δ_q^p is already introduced in (37), while the others two functions are defined as follows

$$\Delta_{qr}^p = \begin{cases} 0 & \text{if} \quad \alpha_j^p = 0 \quad \text{and} \quad \boldsymbol{m_q}_j + \boldsymbol{m_r}_j = 1,2 \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad \Delta_{pqrt} = \begin{cases} 0 & \text{if} \quad \boldsymbol{m_p}_j + \boldsymbol{m_q}_j + \boldsymbol{m_r}_j + \boldsymbol{m_t}_j = 1,2 \\ 1 & \text{otherwise} \end{cases} \tag{42}$$

In the case of the conditional contribution associated to a specific multi-index $\boldsymbol{m_i}$, it holds

$$\begin{aligned} k_{\boldsymbol{m_i}} &= \sum_{k \in K_{\boldsymbol{m_i}}} \beta_k^4 \langle \Psi_k^4(\boldsymbol{\xi}) \rangle + 4 \sum_{p \in K_{\boldsymbol{m_p}}} \beta_p^3 \sum_{\substack{q \in K_{\boldsymbol{m_q}} - \{p\} \\ \boldsymbol{m_p} \boxplus \boldsymbol{m_q} = \boldsymbol{m_i}}} \beta_q \langle \Psi_p^3, \Psi_q \rangle \Delta_q^p + 6 \sum_{p \in K_{\boldsymbol{m_p}}} \beta_p^2 \sum_{\substack{q \in K_{\boldsymbol{m_q}} - \{p\} \\ \boldsymbol{m_p} \boxplus \boldsymbol{m_q} = \boldsymbol{m_i}}} \beta_q^2 \langle \Psi_p^2, \Psi_q^2 \rangle \\ &+12 \sum_{p \in K_{\boldsymbol{m_p}}} \beta_p^2 \sum_{q \in K_{\boldsymbol{m_q}} - \{p\}} \beta_q \sum_{\substack{r \in K_{\boldsymbol{m_r}} \\ r \geq q+1 \\ \boldsymbol{m_{pqr}} = \boldsymbol{m_i}}} \beta_r \langle \Psi_p^2, \Psi_q \Psi_r \rangle \Delta_{qr}^p + 24 \sum_{p \in K_{\boldsymbol{m_p}}} \sum_{\substack{q \in K_{\boldsymbol{m_q}} \\ q \geq p+1}} \sum_{\substack{r \in K_{\boldsymbol{m_r}} \\ r \geq q+1}} \sum_{\substack{t \in K_{\boldsymbol{m_t}} \\ t \geq r+1 \\ \boldsymbol{m_{pqrt}} = \boldsymbol{m_i}}} \beta_p \beta_q \beta_r \beta_t \langle \Psi_p \Psi_q, \Psi_r \Psi_t \rangle \Delta_{pqrt}. \end{aligned} \tag{43}$$

5. Introducing more sensitivity indices

As introduced by Sobol [19], sensitivity indexes for variance can be computed for each conditional contribution following Eq. (14):

$$\sigma_{\boldsymbol{m_i}}^{2,\mathrm{SI}} = \frac{\sigma_{\boldsymbol{m_i}}^2}{\sigma^2}. \tag{44}$$

Here, we introduce new sensitivity indexes, basing on the decomposition of skewness and kurtosis and using the definition of the conditional term in (21) and (24), as follows

$$\begin{aligned} s_{\boldsymbol{m_i}}^{\mathrm{SI}} &= \frac{s_{\boldsymbol{m_i}}}{s} \\ k_{\boldsymbol{m_i}}^{\mathrm{SI}} &= \frac{k_{\boldsymbol{m_i}}}{k}. \end{aligned} \tag{45}$$

If a total sensitivity index is needed, *i.e.* it is necessary to compute the overall influence of a variable, it can be computed summing up all the contributions in which the variable is present

$$\begin{aligned} \mathrm{TSI}_j &= \sum_{\xi_j \in (\boldsymbol{\xi} \cdot \boldsymbol{m_i})} \sigma_{\boldsymbol{m_i}}^{2,\mathrm{SI}} \\ \mathrm{TSI}_j^s &= \sum_{\xi_j \in (\boldsymbol{\xi} \cdot \boldsymbol{m_i})} s_{\boldsymbol{m_i}}^{\mathrm{SI}} \\ \mathrm{TSI}_j^k &= \sum_{\xi_j \in (\boldsymbol{\xi} \cdot \boldsymbol{m_i})} k_{\boldsymbol{m_i}}^{\mathrm{SI}}. \end{aligned} \tag{46}$$

6. Numerical results

In this section, the importance of considering high-order statistics for global sensitivity analysis is demonstrated through some numerical examples. The numerical test cases are chosen in order to highlight the importance of the analysis related to the high-order conditional contributions for the analysis of systems including multiple sources of uncertainties. The focus will be devoted to the analysis of variance-based reduction strategies, where the importance of considering high-order contributions is demonstrated. The numerical test section is organized as follows. In section §6.1, the previous relations for the computation of the high-order conditional terms are demonstrated numerically by showing the convergence properties of PC with respect to the analytical conditional high-order statistics. In section §6.2, a comparison is performed between the information obtained by an analysis based on the variance or on high-order conditional contributions, by solving some numerical test problems with different kind of interactions between parameters. Finally, how reducing the model dimension is addressed in sections §6.3 and §6.4, where the reduction procedure in the truncation (by reducing the number of dimensions) and in the superposition (by reducing the order of interactions) sense is presented, respectively.

6.1. Computing conditional statistics by means of PC

In this section, the problem of the computation of high-order conditional terms is analyzed by means of the PC expansion series (see section §4). If the aim is to compute the high-order statistics (the values of variance, skewness or kurtosis) employing the same set of deterministic evaluation of the models, *i.e.* the same number of functional evaluation $f = f(\boldsymbol{\xi})$ in the same sample points, their values can be obtained only by computing the first order coefficients of the PC expansion for f, f^2, f^3 and f^4 corresponding to the expected values of the four functions. The combination of the expectancies of the function f raised up to an increasing power can be employed to compute the total variance, skewness and kurtosis according to the relations reported in Appendix A. Hereinafter, the combination of central moments in order to obtain high-order statistic employing only evaluation of the function f in the same quadrature points of the PC expansion is referred as *collocation*. It is important to remark that the collocation approach does not provide any kind of metamodel for the function f, nor the possibility to compute conditional terms. Then, this approach is employed only for a comparison with respect to the PC series and for assessing the convergence of the expansion.

Consider the following function

$$f(\boldsymbol{\xi}) = \prod_{i=1}^{d} \sin(\pi \xi_i) \tag{47}$$

where each variable $\xi_i \sim \mathcal{U}(0,1)$ with an increasing dimension d up to three. In the following, statistical moments as well as sensitivity indexes (relative) errors are systematically computed with respect to the analytical solution.

In figure 1, statistical moments convergence are reported as a function of the number of functional evaluations for the dimension $d = 2$. A number of simulations equal to $N = 120$ is needed to reach a relative order of the error of order $O(10^{-4})$ for the kurtosis while for lower statistics is reached with few realizations. Also the collocation approach, as expected, converges faster, but, as already discussed, is limited to the computation of the full central moments.

Now, conditional statistics can be computed using a PC approach using Eqs. (36) and (41). In Figure 2, we show first-order statistics v_1, s_1 and k_1 (where for symmetry $\sigma_1^2 = \sigma_2^2 = v1$, $s_1 = s_2$ and $k_1 = k_2$) and interaction terms (v_{12}, s_{12}, k_{12}) errors computed with respect to the analytical solution. These statistics are well converged at $N = 120$. Then, the case with $d = 3$ is taken into account. In Figures 3 and 4, convergence of statistical moments and conditional statistics are reported, respectively. As observed earlier, collocation gives higher convergence rate with respect to the PC (Figure 3(a)). Convergence, for both statistical moments and conditional statistics, is attained at nearly $N = 1500$. This illustrates how it becomes computationally expensive to have well-converged statistics terms when the number of dimension increases.

In the following sections, the high-order conditional statistics are employed to show the importance of the high-order interactions between uncertain parameters for the reduction of a numerical model.

6.2. On the advantages of high-order indexes for global Sensitivity analysis

The importance of including high-order conditional terms computation in the statistics analysis is demonstrated in this section by means of several model functions. Note that this kind of approach is conceived in order to extend the

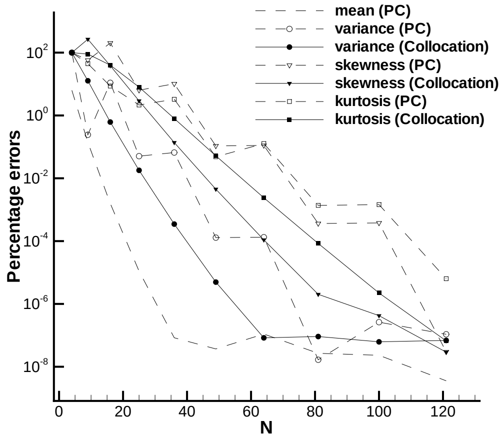


Figure 1: Statistical moments error vs number of function evaluations in the case $d = 2$.

global sensitivity analysis based on the variance. Anyway, in some situations, criteria based on statistical moments are not adequate and moment independent criteria should be adopted. The interested reader should refer to [34] for a discussion in this sense.

The first model function is the well-known Ishigami function

$$f(\boldsymbol{\xi}) = (1 + 0.1\xi_3^4)\sin(\xi_1) + 7\sin(\xi_2)^2 \quad \text{where} \quad \xi_i \sim \mathcal{U}(-\pi,\pi). \tag{48}$$

For this function, the first-order Sensitivity Indexes (SI) contribution (for variance, skewness and kurtosis) computed for to the third variable ξ_3 are equal to zero. In Figure 5, sensitivity indexes are reported for variance, skewness and kurtosis.

Remark that the interaction between the third and the first variable is not negligible, obtaining a $k^{\mathrm{SI}}_{\boldsymbol{m_i}}$, for $\boldsymbol{m_i} = (1,0,1)$, higher than 0.4. Also the interaction between the three variables is almost equal to 0.2 for the kurtosis but it is zero for both variance and skewness. It is also interesting to note that, even if the ranking of the variables it is not directly affected by the choice of the order of Sensitivity Indexes (SI) (*i.e.* for the variance, skewness or kurtosis), the three indexes provide complementary results. For instance, the relative importance of first-order terms is about 0.75 for the variance, while it is only 0.15 for the kurtosis. This different impact on high-order interactions is even more evident when the problem is to reduce the model, as it will be shown in Section 6.4.

