# Quantification of Uncertainties in Compressible Flows

Von der Fakultät für Mathematik und Physik der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

Fabian Meyer

aus Mutlangen

| | |
|---|---|
| Hauptberichter: | Prof. Dr. Christian Rohde |
| Mitberichter: | Prof. Dr. Andrea Barth |
| | Prof. Dr. Mária Lukácová-Medvidová |

Tag der mündlichen Prüfung: 17.12.2019

Institut für Angewandte Analysis und Numerische Simulation der Universität Stuttgart

2019

*To Miriam.*

# Acknowledgements

This thesis has benefited greatly from the support of many people some of whom I would sincerely like to thank. At first I would like to thank my supervisor Prof. Dr. Christian Rohde. Thank you for giving me the opportunity to work on the research project “SEAL” and moreover, to give me the chance to be part of your research team. I would like to thank Prof. Dr. Andrea Barth and Prof. Dr. Mária Lukácová-Medvidová for reviewing this thesis as referee. I would also like to thank my “SEAL”-colleagues, Andrea Beck, Jakob Dürrwächter, Thomas Kuhn and Claus-Dieter Munz from the IAG. Special thanks go to Jakob and Thomas, it has been a pleasure working with you. I really enjoyed our joint discussions and coding camps throughout the last three years. Furthermore, I would like to thank all my present and former colleagues from IANS for the great working atmosphere. I am especially grateful to Prof. Dr. Jan Giesselmann, firstly for our fruitful collaboration and secondly for proofreading parts of this thesis. Moreover, I would like to thank Jakob Dürrwächter and Lukas Ostrowski who have also proofread parts of this thesis.

The work resulting from this thesis would not have been possible without the financial support from the Baden-Württemberg Stiftung within the project “BW-HPC2: SEAL”, for which I am very grateful.

I am grateful to my family which supported me throughout my whole studies, starting from my Bachelor degree up to this Ph.D. thesis. Above all I would like to thank Miriam. For your understanding, patience and never-ending support, even though I spent many nights and weekends only with mathematics. Your love carried me through the last years. Thank you.

# Abstract

Due to rising computing capacities, including and accounting for uncertain (model) parameters in numerical simulations is becoming more and more popular. Uncertainty Quantification (UQ) addresses this issue and provides a variety of different mathematical methods to quantify the influence of uncertain input parameters on numerical solutions and derived quantities of interest. This thesis is concerned with the development and improvement of different UQ methods for numerical simulations of compressible flow problems, described by random conservation laws like the compressible Euler or Navier–Stokes equations. We distinguish between polynomial-based (non-statistical) UQ methods and sampling-based (statistical) UQ methods.

The first part of this thesis investigates non-statistical UQ methods, in particular the Stochastic Galerkin (SG), Non-Intrusive Spectral Projection (NISP) and Stochastic Collocation (SC) method. While SG is a frequently used method for UQ of random partial differential equations, the classical SG approach is not ensured to preserve hyperbolicity of the underlying random hyperbolic conservation law. To this end we develop a hyperbolicity-preserving numerical scheme, which uses a slope limiter to retain admissible solutions of the SG system, while providing high-order approximations in physical and random space. The modified numerical scheme is applied to different challenging numerical examples for which the classical SG approach fails.

An important aspect when considering space-time-stochastic numerical schemes is to quantify the errors that arise from numerical discretization. In this thesis we derive a novel a posteriori error analysis framework for numerical discretizations of random hyperbolic systems of conservation laws, which rely on the Runge–Kutta Discontinuous Galerkin method in combination with polynomial-based UQ methods. Our estimates are based on the relative entropy framework of Dafermos and DiPerna [26] and allow us to quantify the entire space-time-stochastic discretization error. Moreover, due to a splitting of the residual we are able distinguish between spatio-temporal and stochastic errors. Based on the a posteriori error estimates we design novel residual-based, space-stochastic adaptive numerical schemes. We confirm our theoretical findings by various numerical experiments.