The analysis of the table 1, where the total sensitivity indexes are reported, confirms that the ranking of the variables is nearly the same for each statistical moment.

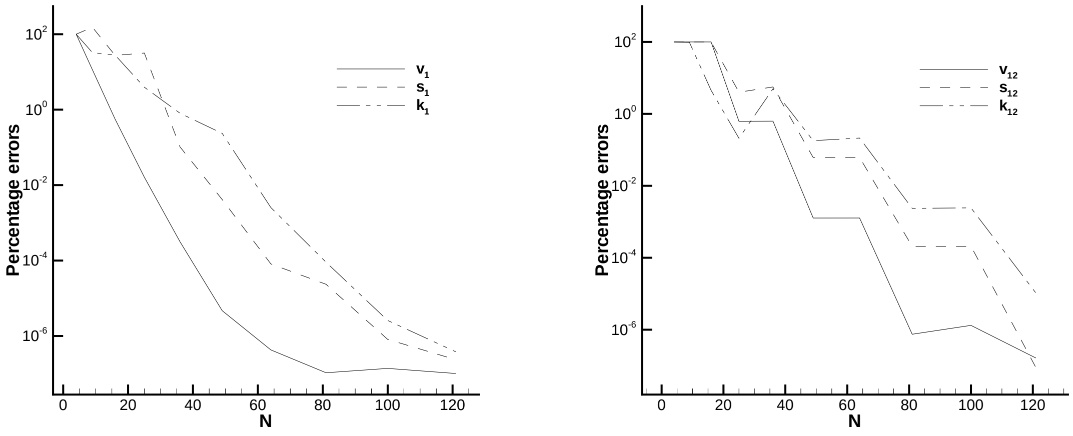
Variable	TSI	TSI^{s}	TSI^{k}
ξ_1	0.57	0.00	0.91
ξ_2	0.43	1.00	0.50
ξ_3	0.25	0.00	0.64

Table 1: Total sensitivity indexes for the Ishigami function (48) based on a PC series with total degree $n_0 = 7$.

Let us consider now, the classical Sobol function (four dimension)

$$f(\boldsymbol{\xi}) = \prod_{i=1}^{4} \frac{|4\xi_i - 2| + a_i}{1 + a_i}, \tag{49}$$

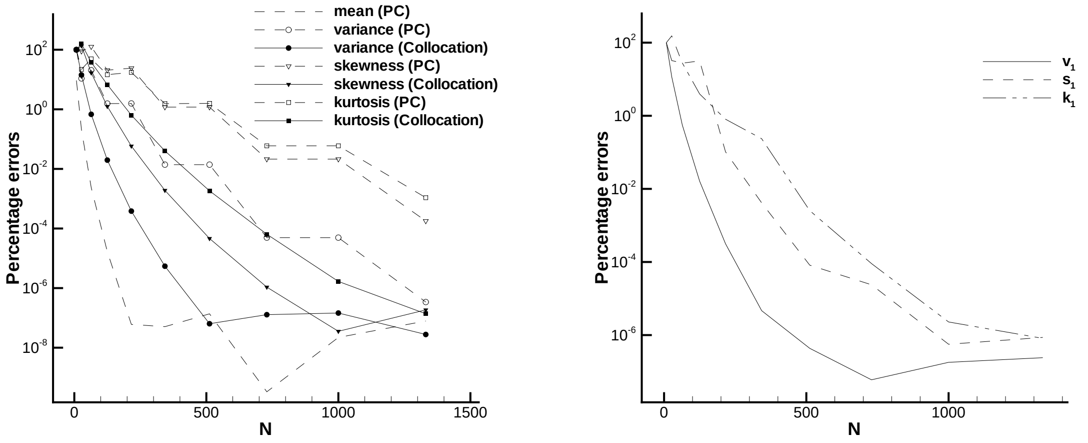
where $\xi_i \sim \mathcal{U}(0,1)$. Two possible choices of the coefficients are considered here



(a)

(b)

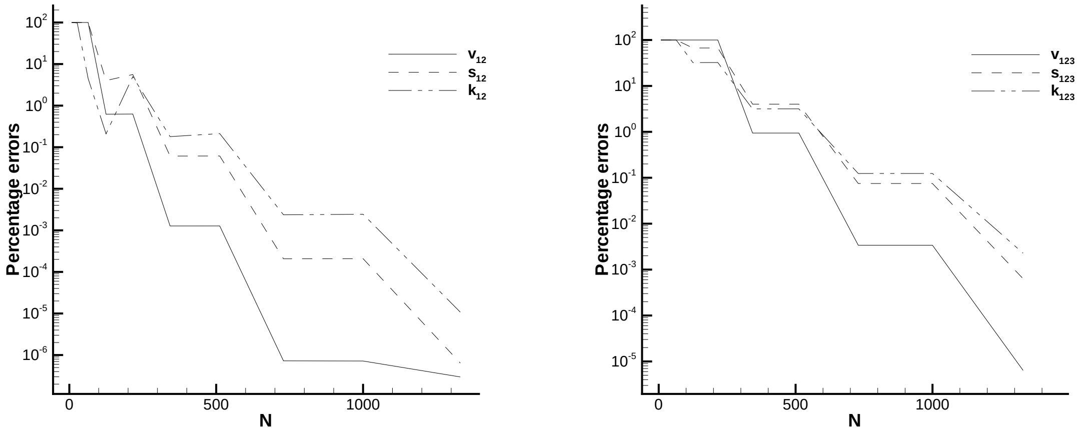
Figure 2: Conditional statistics vs number of function evaluations in the case $d = 2$.



(a)

(b)

Figure 3: Statistical moments (a) and conditional statistics (b) error vs number of function evaluations in the case $d = 3$.



(a)

(b)

Figure 4: Sensitivity indices vs number of function evaluations in the case $d = 3$.

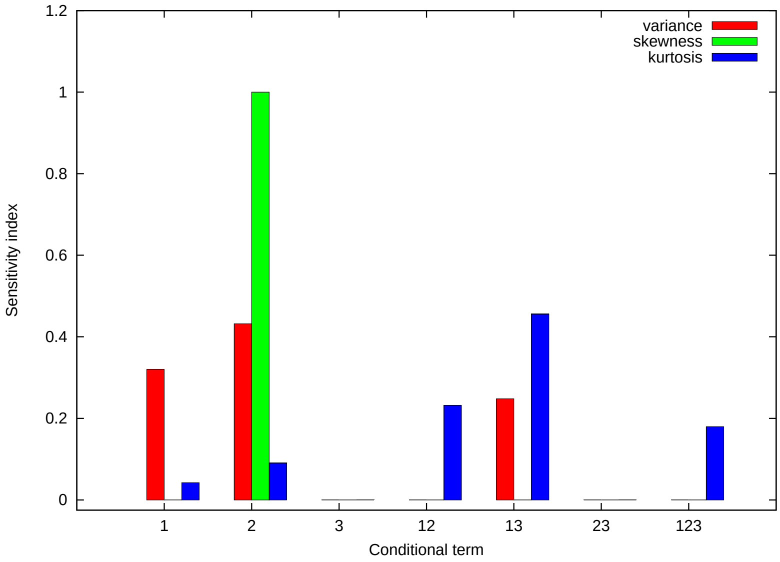


Figure 5: Sensitivity indexes for the Ishigami function (48) obtained with a PC series with total degree $n_0 = 7$.

- $a_i = (i - 1)/2$ the so called linear g-function f_{glin}
- $a_i = i^2$ the so called quadratic g-function f_{gquad}.

In figure 6, Sensitivity Indexes (SI) for the linear g-function f_{glin} are reported. Looking at figure 6, several differences can be noticed between the sensitivity indexes computed on the variance or on other high-order moments. The variance-based ranking illustrates that the first-order sensitivity indexes are higher than the second order one, while these last ones are higher than the third and fourth order ones. This is not the case for skewness and kurtosis, where the second-order contributions are higher than the first-order and third-order ones. This behavior reveals that the variance is able to catch the absolute ranking of the variables in terms of first-order contributions, but the importance associated to higher-order interactions between the parameters is totally lost. From a practical point of view, underestimating the importance of high-order interactions between variables can lead to wrong decisions in a dimension reduction strategy as it will be shown in Sections 6.3 and 6.4. The variance based only on first-order contributions exceeds 0.8, while skewness and kurtosis do not attain 0.1. This can be demonstrated to be very influential if the probability distribution for reduced models is considered. However, in table 2, the total sensitivity indexes for the four variables are reported. It is evident that the ranking of variables is not influenced by the statistical moment, but their relative importance can vary significantly.

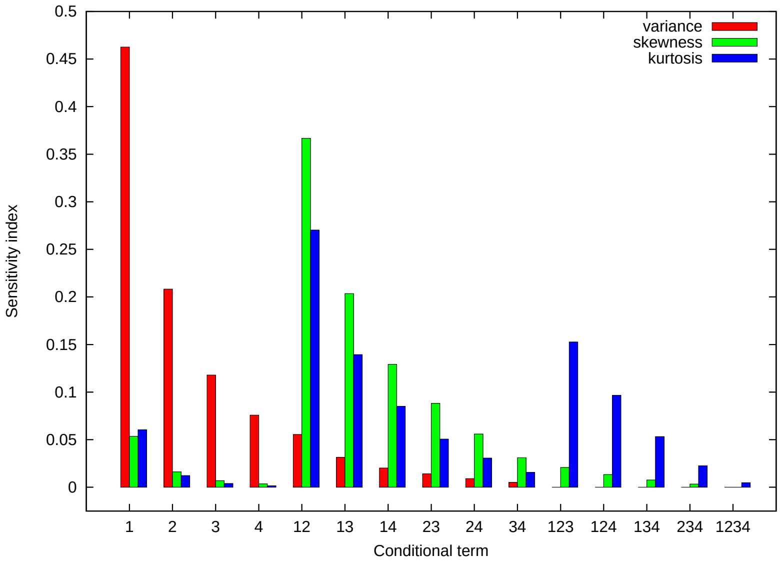


Figure 6: Sensitivity indexes for the linear g-function f_{glin} (49) obtained with a PC series with total degree $n_0 = 5$.