The last part of this thesis is concerned with statistical UQ methods, especially Monte Carlo (MC) type methods. We extend the Multilevel Monte Carlo (MLMC) method to what we call $hp$-MLMC method. Instead of considering a hierarchy of spatially refined meshes, we allow for meshes which are arbitrarily $hp$-refined. The classical complexity analysis of MLMC is extended to the $hp$-MLMC method. Moreover, to increase the robustness and efficiency of an iterative version of $hp$-MLMC, we construct a confidence interval for the optimal number of samples per level. To demonstrate the efficiency of the $hp$-MLMC method combined with

the novel sample estimator we apply our method to two different compressible flow problems described by the random Navier–Stokes equations. In particular, we consider an important problem from computational acoustics that exhibits physical phenomena with high sensitivity with respect to the problem parameters and which poses a challenging problem for UQ.

# Zusammenfassung

Aufgrund stetig steigender Rechenkapazität spielen numerische Simulationsmethoden, welche zusätzlich den Einfluss unsicherer Eingabeparameter quantifizieren, eine immer größere Rolle. Das Forschungsgebiet der Unsicherheitsquantifizierung widmet sich dieser Aufgabe und bietet eine Vielzahl mathematischer Methoden an, welche es ermöglichen den Einfuss unsicherer Parameter auf numerische Lösungen und abgeleiteter Größen zu quantifizieren. Diese Arbeit befasst sich mit der Entwicklung und Erweiterung von Methoden zur Unsicherheitsquantifizierung im Rahmen der Simulation von kompressiblen Strömungsvorgängen, welche durch zufällige Erhaltungsgleichungen, speziell den kompressiblen Euler oder Navier–Stokes Gleichungen, beschrieben werden. Es wird hierbei zwischen polynombasierten (nicht-statistischen) Unsicherheitsquantifizierungmethoden und stichprobenbasierten (statistischen) Unsicherheitsquantifizierungmethoden unterschieden.

Der erste Teil der Arbeit beschäftigt sich mit nicht-statistischen Unsicherheitsquantifizierungsmethoden, nämlich mit der stochastischen Galerkin Methode (SG), der Nichtintrusiven Spektralprojektion (NISP) und der Stochastischen Kollokation (SK). Während die SG Methode häufig zur Unsicherheitsquantifizierung von zufälligen partiellen Differentialgleichungen angewendet wird, stellt der klassiche SG Ansatz nicht sicher dass die Hyperbolizität der zugrunde liegenden hyperbolischen Erhaltungsgleichung erhalten bleibt. Aus diesem Grund wird mit Hilfe eines Limiters ein modifiziertes numerisches Verfahren entwickelt, welches die Hyperbolizität das SG Systems garantiert. Zugleich bleibt eine hohe Approximationsordnung im physikalischen und stochastischen Raum erhalten. Das modifizierte Verfahren wird auf mehrere anspruchsvolle numerische Beispiele angewendet, bei denen das klassische SG Verfahren versagt.

Ein wichtiger Aspekt von Raum-Zeit-stochastischen numerischen Verfahren ist die Quantifizierung des Fehlers der durch die numerische Diskretisierung der zufälligen Erhaltungsgleichung auftritt. In dieser Arbeit leiten wir einen neuen Ansatz zur a posteriori Fehleranalyse von zufälligen hyperbolischen Erhaltungsgleichungen her. Für die numerisch Approximation wird das Runge–Kutta Discontinous Galerkin Verfahren in Kombination mit polynombasierten Unsicherheitsquantifizierungmethoden verwendet. Die Fehlerschätzer basieren auf der relativen Entropiemethode von Dafermos und DiPerna [26] und erlauben es zwischen Raum-Zeit und stochastischem Diskretisierungsfehler zu unterscheiden. Basierend auf dem a posteriori Fehlerschätzer konstruieren wir neue residual-basierte Raum-Zeit-Stochastik adaptive numerische Verfahren. Wir unterstreichen die theoretischen Ergebnisse mit einer Auswahl an verschiedenen numerischen Experimenten.

Der letzte Teil dieser Arbeit widmet sich den statistischen Unsicherheitsquantifizierungmetho-

den, speziell der Monte Carlo (MC) Methode. Wir erweitern die Multilevel Monte Carlo Methode, welche eine Hierarchie von räumlichen Gittern unterschiedlicher Auflösung betrachtet, zu der $hp$-MLMC Methode, welche beliebig $hp$-verfeinerte Gitter zulässt. Die klassiche Komplexitätsanalyse des MLMC Verfahrens wird auf das $hp$-MLMC Verfahren erweitert. Desweiteren wird für eine iterationsbasierte Version des $hp$-MLMC Verfahrens ein Konfidenzintervall hergeleitet, welches eine robuste Abschätzung an die Zahl der optimalen Samples pro Level liefert. Um die Effizienz unseres Ansatzes zu demonstrieren betrachten wir zwei verschiedene kompressible Strömungsprobleme welche mittels den zufälligen Navier–Stokes Gleichungen modelliert werden. Insbesondere betrachten wir sogenannte Fugenströmungen welche eine wichtige Problemklasse im Bereich der computer-gestützten Akustik dartstellen. Da diese Probleme physikalische Effekte mit starker Abhängigkeit von den Eingabeparametern modellieren stellen sie ein anspruchsvolles Problem für Unsicherheitsquantifizierung dar.

# Contents

# 1 Introduction

## Deterministic hyperbolic conservation laws

Many processes in physics and engineering are described by hyperbolic conservation laws, a set of partial differential equations (pdes) which are derived from fundamental principles of physics, namely conservation of mass, momentum and energy. To name a few examples, flows of inviscid ideal gases or fluid flows in rivers and channels can be modeled by hyperbolic conservation laws. The time-evolution of the vector of conserved quantities $u(t,x)$ reads as

$$\begin{cases} \partial_t u(t,x) + \sum_{i=1}^{d} \partial_{x_i} F_i(u(t,x)) = 0, & (t,x) \in (0,T) \times \mathbb{R}^d, \\ u(0,x) = u^0(x), & x \in \mathbb{R}^d, \end{cases} \tag{1.1}$$

where $t$ denotes the variable with respect to (w.r.t.) time and $x_i$ are the spatial variables corresponding to $d$ spatial dimensions. The set $\mathscr{U} \subset \mathbb{R}^m$ is called state space, $F_i \in C^1(\mathscr{U};\mathbb{R}^m)$, is the flux function for the $i$-th direction, and $u^0(x) \in \mathscr{U}$ are some suitable initial conditions. The system (1.1) is called hyperbolic if the flux Jacobian $\mathrm{D}F := \sum_{i=1}^{d} \alpha_i \mathrm{D}F_i$, has $m$ real eigenvalues and a set of $m$ linearly independent eigenvectors for each $\alpha \in \mathbb{R}^d$, $|\alpha| = 1$. Here $\mathrm{D}F_i$ denotes the matrix which contains all partial derivatives of $F_i$ w.r.t. to $u$ in each direction $i = 1,\dots,d$. The system (1.1) is called strictly hyperbolic if the flux Jacobian $\mathrm{D}F$ has $m$ distinct real eigenvalues.