Variable	TSI	TSI^s	TSI^k
ξ_1	0.57	0.79	0.86
ξ_2	0.29	0.56	0.64
ξ_3	0.17	0.36	0.44
ξ_4	0.11	0.24	0.31

Table 2: Total sensitivity indexes for the linear g-function function (49) based on a PC series with total degree $n_0 = 5$.

The same functional form can lead to slightly different results if the quadratic function coefficients are considered. In Figure 7, the sensitivity indexes for the g-function with a quadratic dependence of the coefficients are reported. In this case, the difference between the first order contribution and high-order terms is even more evident. Considering the variance, first-order contributions exceed 0.98, while a value larger than 0.5 is computed for high-order interactions when considering skewness and kurtosis. In this case, the contribution of the first variable exceeds 0.8, but in order to attain this level, it is necessary to include contributions related to the first variable and the second-order interaction between the first and second variable. In the table 3, total sensitivity indexes are reported for the four variables. In this case, variance contributions for both the third and fourth variables are below 0.05, while for both skewness and kurtosis, only the fourth variable contribution takes a TSI value of 0.04. A low level of TSI for the variables ξ_3 and ξ_4 could suggest to truncate the dimensionality of the model to the first two variables or neglect the contributions related to the order higher than one. This case is analyzed in the following section in order to demonstrate the importance of high-order sensitivity indexes analysis.

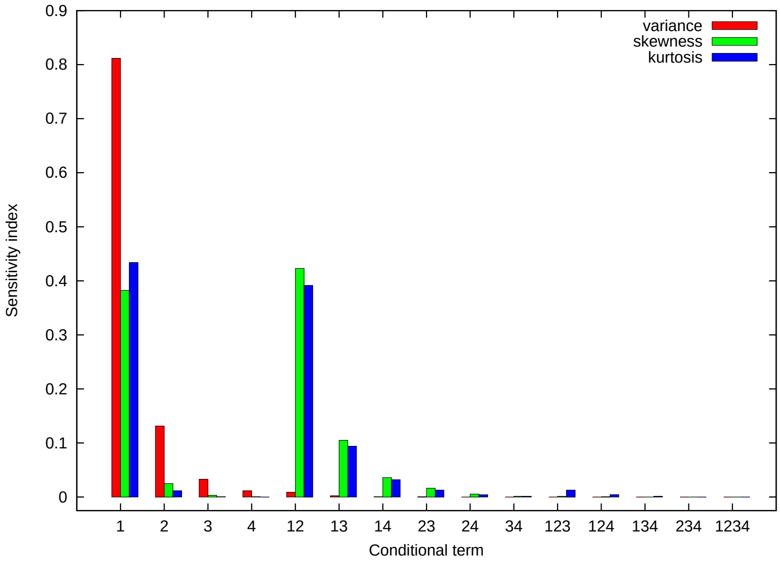


Figure 7: Sensitivity indexes for the quadratic g-function f_{gquad} (49) obtained with a PC series with total degree $n_0 = 5$.

Variable	TSI	TSI^s	TSI^k
ξ_1	0.82	0.95	0.97
ξ_2	0.14	0.47	0.44
ξ_3	0.04	0.13	0.12
ξ_4	0.01	0.04	0.04

Table 3: Total sensitivity indexes for the quadratic g-function f_{gquad} (49) based on a PC series with total degree $n_0 = 5$.

Let us now consider the following functions:

$$\begin{aligned} f_1 &= \xi_1 e^{\frac{\xi_2}{\xi_3^2+1}} + \xi_1\xi_2 \\ f_2 &= \prod_{i=1}^{3} \frac{2\xi_i + 1}{2}, \end{aligned} \tag{50}$$

where the parameters are $\xi_i \sim \mathcal{U}(0, 1)$.

Sensitivity indexes associated to the first function f_1 are reported in Figure 8. For the function f_1, the most important variable is ξ_1. For the variance, the first-order sensitivity index relative to ξ_1 is also the most important SI. On the contrary, for both skewness and kurtosis, the highest SI is associated to the second-order interaction between the first and the second variable. In this case, the inspection of the total sensitivity indexes, reported in the table 4, suggests that the third variable ξ_3 is meaningless with respect to the variance. The TSI associated to ξ_3 are lower than the limit proposed in [35] to identify a negligible uncertainty that could be frozen. However, if this information is used together with the high-order total sensitivity indexes information, the choice of freezing the third variable should be considered more carefully. This reflects the importance of the third variable in the actual form of the probability density function of f_1 even if its variance is not heavily influenced by it. The results of a model reduction decision, totally based on variance measures, is further discussed in the following section.

The last example, *i.e.* the function f_2, is reported here to underline the difference between the measure of sensitivity associated to the variance and to the higher-order moments. In particular, the functional form of f_2 (50) includes an equal contribution of three variables. However, looking at the figure 9, it is possible to note that the variance is concentrated only on first-order contributions of the single variables and their sum exceeds 0.9. The skewness and kurtosis contributions, on the contrary, are concentrated on second-order interaction. For kurtosis, the third-order

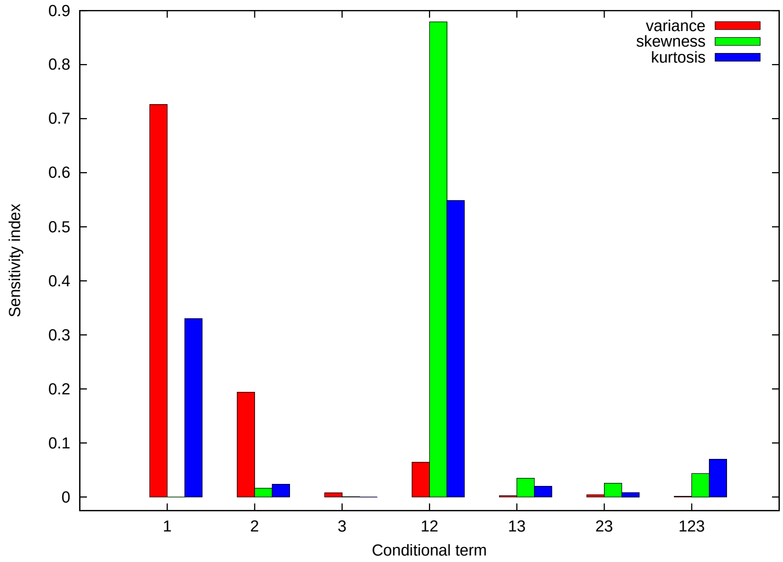


Figure 8: Sensitivity indexes for the first function f_1 (50) obtained with a PC series with total degree $n_0 = 7$.

Variable	TSI	TSI^s	TSI^k
ξ_1	0.79	0.96	0.97
ξ_2	0.26	0.96	0.67
ξ_3	0.02	0.10	0.10

Table 4: Total sensitivity indexes for the first function f_1 (50) based on a PC series with total degree $n_0 = 7$.

interaction is the highest contribution. Remark that even if the sum of the first-order variance contribution exceeds 0.9, a reduction of the model in the superposition sense (*i.e.* by neglecting the high orders of interaction), could lead to wrong conclusions, as explained in Section 6.4. The skewness associated to a model including only first-order contribution does not include the skewness information about the probability distribution of the output.

Values for the total sensitivity indexes are reported in table 5 for this case. It is interesting to note that the sum of the total sensitivity indexes over the three variables is much more higher for skewness and kurtosis with respect to the variance. Then, they refer, correctly, to an intrinsically high-order (of interaction) function (see equation (50) for f_2 definition).

Variable	TSI	TSI^s	TSI^k
ξ_1	0.36	0.70	0.71
ξ_2	0.36	0.70	0.71
ξ_3	0.36	0.70	0.71

Table 5: Total sensitivity indexes for the first function f_2 (50) based on a PC series with total degree $n_0 = 7$.

Numerical test-cases presented in this section illustrate how information relative to variance-based sensitivity indexes seem to be incomplete in order to understand the true dependence of a model from its variables. In particular, variance seems to be more associated to low order interaction with respect to the sensitivity indexes associated to skewness and kurtosis. Then, for a whatever function that is known by points, *i.e.* for example experimental observations or computer runs of a code, the sensitivity indexes on the skewness and on the kurtosis could be very helpful to capture some interactions between subset of variables, much more than the variance.

This could be even more important if the aim is to reduce the dimensionality of the problem and to build an accurate metamodel. The following sections are focused on the advantages in using the high-order sensitivity indexes

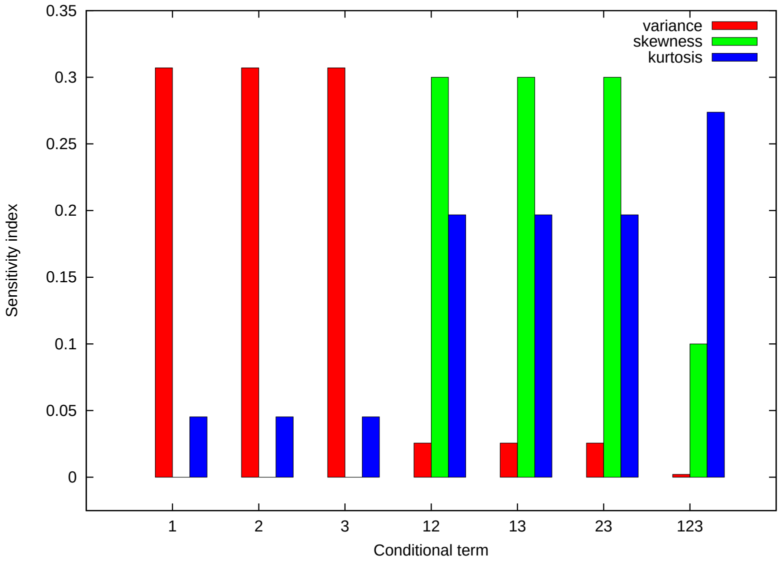


Figure 9: SI for the function f_2.

in a truncation strategy framework.