For nonlinear flux functions, solutions of (1.1) may develop discontinuities in finite time and therefore one has to consider weak solutions of (1.1). We call a function $u \in L^\infty((0,T) \times \mathbb{R}^d, \mathbb{R}^m)$, a weak solution of (1.1) if it satisfies

$$\int_0^T \int_{\mathbb{R}^d} \left( u(t,x) \cdot \partial_t \phi(t,x) + \sum_{i=1}^{d} F_i(u(t,x)) \cdot \partial_{x_i} \phi(t,x) \right) \mathrm{d}x \, \mathrm{d}t + \int_{\mathbb{R}^d} u^0(x) \cdot \phi(0,x) \, \mathrm{d}x = 0, \tag{1.2}$$

for all $\phi \in C_c^\infty(\mathbb{R}^d \times (0,T);\mathbb{R}^m)$. Weak solutions of (1.1) are not necessarily unique and thus, to ensure uniqueness, one has to consider additional admissibility criteria. A common criterion is the so-called entropy admissibility criterion [26] which can be posed provided (1.1) is equipped

with an entropy/entropy flux pair, $(\eta, q)$, with $\eta \in C^2(\mathscr{U};\mathbb{R})$ strictly convex and $q_i \in C^2(\mathscr{U};\mathbb{R})$, satisfying

$$\mathrm{D}\, q_i = \mathrm{D}\,\eta\, \mathrm{D}\, F_i \tag{1.3}$$

for all $i = 1,\dots,d$. In addition to (1.2), an entropy admissible weak solution of (1.1) is supposed to satisfy the following weak entropy inequality

$$\int_0^T \int_{\mathbb{R}^d} \Big( \eta(u)\partial_t \Phi(t,x) + \sum_{i=1}^d q_i(u)\partial_{x_i}\Phi(t,x) \Big) \mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}^d} \eta(u)\Phi(0,x)\mathrm{d}x \geq 0, \tag{1.4}$$

for all $\Phi \in C_c^\infty(\mathbb{R}^d \times (0,T);\mathbb{R}_+)$.

Well-posedness of (1.1), i.e. existence, uniqueness and continuous dependence on initial data, of entropy solutions for scalar conservation laws (i.e. $m = 1$) in arbitrary space dimensions has been proven by Kružkov in [72]. The proof relies on the fact that scalar conservation laws admit infinite many entropy/entropy flux pairs. In contrast to scalar conservation laws most systems of hyperbolic conservation laws are endowed with only one entropy/entropy flux pair. In the case of spatially one-dimensional ($d = 1$) systems with initial data with small total variation, existence of weak entropy admissible solutions has been shown by Glimm [53]. Later, Bressan and coauthors [15, 16, 17] proved that the entropy admissible solutions constructed by the Glimm scheme (or equivalently by wave-front tracking) are unique in the sense that they are the only entropy admissible solutions satisfying additional stability properties such as certain bounds on the growth of their total variation. For multi-dimensional systems of linear hyperbolic conservation laws

$$\begin{cases} \partial_t u(t,x) + \sum_{i=1}^d A_i \partial_{x_i} u(t,x) = 0, & (t,x) \in (0,T)\times\mathbb{R}^d, \\ u(0,x) = u^0(x), & x \in \mathbb{R}^d, \end{cases} \tag{1.5}$$

where $A_i \in \mathbb{R}^{m\times m}$ are constant matrices for $i = 1,\dots,d$, existence and uniqueness results of weak solutions for (1.5) are also available, cf. [58, Thm. 5.3.2].

However, when considering nonlinear flux functions, results concerning existence and uniqueness of entropy admissible weak solution for multi-dimensional systems are not available. In contrast, recent results of the authors of [21, 27] have shown that entropy admissible weak solutions in multiple space dimensions for the incompressible and compressible Euler equations are not necessarily unique.

Although for nonlinear multi-dimensional systems the entropy admissibility criterion (1.4) does not guarantee uniqueness of entropy admissible weak solutions, the existence of a convex entropy provides at least local-wellposedness of the Cauchy problem (1.1) in terms of classical

solutions. This means that an entropy admissible weak solution coincides with the classical solution of (1.1) as long as the latter exists, cf. [26, Thm. 5.2.1]. The stability estimate is based on the relative entropy $\eta(\cdot|\cdot) : \mathscr{U} \times \mathscr{U} \to \mathbb{R}$ and the relative entropy flux $q(\cdot|\cdot) : \mathscr{U} \times \mathscr{U} \to \mathbb{R}$, which are defined as

$$\eta(u|v) := \eta(u) - \eta(v) - \mathrm{D}\,\eta(v)(u-v), \tag{1.6}$$

$$q(u|v) := q(u) - q(v) - \mathrm{D}\,\eta(v)(F(u) - F(v)). \tag{1.7}$$

As $\eta$ is strictly convex, $\eta(u|v)$ is also strictly convex and this allows to estimate

$$\eta(u|v) \geq c_v |u-v|^2, \tag{1.8}$$

for some $c_v > 0$, when $u$ is in a compact neighborhood of $v$, cf. [93]. Based on (1.8), it is possible to compare the distance of two solutions $u$, $v$ of (1.1) w.r.t. the $L^2$-norm. If one of the two solutions is at least Lipschitz continuous (and in particular satisfies (1.4) as equality), and the other solution is entropy admissible, one can derive the following stability estimate [26, Thm. 5.2.1]

$$\int_{|x|<r} \eta(u(\cdot,t)|v(\cdot,t))\,\mathrm{d}x \leq a e^{bt} \int_{|x|<r+st} \eta(u^0(\cdot)|v^0(\cdot))\,\mathrm{d}x, \tag{1.9}$$

for some suitable positive constants $a, b, s$ and any $r > 0$. The estimate (1.9) shows on the one hand that entropy admissible weak solutions and classical solutions of (1.1), which have the same initial data, coincide as long as the classical solution exists. This property of entropy admissible weak solutions is also called weak-strong uniqueness. On the other hand, the stability estimate (1.9) can be also be interpreted in an "a posteriori sense". It is possible to bound the approximation error between the entropy admissible weak solution and a suitable numerical approximation, or more precisely, a reconstruction thereof, in the same manner as (1.9) in terms of a computable residual.

While this approach has been successfully applied for deterministic conservation laws cf. [28, 46] our aim is to extend the existing a posteriori error estimate to a probabilistic version of (1.1), where we allow for random variations in initial condition and flux function. To achieve this aim we derive in Section 2.4 a general stability estimate between a so-called random entropy admissible solution and a Lipschitz continuous solution of a perturbed problem. With this error estimate at hand we are able to derive different a posteriori error estimates for various numerical approximations of the random entropy admissible solution. The description of random conservation laws and numerical approximation schemes (also known as UQ methods) for random conservation laws are described in the next section "Uncertainty Quantification".

# Uncertainty Quantification

Although conservation laws accurately describe the evolution of physical quantities like mass density, momentum or temperature, a discrepancy between experimental measurements and high-resolution numerical simulations is often observable. For an example in the context of aeroacoustics we refer to [74, Sec. 4.3.3.], where the authors report a discrepancy between a measured and a simulated frequency spectrum for a flow of air over an open cavity. This misfit occurred because the frequency spectrum exhibits high sensitivity to geometry parameters, which the deterministic numerical simulation did not account for. This example illustrates that apart from deterministic numerical approximation errors, a more fundamental source of error is uncertainty in model parameters or measurements. To name a few more examples, it is often not feasible or possible to accurately describe model parameters [9, 66, 67, 74, 78], or initial and boundary conditions [55, 79, 80]. To reach a satisfactory level of validity and reliability of high fidelity simulations, it is thus inevitable to take systematic uncertainties into account. A possible approach to achieve this aim is to embed (1.1) into a probabilistic framework and model uncertain parameters, initial conditions, or boundary conditions as random fields. The probabilistic versions of (1.1) are called random hyperbolic conservation laws and their solutions are in fact random fields, for which it is possible to compute important statistical information, such as mean, variance or higher order moments. In Chapter 2 we establish the necessary probabilistic framework for the description of random conservation laws and introduce the notion of random entropy admissible weak solutions. Furthermore, we address the issue of well-posedness of random conservation laws with uncertain initial condition and random flux functions.