6.3. *Dimensional reduction in the truncation sense*

In this section, the problem of reducing the number of dimensions is analyzed through some numerical test-cases based on the results obtained in the previous section. The first test-case is represented by the quadratic g-function (49). From the analysis conducted in the previous section (see table 3), note that the third and fourth variables seem to be meaningless for the variance-based indexes. Their total sensitivity indexes sum up to 0.05 for the variance, while exceed 0.15 for both skewness and kurtosis. Considering only the sensitivity indexes computed on the variance, the decision-maker could be tempted to neglect the variables ξ_3 and ξ_4. In this case, the ANOVA expansion does not include the terms containing ξ_3 and ξ_4, as follows

$$\begin{aligned} f_{G1} &= f_0 + f_1(\xi_1) + f_2(\xi_2) + f_{12}(\xi_1, \xi_2) \\ f_{G2} &= f_0 + f_1(\xi_1) + f_2(\xi_2) + f_{12}(\xi_1, \xi_2) + f_3(\xi_3) + f_{13}(\xi_1, \xi_3) + f_{23}(\xi_2, \xi_3) + f_{123}(\xi_1, \xi_2, \xi_3), \end{aligned} \tag{51}$$

where in the first case f_{G1} both are neglected; on the contrary for f_{G2} only ξ_4 is neglected. In this case, the ANOVA terms and the statistics can be computed analytically. In the table 6, the percentage errors, for the first four central moments, are reported with respect to the analytical exact solution for both the reduced models f_{G1} and f_{G2}.

Function	Variance	Skewness	Kurtosis
f_{G1}	4.7997	29.236	15.039
f_{G2}	1.2369	7.7705	4.0632

Table 6: Percentage $\left(\frac{abs(\mu - \mu_{ex})}{\mu_{ex}} \times 100\right)$ errors related to the reduced g-function f_{G1} and f_{G2}.

In table 6, it is evident that an error of only 5% on the variance can correspond to a much greater error on the higher moments. This behavior is justified looking at the Figure 10, where the probability density function is computed for both f_{G1} and f_{G2} and compared with the complete function (49). In this case, the model with only the first two variables can not reproduce the tails while a good approximation is attained in the middle part. However, this test-case clearly shows that considering only the sensitivity indexes based on the variance could be very risky in a decision-making process. In this case, the pdf results to be analytically bounded between 0.4 and 1.8. If the third variable is included in the reduced model, both variance and skewness are computed with an error lower than 5%, while the error on the kurtosis remains lower than 8%. The total sensitivity indexes associated to the fourth variable is reported in table 3 and it is lower than 5% for the three moments. The improvement of the model given by including the third variable is evident in Figure 10, where the pdf of the reduced model allows recovering much better the pdf of the complete function.

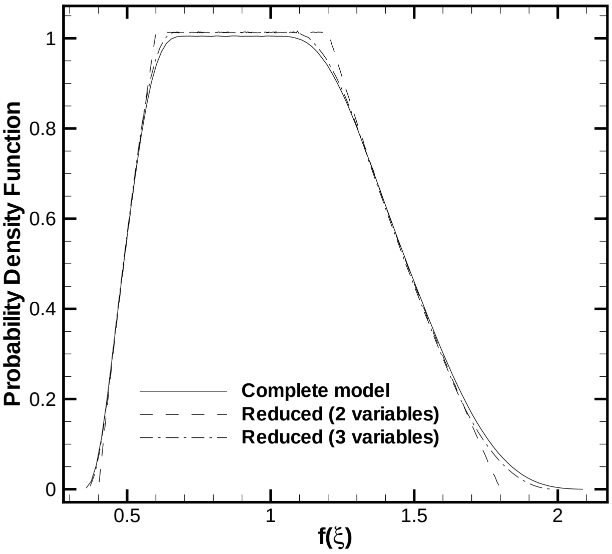


Figure 10: PDFs for the complete g-function and the reduced models (see equations 51).

From a practical point-of-view, the dimension reduction is commonly applied by freezing the neglected parameters. For an analytical function, it is possible to compute the constant values to choose, for both ξ_3 and ξ_4, in order to obtain a reduced model that preserves both the expectancy and the variance of the original complete model. Of course, both requirements cannot be satisfied at the same time, but a set of values satisfying the mean and the variance can be obtained analytically requiring that

$$\begin{aligned} \frac{|4\overline{\xi}_j - 2| + a_j}{1 + a_j} &= 1 \\ \left(\frac{|4\overline{\xi}_j - 2| + a_j}{1 + a_j}\right)^2 &= \int_0^1 \left(\frac{|4\xi_j - 2| + a\mathrm{J}}{1 + a_j}\right)^2 \mathrm{d}\xi_j. \end{aligned} \tag{52}$$

The following values can be analytically computed for the two variables: $\xi_3 = \{1/4, 3/4, 91/120, 29/120\}$ and $\xi_4 = \{1/4, 3/4, 77/102, 25/102\}$.

In Figure 11, the pdf associated to the complete quadratic g-function with parameters ξ_3 and ξ_4 frozen, are reported with the complete pdf and the totally reduced one.

From Figure 11, it is evident that freezing some parameters in order to assure the correctness of the mean and the variance, yields pdf very close to that one obtained by neglecting entirely the ANOVA terms. From a practical point-of-view, the analysis of the reduced model can be carried out both with the ANOVA reduced model (if it is analytically possible to compute the integrals) and by freezing the parameter to neglect by satisfying the requirement on the expectancy and variance. In both cases, results make evident that a variance-based sensitivity analysis should be supplemented by high-order sensitivity analysis for building a reduced model which does not deteriorate the distribution of the realizations, especially in the tails.

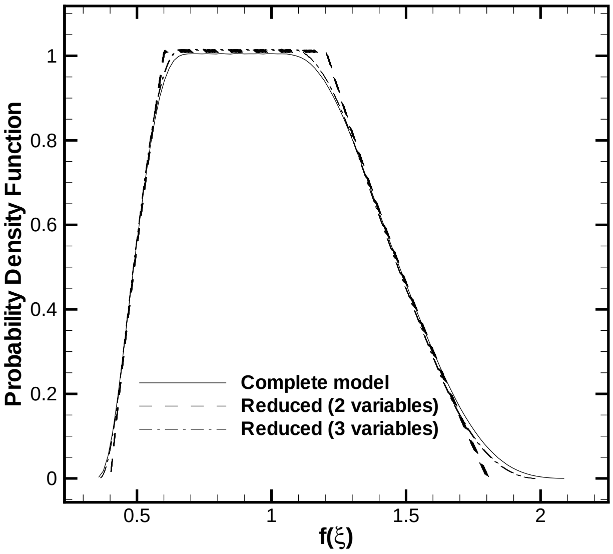


Figure 11: PDFs for the complete g-function and the reduced models.

Now, let us analyze the function f_1. In the previous section, the total sensitivity indexes for the three variables has been reported in the table 4. For the third variable, the level of the TSI of 1.55%, is inferior to the threshold of 2%, indicated in [35], to detect meaningless parameters. A reduced model can be obtained by freezing the third parameter, or equivalently as shown in the first part of this section, by neglecting all the ANOVA terms in which the variable ξ_3 is present

$$\hat{f}_1 = f_0 + f_1(\xi_1) + f_2(\xi_2) + f_{12}(\xi_1, \xi_2). \tag{53}$$

For a variable ξ_3 recovering the exact value of the mean and the variance of the complete model, the following values are obtained: $\xi_3 = \{0.4283, 0.4166\}$. In the figure 12, the pdf for the complete model and the one obtained by freezing the third parameters are reported. Even in this case, *i.e.* with a model permitting to obtain an error on the variance inferior to 2%, the information about the tails of the distributions are, again, totally lost. This is a further confirmation that the information about high-order sensitivity indexes should be considered for building an accurate metamodel. In all the case proposed here and in others not reported here for brevity, it appears evident that only when even the high-order sensitivity indexes have reached a safety threshold (about 5%), the model can be really (and more safely) truncated.

6.4. Dimensional model reduction in the superposition sense

In this section, the problem of the truncation is analyzed from a different perspective, in a so-called *superposition sense*. This means that the dimension of the model is not reduced in terms of number of variables, but in terms of order of interaction between variables. Note also that, if the function is approximated by means of a PC series of total degree n_0, all the interactions of order $n_0 + 1$ are lost.

One could choose to truncate the model, neglecting all the interaction of higher order, if the error on the statistics has already attained a specific threshold associated to the application of interest.

Consider the contribution up to the order t, *i.e.*

$$f(\boldsymbol{\xi}) = \sum_{\boldsymbol{m}_i} f_{\boldsymbol{m}_i} \simeq \sum_{|\boldsymbol{m}_i| \leq t} f_{\boldsymbol{m}_i} = \hat{f}(\boldsymbol{\xi}). \tag{54}$$

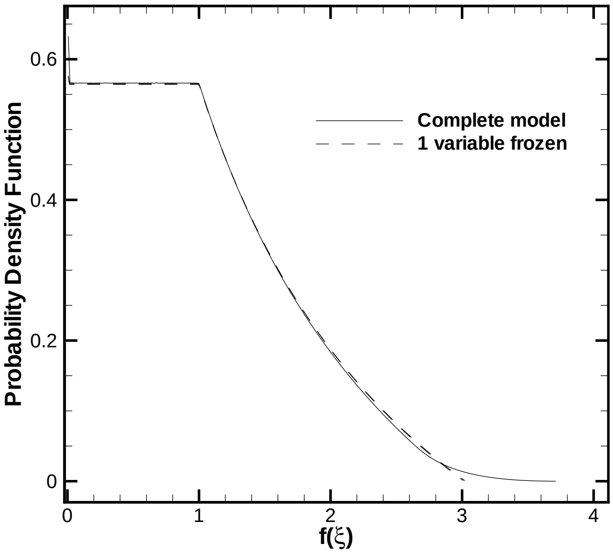


Figure 12: PDFs for the complete f_1 and the reduced models freezing the third parameter.

From a practical point of view, the error related to the truncation of the PC series is greater than the error of the truncation of the ANOVA approximation at a certain order. This is due to the approximation of each single term of the ANOVA expansion via a truncated polynomial series so the ANOVA functional decomposition contains only 2^d terms while they are approximated by a finite PC series. This means that the PC approximation up to a certain order t could be considered a good approximation to the ANOVA functional expansion up to the order t only if the series approximating each single function in the ANOVA is well converged.

For this reason, even if the analysis is based on the information presented in the previous section and based on the computation of the sensitivity indexes by a PC approximation, the reduced model are computed analytically. This represent the best case scenario, since it relies on a perfect knowledge of the reduced model. However, the analytical knowledge of the model represent the case where the differences between the complete and reduced model are minimized. To obtain an equivalent result for a generic function a non truncated PC series would be necessary.