Apart from discretizing space and time, the numerical approximation of solutions of random conservation laws requires discretizing the random space. Uncertainty Quantification (UQ) provides a large range of different mathematical methods for stochastic discretization. In this thesis, we focus on the forward propagation of uncertainties and consider four different UQ methods, which can be roughly divided into statistical and non-statistical methods. Statistical methods approximate moments of the underlying random field by sampling. The most prominent and most frequently applied method in this class is the Monte Carlo (MC) method. As the asymptotic rate of convergence of the classical MC method is rather slow, Heinrich [61] and later Giles [51] extended the MC method to the Multilevel Monte Carlo (MLMC) method, where they considered spatial mesh hierarchies instead of one fixed mesh to discretize the deterministic equation of interest. The main idea of the MLMC method is that the global behavior of the exact expectation can be approximated by the behavior of the expectation of numerical solutions with a low spatial resolution, which can be computed at low cost. The coarse expectation

is then subsequently corrected by computations on finer levels, which are computationally more expensive per sample. The number of these simulations at full resolution is significantly reduced compared to the original MC method, resulting in a considerably lower overall computational cost. In Chapter 6 we present the classical MLMC method and its extension to arbitrarily $hp$-refined meshes. We extend the classical complexity analysis of MLMC to $hp$-MLMC and we present a novel confidence interval for a robust estimate of the number of additional samples on each level. The number of additional samples on each level plays a crucial role in the efficiency of an iterative version of the $hp$-MLMC method.

For non-statistical methods, we focus on Polynomial Chaos (PC) and Stochastic Collocation (SC) methods. Both approaches are polynomial-based approximations, which derive deterministic models for the stochastic modes. Their description is content of Chapter 4. The theoretical foundation for the PC expansion was laid in [103] and can be described as an approximation of Gaussian random variables by polynomials that are orthogonal with respect to the inner product induced by the probability density function of the uncertain parameters. Later, the approach has been generalized to a larger class of distributions [109], which is now known as generalized Polynomial Chaos (gPC). The modes in the gPC expansion can be either computed by a Galerkin projection, which is also known as Stochastic Galerkin (SG) approach [45], or via a pseudo-spectral projection, which approximates the modes by numerical quadrature [77]. The pseudo-spectral projection is also called Non-intrusive Spectral Projection (NISP) and it is a non-intrusive method. This means that the random conservation law only needs to be evaluated at so-called collocation points, resulting in a finite set of deterministic hyperbolic conservation laws, which can be discretized using already existing numerical solvers. Thus, efficient numerical solver and existing parallelization strategies can readily be used and do not need to be modified. In combination with sparse grids [19] the NISP method proves to be very efficient, especially for high-dimensional random inputs.

In contrast to NISP, the SG approach is an intrusive method because the random conservation law is transformed into a modified deterministic system of conservation laws, which makes the adaption of existing numerical solvers inevitable. For low-dimensional random inputs, the SG method promises higher resolution for fewer degrees of freedom than the NISP method. However, for high-dimensional random inputs this method becomes computationally infeasible due to a very expensive evaluation of the modified flux function. We examine the performance of both methods in detail in Section 4.3. Another flaw of the SG method applied to nonlinear hyperbolic conservation laws is, that the resulting SG system may lose its hyperbolicity, although the original deterministic system is hyperbolic [87]. We address this issue for the SG method in Section 4.2 and present a modified numerical scheme based on the Runge–Kutta Discontinuous

Galerkin method which ensures the hyperbolicity of the underlying SG system.

The SC method approximates the random field by a series of Lagrange polynomials associated with a set of prescribed collocation points. Similarly to the NISP method, SC is a non-intrusive method and hence, deterministic numerical solver can readily be used. Its description is also part of Chapter 4.