The first example considered is the linear g-function (49). Results presented in Section §6.2 in terms of sensitivity indexes (see figure 6) and total sensitivity indexes (see table 2) illustrate some main features: the first order interaction seems to be enough to represent the model; the contributions related to the first order interactions exceeds 0.8 for the variance, but it is much more reduced for both skewness and kurtosis. Two different reduced model are considered in this case: the first-order model f_{O1} and the second-order one f_{O2}, described by the following equations

$$\begin{aligned} f_{O1} &= f_0 + f_1(\xi_1) + f_2(\xi_2) + f_3(\xi_3) + f_4(\xi_4) \\ f_{O2} &= f_{O1} + f_{12}(\xi_1,\xi_2) + f_{13}(\xi_1,\xi_3) + f_{14}(\xi_1,\xi_4) + f_{23}(\xi_2,\xi_3) + f_{24}(\xi_2,\xi_4) + f_{34}(\xi_3,\xi_4) \end{aligned} \tag{55}$$

In the table 7, the error relative to the first and second order models are reported, where the models are obtained by a truncated PC series and their exact counterparts, f_{O1}^{ex} and f_{O2}^{ex} are computed analytically.

The figure 13 illustrates the PDF for the complete and the reduced model. Note that including a large amount of variance could lead to very bad metamodel if the information on the variance are not supplemented by those obtained by the analysis of high-order methods. From the table 7, it is evident that even if the variance related to the first order terms exceeds the 80% of the total variance, the corresponding skewness and kurtosis is very low. The situation is evident in the probability density function associated to the reduced model f_{O1} reported in figure 13 where the pdf corresponding to f_{O1} completely looses information about the skewness (it is perfectly symmetric) and reveals to be a

Function	Variance	Skewness	Kurtosis
f_{O1}	86.46	8.02	7.83
f_{O2}	100.00	95.47	67.00
f_{O1}^{ex}	82.76	0.00	31.51
f_{O2}^{ex}	98.78	81.32	76.52

Table 7: Total contribution for the variance, skewness and kurtosis up to the first and second order (total degree 5) as computed in section §6.2 and their analytical counterparts.

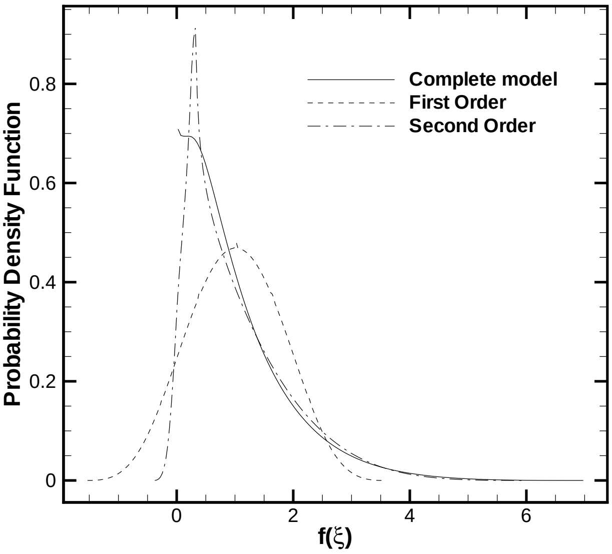


Figure 13: PDFs for the complete linear g-function f_{glin} (see equation (49)) and the reduced models f_{O1}^{ex} and f_{O2}^{ex} (see equations (55)).

very inaccurate approximation of the complete function. The situation greatly improves including contributions up to the second order of interactions between variables. This is supported by the value of the statistics reported in the table 7 where, even if the improvement in terms of variance is reduced, a better approximation of both skewness and kurtosis are achieved. This first example again demonstrates as even in the case of reduction of a model in the superposition sense, the higher-order sensitivity index can furnish useful information on the quality of reduced metamodels.

The second example is the function f_2 (see equation (50)). The results reported in the section §6.2 show than the first order terms represent more than 90% of the variance while they correspond to the 0% for the skewness and they contribute to a value inferior than 15% for the kurtosis. In this case, looking at the sensitivity indexes relative to the variance, a first order model could appear as a good approximation of the complete function. However, the quantification of the higher moments sensitivity indexes is instead very important. Considering the first and the second order defined as follows

$$\begin{aligned} f_{O1} &= f_0 + f_1(\xi_1) + f_2(\xi_2) + f_3(\xi_3) \\ f_{O2} &= f_{O1} + f_{12}(\xi_1,\xi_2) + f_{13}(\xi_1,\xi_3) + f_{23}(\xi_2,\xi_3), \end{aligned} \tag{56}$$

the computation of the pdf clearly reveals the importance of the high order terms.

In the figure 14, the pdf for the complete model and the first and second orders are reported. Even if more than 90% of the variance is included in the first order model, its pdf contains no information about the skewness and the tails appear to be totally lost. However, including the second order interactions between variables, the quality of the pdf improves a lot as it can be observed.

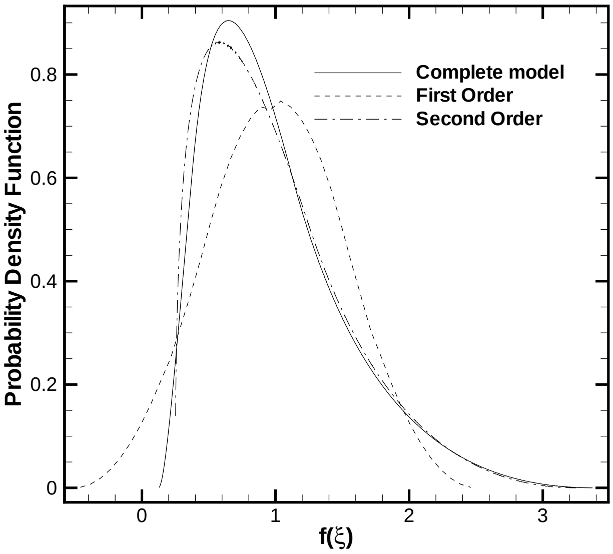


Figure 14: PDFs for the complete f_2 and the reduced models up to the first and second orders.

7. Conclusions

This paper deals with the decomposition of high-order statistics and with the importance of using this information for reducing the complexity of a stochastic problem.

First, it is illustrated how third and fourth-order statistical moments, *i.e.* skewness and kurtosis, can be decomposed. Secondly, a correlation was found between the functional decomposition, as depicted by Sobol, and the polynomial chaos development. This permitted to identify clearly each term of the decomposition, drawing also a practical way to compute all these terms. This procedure is assessed on several test-cases computing the convergence curves obtained by using PC with respect to the reference solution, that is the exact analytical one.

Moreover, sensitivity indices based on skewness and kurtosis decomposition were introduced. The importance of ranking the predominant uncertainties in terms not only of the variance but also of higher order moments (then extending the ANOVA analysis also to higher order statistic moments), was demonstrated with several functions, where all the decomposition terms can be calculated analytically.

Two different strategies for reducing the complexity of the stochastic problem are considered: i) to reduce the number of dimensions, or to reduce the order of interactions between different variables. For the proposed test-cases, the influence of different choices in terms of some simplifying assumptions, is assessed by computing the error between the global and the reduced problem. Considering high-order statistics is shown to be of fundamental importance for saving the statistics properties of the reduced problem with respect to the complete one.

Future works will be directed towards adaptive strategies for the reduction of the global computational cost, and the use of High-Order statistics in robust design optimization.

References

[1] D. Xiu, G. E. Karniadakis, The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations, SIAM Journal on Scientific Computing 24 (2002) 619–644.

[2] D. Xiu, J. S. Hesthaven, High-Order Collocation Methods for Differential Equations with Random Inputs, SIAM Journal on Scientific Computing 27 (2005) 1118.

[3] X. Wan, G. E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, Journal of Computational Physics 209 (2005) 617–642.

[4] R. G. Ghanem, P. D. Spanos, Stochastic Finite Elements. A spectral approach, Springer Verlag, 1991.

[5] F. Nobile, R. Tempone, C. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM Journal on Numerical Analysis 46 (2008) 2309–2345.

[6] J. Foo, G. E. Karniadakis, Multi-element probabilistic collocation method in high dimensions, Journal of Computational Physics 229 (2010) 1536–1557.

[7] N. H. Kim, H. Wang, N. V. Queipo, Efficient Shape Optimization Under Uncertainty Using Polynomial Chaos Expansions and Local Sensitivities, Technical Report 5, 2006.

[8] M. Eldred, Recent Advances in Non-Intrusive Polynomial Chaos and Stochastic Collocation Methods for Uncertainty Analysis and Design , AIAA Paper 2009-2274 (2009) –37.

[9] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Soviet Math. Dokl. (1963) 240–243.

[10] O. Le Maître, Uncertainty propagation using WienerHaar expansions, Journal of Computational Physics 197 (2004) 28–57.

[11] O. Le Maître, O. Knio, H. Najm, R. Ghanem, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, Journal of Computational Physics 197 (2004) 502–531.

[12] I. Babuška, R. Tempone, G. Zouraris, Galerkin finite element approximations of stochastic elliptic differential equations, SIAM Journal on Numerical Analysis 42 (2004) 800–825.

[13] M. H. Keese A., Numerical methods and Smolyak quadrature for nonlinear partial differential equations, Informatikbericht 2003-5 (2003).

[14] E. Novak, K. Ritter, Simple cubature formulas with high polynomial exactness, Constr. Approx. (1999) 499–522.

[15] V. Barthelmann, E. Novak, K. Ritter, High dimensional polynomial interpolation on sparse grids, Adv. Comput. Math., Vol. 12, No. 4 (2000) 273–288.

[16] E. Novak, K. Ritter, High dimensional integration of smooth functions over cubes, Numerische Mathematik (1996) 79–97.

[17] K. Petras, On the Smolyak cubature error for analytic functions, Advances in Computational Mathematics 12 (2000) 71–93.

[18] E. Borgonovo, G. E. Apostolakis, S. Tarantola, A. Saltelli, Comparison of global sensitivity analysis techniques and importance measures in PSA, Reliability Engineering & System Safety 79 (2003) 175–185.

[19] I. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and Computers in Simulation 55 (2001) 271–280.

[20] H. Rabitz, O. Alis, J. Shorter, K. Shim, Efficient input-output model representations., Comput. Phys. Commun. 117 (1999) 11–20.

[21] A. Saltelli, Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index., Comput. Phys. Commun. 181 (2010) 259–270.

[22] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering and System Safety 93 (2008) 964–979.

[23] T. Crestaux, O. Le Maître, J.-M. Martinez, Polynomial chaos expansion for sensitivity analysis, Reliability Engineering & System Safety 94 (2009) 1161–1172.

[24] G. Blatman, B. Sudret, A comparison of three metamodel-based methods for global sensitivity analysis: GP modelling, HDMR and LAR-gPC, Procedia - Social and Behavioral Sciences 2 (2010) 7613–7614.

[25] X. Yanga, M. Choi, G. Lin, G. E. Karniadakis, Adaptive ANOVA Decomposition of Stochastic Incompressible and Compressible Flows, Journal of Computational Physics 231 (2012) 1587–1614.

[26] G. Blatman, B. Sudret, Efficient computation of global sensitivity indices using sparse polynomial chaos expansions, Reliability Engineering and System Safety 95 (2010) 1216–1229.

[27] C. Xu, G. Gertner, Extending a global sensitivity analysis technique to models with correlated parameters., Computational Statistics & Data Analysis 51 (2007) 5579–5590.

[28] C. Xu, G. Gertner, Uncertainty and sensitivity analysis for models with correlated parameters., Reliability Engineering & System Safety 93 (2008) 1563–1573.

[29] E. Borgonovo, A new uncertainty importance measure, Reliability Engineering & System Safety 92 (2007) 771–784.

[30] J. Y. Caniou, B. Sudret, Distribution-based global sensitivity analysis in case of correlated input parameters using polynomial chaos expansions, in: ICASP2011.

[31] E. Borgonovo, W. Castaings, S. Tarantola, Model emulation and moment-independent sensitivity analysis: An application to environmental modelling, Environmental Modelling & Software 34 (2012) 105–115.

[32] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, in: Proc. 5th Int. Conf. on Comp. Stoch. Mech (CSM5), Rhodos, Greece.

[33] B. Sudret, Uncertainty propagation and sensitivity analysis in mechanical models Contributions to structural reliability and stochastic spectral methods, Habilitation à Diriger des Recherches, Université BLAISE PASCAL - Clermont II (2007) 1–252.

[34] E. Plischke, E. Borgonovo, C. L. Smith, Global sensitivity measures from given data, European Journal of Operational Research 226 (2013) 536–550.

[35] Z. Gao, J. S. Hesthaven, Efficient solution of ordinary differential equations with high-dimensional parametrized uncertainty, Communications in computational physics (2010) 1–33.

Appendix A. Definition of High-order statistics

This section illustrates how statistics (of order n) of f can be computed from the conditional expectancy of n-powers of f. First, let us consider the definition of the variance

$$\sigma^2 = \int_{\Xi^d} (f(\boldsymbol{\xi}) - E(f))^2 p(\boldsymbol{\xi}) \mathrm{d}\xi. \tag{A.1}$$

As a consequence, it can be easily computed that

$$\sigma^2 = E(f^2) - E(f)^2. \tag{A.2}$$

In the same way, starting from the definition of the skewness, the following formula can be obtained

$$\begin{aligned} s &= \int_{\Xi^d} (f(\boldsymbol{\xi}) - E(f))^3 p(\boldsymbol{\xi}) \mathrm{d}\xi \\ &= \int_{\Xi^d} \left(f^3(\boldsymbol{\xi}) - 3f^2(\boldsymbol{\xi})E(f) + 3f(\boldsymbol{\xi})E(f)^2 - E(f)^3 \right) p(\boldsymbol{\xi}) \mathrm{d}\xi \\ &= E(f^3) - 3E(f^2)E(f) + 3E(f)E(f)^2 - E(f)^3 \\ &= E(f^3) - 3E(f^2)E(f) + 2E(f)^3. \end{aligned} \tag{A.3}$$

This means that skewness, as defined in Eq. A.3, depend only on the expected values of the function f, f^2 and f^3. Using the formula for $E(f^2)$ obtained from Eq. A.2, equation A.3 becomes

$$\begin{aligned} s &= E(f^3) - 3E(f^2)E(f) + 2E(f)^3 \\ &= E(f^3) - 3\sigma^2 E(f) - 3E(f)^3 + 2E(f)^3 \\ &= E(f^3) - 3\sigma^2 E(f) - E(f)^3. \end{aligned} \tag{A.4}$$

Following the same procedure, kurtosis can first be written as follows

$$\begin{aligned} k &= \int_{\Xi^d} (f(\boldsymbol{\xi}) - E(f))^4 p(\boldsymbol{\xi}) \mathrm{d}\xi \\ &= \int_{\Xi^d} \left(f^4(\boldsymbol{\xi}) - 4f^3(\boldsymbol{\xi})E(f) + 6f(\boldsymbol{\xi})^2 E(f)^2 - 4E(f)^3 f(\boldsymbol{\xi}) + E(f)^4 \right) p(\boldsymbol{\xi}) \mathrm{d}\xi \\ &= E(f^4) - 4E(f^3)E(f) + 6E(f^2)E(f)^2 - 4E(f)^4 + E(f)^4 \\ &= E(f^4) - 4E(f^3)E(f) + 6E(f^2)E(f)^2 - 3E(f)^4. \end{aligned} \tag{A.5}$$

Then, using the value of $E(f)^3$ obtained from Eq. A.4 and the value of $E(f^2)$ from Eq. A.2, kurtosis can be computed as follows

$$k = E(f^4) - 4E(f)s - 6\sigma^2 E(f)^2 - E(f)^4. \tag{A.6}$$

Appendix B. Third central moment expression for the ANOVA functional decomposition

In this section, the final expression of the skewness as reported in (19), is computed using the expression obtained via the multinomial theorem applied to the functional ANOVA decomposition (16).

Equation (16), *i.e.*

$$s = \sum_{p=1}^{N} \int_{\hat{\Xi}_p} f_{\boldsymbol{m}_p}^3 \mathrm{d}\mu_p + 3 \sum_{p=1}^{N} \sum_{\substack{q=1 \\ q \neq p}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m}_p}^2 f_{\boldsymbol{m}_q} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m}_p} f_{\boldsymbol{m}_q} f_{\boldsymbol{m}_r} \mathrm{d}\mu_{pqr},$$

displays interaction between variables and sub-sets of variables for the third central moment of the function decomposed via the multinomial theorem. In particular, the second and the third term of (16) could be simplified to highlight the contributions that are always equal to zero.

The second term of the right hand side of the previous equation can be simplified as follows

$$3 \sum_{p=1}^{N} \sum_{\substack{q=1 \\ q \neq p}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m}_p}^2 f_{\boldsymbol{m}_q} \mathrm{d}\mu_{pq} = 3 \sum_{\boldsymbol{m}_p} \sum_{\boldsymbol{m}_q \subset \boldsymbol{m}_p} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m}_p}^2 f_{\boldsymbol{m}_q} \mathrm{d}\mu_{pq}. \tag{B.1}$$

Proof. This term presents the interaction between two multi-indexes $\boldsymbol{m_p}$ and $\boldsymbol{m_q}$, where one of them is raised to the second power. The two multi-indexes should be different for construction: if $\boldsymbol{m_q}$ is not a subset of $\boldsymbol{m_p}$ then a set of coordinates $\boldsymbol{m_q} - \cap_{pq}$ belong only to $\boldsymbol{m_q}$ (if the set is totally disjointed, the term $\cap_{pq}$ is the null set). Note that the symbol $\cap_{pq}$ indicate the coordinates contained in both the multi-indexes p and q.

The integral can be than reformulated into the form

$$\int_{\hat{\Xi}_{pq}} f^2_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} = \int_{\hat{\Xi}_{p+\cap_{pq}}} f^2_{\boldsymbol{m_p}} \left(\int_{\hat{\Xi}_{q-\cap_{pq}}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_{q-\cap_{pq}} \right) \mathrm{d}\mu_{p+\cap_{pq}} = 0, \tag{B.2}$$

where the internal integral is equal to zero due to the orthogonality of the ANOVA contributions (see equation (8)). □

Now, the case of the identification of the contributions to a specific multi-index $\boldsymbol{m_i}$ is addressed, where $\boldsymbol{m_{pq}} = \boldsymbol{m_i}$ must hold. If all the sub-sets of variables and their interactions of $\boldsymbol{m_i}$ are collected in the set $\mathcal{P}_i$, the contributions to $\boldsymbol{m_i}$ are a sub set of the $2^{|\mathcal{P}_i|} - 1$ simple combinations of class two. For instance, in the multi-index $\boldsymbol{m_i} = (1, 1, 1)$, even if the two subset $(1, 0, 0)$ and $(1, 1, 0)$ are contained in $\mathcal{P}_i$, their interactions does not contribute to the conditional term of the skewness $s_{\boldsymbol{m_i}}$, *i.e.* $(1, 0, 0) \boxplus (1, 1, 0) \neq \boldsymbol{m_i}$. Requirements $\boldsymbol{m_q} \subset \boldsymbol{m_p}$ and $\boldsymbol{m_{pq}} = \boldsymbol{m_i}$ allow to identify the only non-null contributions as follows

$$3 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_{pq}} = \boldsymbol{m_i} \\ \mathcal{P}_i \ni \boldsymbol{m_q} \subset \boldsymbol{m_p}}} \int_{\hat{\Xi}_{pq}} f^2_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} = 3 \int_{\hat{\Xi}_i} f^2_{\boldsymbol{m_i}} \sum_{\boldsymbol{m_q} \in \mathcal{P}_{i,\neq}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i, \tag{B.3}$$

where $\mathcal{P}_{i,\neq}$ is employed as shorthand for $\mathcal{P}_{i,\neq} = \mathcal{P}_i - \{\boldsymbol{m_i}\}$.

The last term of (16) can be written as follows

$$6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} = 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{\substack{r=q+1 \\ \boldsymbol{m_{pq}} = \boldsymbol{m_r}}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pq}. \tag{B.4}$$

Proof. This case can be demonstrated (extending what already done for the dyadic interaction between multi-indexes) easily as follows

$$\int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} = \int_{\hat{\Xi}_{pq+\cap pqr}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \left(\int_{\hat{\Xi}_{r-\cap pqr}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{r-\cap_{pqr}} \right) \mathrm{d}\mu_{pq+\cap_{pqr}} = 0, \tag{B.5}$$

by using the orthogonality property. □

If a specific index $\boldsymbol{m_i}$ is of interest, the conditional contribution is identified requiring $\boldsymbol{m_{pqr}} = \boldsymbol{m_i}$ (then to obtain only non null contribution, $\boldsymbol{m_{pq}} = \boldsymbol{m_r} = \boldsymbol{m_i}$ should be considered):

$$s_{\boldsymbol{m_i}} = 6 \sum_{\boldsymbol{m_p} \neq \boldsymbol{m_i}} \sum_{\substack{\boldsymbol{m_p} \neq \boldsymbol{m_q} \neq \boldsymbol{m_i} \\ \boldsymbol{m_{pq}} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_i}} \mathrm{d}\mu_i. \tag{B.6}$$

In conclusion, the final form for the skewness is equal to

$$s = \sum_{p=1}^{N} \int_{\hat{\Xi}_p} f^3_{\boldsymbol{m_p}} \mathrm{d}\mu_p + 3 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_q} \subset \boldsymbol{m_p}} \int_{\hat{\Xi}_{pq}} f^2_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{\substack{r=q+1 \\ \boldsymbol{m_{pq}} = \boldsymbol{m_r}}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pq},$$

where each conditional contribution is as follows

$$s_{\boldsymbol{m_i}} = \int_{\hat{\Xi}_i} f^3_{\boldsymbol{m_i}} \mathrm{d}\mu_i + 3 \int_{\hat{\Xi}_i} f^2_{\boldsymbol{m_i}} \sum_{\boldsymbol{m_q} \in \mathcal{P}_{i,\neq}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i + 6 \sum_{\boldsymbol{m_p} \neq \boldsymbol{m_i}} \sum_{\substack{\boldsymbol{m_p} \neq \boldsymbol{m_q} \neq \boldsymbol{m_i} \\ \boldsymbol{m_{pq}} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_i}} \mathrm{d}\mu_i.$$

Appendix C. Fourth central moment expression for the ANOVA functional decomposition

In this section, the equation (24) is obtained starting from the equation (22), *i.e.*

$$
\begin{aligned}
k = {} & \sum_{p=1}^{N} \int_{\hat{\Xi}_p} f_{\boldsymbol{m_p}}^4 \mathrm{d}\mu_p + 4 \sum_{p=1}^{N} \sum_{\substack{q=1\\ q\neq p}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^3 f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}}^2 \mathrm{d}\mu_{pq} \\
& + 12 \sum_{p=1}^{N} \sum_{\substack{q=1\\ q\neq p}}^{N} \sum_{\substack{r=j+1\\ r\neq p}}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} + 24 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \sum_{t=r+1}^{N} \int_{\hat{\Xi}_{pqrt}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_{pqrt},
\end{aligned}
\tag{C.1}
$$

after having identified the orthogonal contributions, *i.e.* the terms always equal to zero.

Looking at the kurtosis, the first three terms are easy to handle: the first and the third on the right side of the previous equation cannot be further simplified, while the second one can be analyzed as already done for the similar skewness term. It is then possible to write

$$
4 \sum_{p=1}^{N} \sum_{\substack{q=1\\ q\neq p}}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^3 f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} = 4 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_q} \subset \boldsymbol{m_p}} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^3 f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq},
\tag{C.2}
$$

as already demonstrated for the skewness term.

More attention should be devoted to the last two terms. The first one can be written as follows

$$
12 \sum_{p=1}^{N} \sum_{\substack{q=1\\ q\neq p}}^{N} \sum_{\substack{r=j+1\\ r\neq p}}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} = 12 \sum_{p=1}^{N} \sum_{\substack{q=1\\ q\neq p}}^{N} \sum_{\substack{r=q+1\\ \boldsymbol{m_{qr}} \setminus \cap_{qr} \subseteq \boldsymbol{m_p}}}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr}.
\tag{C.3}
$$

Proof. It is easy to note that if the multi-indexes $\boldsymbol{m_q}$ and $\boldsymbol{m_r}$ are totally independent from the variables contained in the multi-index $\boldsymbol{m_p}$, $\int_{\hat{\Xi}_{qr}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{qr} = 0$ holds (due to the orthogonality of the ANOVA functional components). In the general case, when $\boldsymbol{m_{qr}} \cap \boldsymbol{m_p} \neq 0$, the existence of a null integral is related to the presence of variables in the multi-indexes $\boldsymbol{m_q}$ or $\boldsymbol{m_r}$ not contained in $\boldsymbol{m_p}$. If $\boldsymbol{m_{pq}} \setminus \cap_{qr} \nsubseteq \boldsymbol{m_p}$ then it is possible to write as follows

$$
\int_{\hat{\Xi}_p} f_{\boldsymbol{m_p}}^2 \left(\int_{\hat{\Xi}_{\boldsymbol{m_{pq}} \setminus \cap_{qr}}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}(\boldsymbol{m_{pq}} \setminus \cap_{qr}) \right) \mathrm{d}\mu_p = 0.
\tag{C.4}
$$

Note that the internal integral is carried out with respect to a variable contained only in one of $\boldsymbol{m_p}$ or $\boldsymbol{m_q}$, then is always zero due to orthogonality. Obviously, the case with the subsets related to $\boldsymbol{m_{qr}}$ and $\boldsymbol{m_p}$ totally disjointed, is included in the previous condition. □

If the specific multi-index $\boldsymbol{m_i}$ is provided, then in this case the contribution of this term is computed as follows

$$
12 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_r} \in \mathcal{P}_i, r>q \\ \boldsymbol{m_p} \boxplus \cap_{qr} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_i.
\tag{C.5}
$$

Proof. The previous equation can be obtained considering the requirements $\boldsymbol{m_{pqr}} = \boldsymbol{m_i}$ and $\boldsymbol{m_{qr}} \setminus \cap_{qr} \subseteq \boldsymbol{m_p}$. It is easy to verify that if the second equation is true, then it must follow $\boldsymbol{m_p} \boxplus \boldsymbol{m_{qr}} \setminus \cap_{qr} \subseteq \boldsymbol{m_p} \boxplus \boldsymbol{m_p} = \boldsymbol{m_p}$, from which $\boldsymbol{m_i} \setminus \cap_{qr} \subseteq \boldsymbol{m_p}$. Finally, $\boldsymbol{m_i} = \boldsymbol{m_p} \boxplus \cap_{qr}$ holds (the equality sign follows from $\boldsymbol{m_p}$, $\boldsymbol{m_q}$ and $\boldsymbol{m_r} \in \mathcal{P}_i$). Remark that great attention must be paid in manipulating expressions with the summation of multi-indexes $\boxplus$. Generally, consider that $\boldsymbol{m_p} \setminus \boldsymbol{m_q} = \boldsymbol{m_r} \Rightarrow \boldsymbol{m_p} = \boldsymbol{m_r} \boxplus \boldsymbol{m_q}$ holds but the contrary is not guaranteed. □

The last term of the kurtosis can be also reformulated as follows

$$
24 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \sum_{t=r+1}^{N} \int_{\hat{\Xi}_{pqrt}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_{pqrt} = 24 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \sum_{\substack{t=r+1 \\ \boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}}}^{N} \int_{\hat{\Xi}_{pqrt}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_{pqrt}.
\tag{C.6}
$$

Proof. This term can be obtained with some constraints on the multi-indexes: they should share, two by two, some sets of coordinates: $\boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}}$ and $\boldsymbol{m_{rt}} \setminus \cap_{rt} \subseteq \boldsymbol{m_{pq}}$. The fulfillment of the previous conditions assures that the integral cannot be divided into a product between integrals of orthogonal contributions. These conditions could be applied choosing randomly two couple of indexes. Using the second constraint, $\boldsymbol{m_{rt}} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}$ is obtained. Using this relation with the first requirement, $\boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}$ holds. □

If a set of variables is specified by using the multi-index $\boldsymbol{m_i}$, then the conditional contribution that arises from the previous term can be identified as follows

$$24 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\boldsymbol{m_q} \in \mathcal{P}_i, q>p} \sum_{\boldsymbol{m_r} \in \mathcal{P}_i, r>q} \sum_{\substack{t>r, \boldsymbol{m_r} \in \mathcal{P}_i \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt} \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{rt}} \boxplus \cap_{pq}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_i. \tag{C.7}$$

Proof. As already shown, the set of all the possible sub-sets of variables relative to a multi-index $\boldsymbol{m_i}$ is represented by $\mathcal{P}_i$, so all the contributions should belong to $\mathcal{P}_i$. However, not all the possible combinations of four elements selected from the set $\mathcal{P}_i$ are relative to the multi-index $\boldsymbol{m_i}$; a first condition is to require that $\boldsymbol{m_{pqrt}} = \boldsymbol{m_i}$. From the previous proof, it is clear that only the non-null elements need to satisfy the two requirements $\boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}}$ and $\boldsymbol{m_{rt}} \setminus \cap_{rt} \subseteq \boldsymbol{m_{pq}}$. If the two latter requirements are manipulated as $\boldsymbol{m_{rt}} \boxplus \boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} \boxplus \boldsymbol{m_{rt}}$ and $\boldsymbol{m_{pq}} \boxplus \boldsymbol{m_{rt}} \setminus \cap_{rt} \subseteq \boldsymbol{m_{pq}} \boxplus \boldsymbol{m_{pq}}$, the following conditions can be written

$$\begin{aligned} \boldsymbol{m_i} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} &\Rightarrow \boldsymbol{m_i} \subseteq \boldsymbol{m_{rt}} \boxplus \cap_{pq} \\ \boldsymbol{m_i} \setminus \cap_{rt} \subseteq \boldsymbol{m_{pq}} &\Rightarrow \boldsymbol{m_i} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}. \end{aligned} \tag{C.8}$$

□

In conclusion, summing up all the contributions, the final form for the kurtosis can be written as

$$\begin{aligned} k = {} & \sum_{p=1}^{N} \int_{\hat{\Xi}_p} f_{\boldsymbol{m_p}}^4 \mathrm{d}\mu_p + 4 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_q} \subset \boldsymbol{m_p}} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^3 f_{\boldsymbol{m_q}} \mathrm{d}\mu_{pq} + 6 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \int_{\hat{\Xi}_{pq}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}}^2 \mathrm{d}\mu_{pq} \\ & + 12 \sum_{p=1}^{N} \sum_{\substack{q=1 \\ q \neq p}}^{N} \sum_{\substack{r=q+1 \\ \boldsymbol{m_{qr}} \setminus \cap_{qr} \subseteq \boldsymbol{m_p}}}^{N} \int_{\hat{\Xi}_{pqr}} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_{pqr} + 24 \sum_{p=1}^{N} \sum_{q=p+1}^{N} \sum_{r=q+1}^{N} \sum_{\substack{t=r+1 \\ \boldsymbol{m_{pq}} \setminus \cap_{pq} \subseteq \boldsymbol{m_{rt}} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt}}}^{N} \int_{\hat{\Xi}_{pqrt}} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_{pqrt}, \end{aligned} \tag{C.9}$$

where each conditional contribution, with respect to a fixed multi-index $\boldsymbol{m_i}$, is equal to

$$\begin{aligned} k_{\boldsymbol{m_i}} = {} & \int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}}^4 \mathrm{d}\mu_i + 4 \int_{\hat{\Xi}_i} f_{\boldsymbol{m_i}}^3 \sum_{\boldsymbol{m_q} \in \mathcal{P}_{i,\neq}} f_{\boldsymbol{m_q}} \mathrm{d}\mu_i + 6 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_i \\ \boldsymbol{m_{pq}} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}}^2 \mathrm{d}\mu_i \\ & + 12 \sum_{\boldsymbol{m_p}} \sum_{\boldsymbol{m_p} \neq \boldsymbol{m_q} \in \mathcal{P}_i} \sum_{\substack{\boldsymbol{m_r} \in \mathcal{P}_i, r>q \\ \boldsymbol{m_p} \boxplus \cap_{qr} = \boldsymbol{m_i}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}}^2 f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} \mathrm{d}\mu_i \\ & + 24 \sum_{\boldsymbol{m_p} \in \mathcal{P}_i} \sum_{\boldsymbol{m_q} \in \mathcal{P}_i, q>p} \sum_{\boldsymbol{m_r} \in \mathcal{P}_i, r>q} \sum_{\substack{t>r, \boldsymbol{m_r} \in \mathcal{P}_i \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{pq}} \boxplus \cap_{rt} \\ \boldsymbol{m_i} \subseteq \boldsymbol{m_{rt}} \boxplus \cap_{pq}}} \int_{\hat{\Xi}_i} f_{\boldsymbol{m_p}} f_{\boldsymbol{m_q}} f_{\boldsymbol{m_r}} f_{\boldsymbol{m_t}} \mathrm{d}\mu_i. \end{aligned} \tag{C.10}$$

Appendix D. Skewness from the PC expansion

In this section, the final form for the skewness relying on the PC series expansion is presented. By applying the multinomial theorem, the skewness can be written as a sum of contributions generated by the interactions of the polynomial basis components. This yields (equal to (35))

$$\begin{aligned} s &= \int_{\Xi} \left(f(\boldsymbol{\xi}) - \beta_0\right)^3 p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\Xi} \left(\sum_{p=1}^{P} \beta_p \Psi_p(\boldsymbol{\xi}) \right)^3 p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \\ &= \sum_{p=1}^{P} \beta_p^3 \langle \Psi_p^3(\boldsymbol{\xi})\rangle + 3 \sum_{p=1}^{P} \beta_p^2 \sum_{\substack{q=1\\ q\neq p}}^{P} \beta_q \langle \Psi_p^2(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\rangle + 6 \sum_{p=1}^{P} \sum_{q=p+1}^{P} \sum_{r=q+1}^{P} \beta_p\beta_q\beta_r \langle \Psi_p(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi})\rangle. \end{aligned} \tag{D.1}$$

Looking at this equation, it seems that no orthogonal contributions are present, because the interactions involve only polynomial forms raised to a power higher than one or *triadic* interaction. However, the second and third terms should be further investigated.

Following from the definition of each polynomial term (27), the product between two polynomial terms of the basis, where the first one is raised to the power n with $1 < n \in \mathbb{N}$, can be written as follows

$$\begin{aligned} \langle \Psi_p^n(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\rangle &= \int_{\Xi} \left(\prod_{i=1}^{d} \psi^n_{\alpha_i^p}(\xi_i) \right) \left(\prod_{i=1}^{d} \psi_{\alpha_i^q}(\xi_i) \right) p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \\ &= \int_{\Xi} \left(\prod_{i=1}^{d} \psi^n_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i) \right) p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \prod_{i=1}^{d} \int_{\Xi_i} \psi^n_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i)\, p(\xi_i)\mathrm{d}\xi_i. \end{aligned} \tag{D.2}$$

Due to the orthogonality of the PC basis with respect to $\Psi_0 = 1$, it follows that if $\alpha_i^p = 0$ then the integral with respect to the variable ξ_i becomes

$$\int_{\Xi} \psi_{\alpha_i^q}(\xi_i)\, p(\xi_i)\mathrm{d}\xi_i = 0 \quad \text{for} \quad \alpha_i^q \neq 0. \tag{D.3}$$

From this relation, the orthogonality of $\langle \Psi_p^n(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\rangle$ follows. The non-null existence of the corresponding skewness (and kurtosis term) can be efficiently identified by means of the function Δ_q^p defined in §4.2.

The third term of the skewness from the PC series involves *triadic* interaction of polynomial terms raised to the power one

$$\langle \Psi_p(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi})\rangle = \int_{\Xi} \left(\prod_{i=1}^{d} \psi_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i)\, \psi_{\alpha_i^r}(\xi_i) \right) p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \prod_{i=1}^{d} \int_{\Xi_i} \psi_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i)\, \psi_{\alpha_i^r}(\xi_i)\, p(\xi_i)\mathrm{d}\xi_i. \tag{D.4}$$

The last equation suggest that the term can be analyzed after the inspection of the relative multi-indexes $\boldsymbol{m_p}$, $\boldsymbol{m_q}$ and $\boldsymbol{m_r}$. If the sum of the respective components of the multi-indexes, *i.e.* $\boldsymbol{m_{p_i}} + \boldsymbol{m_{q_i}} + \boldsymbol{m_{r}}_i$, is equal to zero, then the variable is not present and no information can be obtained (the previous integral would be equal to 1 in such a case). If the sum $\boldsymbol{m_{p_i}} + \boldsymbol{m_{q_i}} + \boldsymbol{m_{r}}_i$ is equal to 1, this means that the variable is present in only one polynomial term between ψ_p, ψ_q and ψ_r, while it should not be present in the others (the relative coefficient $\alpha_i = 0$). This leads to a null integral due to the orthogonality of the basis with respect to the probability density function. However, another possibility can be associated to a null integral: if the sum $\boldsymbol{m_{p_i}} + \boldsymbol{m_{q_i}} + \boldsymbol{m_{r}}_i = 2$, the orthogonality between two polynomial terms guarantees that the integral is zero. The previous results can be resumed in the function Δ_{pqr} introduced in §4.2.

Appendix E. Kurtosis from the PC expansion

In this section, as already shown for the skewness in §Appendix D, the kurtosis structure relying on the PC series is described. By applying the multinomial theorem to the PC series expansion, the kurtosis is computed as follows

$$\begin{aligned} k &= \sum_{p=1}^{P} \beta_p^4 \langle \Psi_p^4(\boldsymbol{\xi})\rangle + 4 \sum_{p=1}^{P} \beta_p^3 \sum_{\substack{q=1\\ q\neq p}}^{P} \beta_q \langle \Psi_p^3, \Psi_q\rangle + 6 \sum_{p=1}^{P} \beta_p^2 \sum_{q=p+1}^{P} \beta_q^2 \langle \Psi_p^2, \Psi_q^2\rangle \\ &+ 12 \sum_{p=1}^{P} \beta_p^2 \sum_{\substack{q=1\\ q\neq p}}^{P} \beta_q \sum_{\substack{r=q+1\\ r\neq p}}^{P} \beta_r \langle \Psi_p^2, \Psi_q\Psi_r\rangle + 24 \sum_{p=1}^{P} \sum_{q=p+1}^{P} \sum_{r=q+1}^{P} \sum_{t=r+1}^{P} \beta_p\beta_q\beta_r\beta_t \langle \Psi_p\Psi_q, \Psi_r\Psi_t\rangle. \end{aligned} \tag{E.1}$$

The first, the second and the third terms are easy to handle because they cannot be identified as null (first and third) or has been already analyzed (second) in the case of the skewness.

The terms with a different structure with respect to the terms existing in the variance or in the skewness are the last two ones. The first contains the interaction between three polynomial terms, where the first of them is raised to the second power. In the general case of $1 < n \in \mathbb{N}$, it holds that

$$\langle \Psi_p^n(\boldsymbol{\xi}), \Psi_q(\boldsymbol{\xi})\Psi_r(\boldsymbol{\xi})\rangle = \int_{\Xi} \left(\prod_{i=1}^{d} \psi^n_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i)\, \psi_{\alpha_i^r}(\xi_i) \right) p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \prod_{i=1}^{d} \int_{\Xi_i} \psi^n_{\alpha_i^p}(\xi_i)\, \psi_{\alpha_i^q}(\xi_i)\, \psi_{\alpha_i^r}(\xi_i)\, p(\xi_i)\mathrm{d}\xi_i. \tag{E.2}$$

From the last equation, it is possible to note that the orthogonality between polynomial term can be advocated if the term raised to the $n-$th power is zero. If $\alpha_i^p = 0$ and the sum of the remaining terms $\boldsymbol{m}_{\boldsymbol{q}i} + \boldsymbol{m}_{\boldsymbol{r}i} \neq 0$, then a null integral exist irrespectively of the coefficients α_i^p and α_i^q

$$\int_{\Xi_i} \psi_{\alpha_i^q}(\xi_i)\, \psi_{\alpha_i^r}(\xi_i)\, p(\xi_i)\mathrm{d}\xi_i = 0. \tag{E.3}$$

Note that the function Δ_{qr}^p has been introduced in §4.3.

The last term of the kurtosis expansion involves the interaction of four polynomials terms. This case represent an extension of the term already analyzed for the skewness (see Appendix D) where the interaction between three polynomial terms has been discussed. Even in the case of interaction between four polynomial terms, by inspecting the sum of the coefficients of the multi-indexes $\boldsymbol{m_p}$, $\boldsymbol{m_q}$, $\boldsymbol{m_r}$ and $\boldsymbol{m_t}$, it is possible to determine if the integral are always equal to zero. In particular, if the sum relative to the $i-$th coordinates, *i.e.* $\boldsymbol{m}_{\boldsymbol{p}i} + \boldsymbol{m}_{\boldsymbol{q}i} + \boldsymbol{m}_{\boldsymbol{r}i} + \boldsymbol{m}_{\boldsymbol{t}i}$, is equal to 1 or 2, the orthogonal properties of each terms with respect to the pdf or a couple of them with respect to the pdf can be employed to identify a non null integral (this is true irrespectively of the values of the α_i^k coefficients). This result has been used in the section §4.3 to define the function Δ_{pqrt}.