## A posteriori error analysis and adaptivity

In recent years stochastic adaptivity has become an important issue for random conservation laws (cf. [18, 57, 69, 99, 101, 104]), mainly because weak solutions of random conservation laws develop discontinuities in finite time and the discontinuities propagate into the random space. In this case it is desirable to locally increase the resolution around discontinuities in spatial as well as in stochastic domain. This is typically achieved using indicators which mark cells for refinement or coarsening. Refinement and coarsening indicators are either derived from rigorous a posteriori error estimates, or they stem from heuristic quantities, for example the gradient of the solution or other more sophisticated quantities. The theory for adaptive mesh refinements (in random and physical space) based on a posteriori error estimates for random elliptic and random parabolic partial differential equations has already reached a very satisfying state [13, 35, 52, 56, 90]. However, compared to elliptic and parabolic equations, the theory for adaptive mesh refinements for random hyperbolic conservation laws is still in its infancy. Although results based on heuristic indicators are available (cf. the list from above), adaptive mesh refinements strategies which rely on rigorous a posteriori error estimates are not available. This is mainly due to the fact that even for deterministic hyperbolic conservation laws a rigorous a posteriori error analysis is still an open problem.

A major contribution of this thesis is the derivation of a rigorous a posteriori error analysis for random hyperbolic systems of conservation laws. Based on the generalized relative entropy method which we introduce in Section 2.4, we derive in Chapter 5 different a posteriori error estimators for the entire space-stochastic discretization error for numerical approximations which rely on the Runge–Kutta Discontinuous Galerkin method in combination with non-statistical UQ methods. More specifically, we discuss the a posteriori error estimates in the context of SG, SC and NISP methods. Furthermore, we prove a splitting of the resulting residual into a spatio-temporal and a stochastic part, which allows us to quantify the errors arising from spatio-temporal and stochastic discretization. For the SC method we use the splitting of the corresponding residual to propose a novel residual-based space-stochastic adaptive numerical

scheme, where the residuals are used as error indicators for local mesh refinement in physical and random space.

# Outline

This thesis is structured as follows. Chapter 2 recapitulates the necessary probabilistic framework, introduces random hyperbolic conservation laws with uncertain initial data and flux function and establishes the notion of random entropy admissible weak solutions. Well-posedness results for random scalar conservation laws are reviewed and the well-posedness for spatially one-dimensional random conservation laws with uncertain initial data and flux function from [47] is presented. The relative entropy method and the stability estimate between weak and strong random entropy admissible solutions are also part of this chapter. Chapter 3 describes the spatio-temporal discretization based on the Runge–Kutta Discontinuous Galerkin (RKDG) method. Moreover, we give a short review of existing deterministic a posteriori error estimates for conservation laws which are approximated with the Finite Volume method or the DG method. In Chapter 4 the Stochastic Galerkin (SG), Non-intrusive Spectral Projection (NISP) and Stochastic Collocation (SC) method are reviewed and the efficiency of NISP and SG up to three random dimensions is assessed by means of a manufactured solution. Chapter 4 is also concerned with the possible loss of hyperbolicity of the SG system and presents a hyperbolicity-preserving numerical scheme for a SG-DG discretization of random hyperbolic conservation laws. The numerical scheme is based on the work [34]. Chapter 5 presents the a posteriori error analysis framework for non-statistical UQ methods from [47, 48, 49]. The last chapter is concerned with the *hp*-MLMC method, an extension of the classical MLMC method to arbitrarily *hp*-refined meshes. It is based on the work [11], and the *hp*-MLMC method is applied to the cavity flow problem from [74]. Finally, Chapter 7 summarizes the results of this thesis and provides a list of directions for further research in the area of UQ of compressible flows.

# 2 Random hyperbolic conservation laws

This section is concerned with the description of random hyperbolic conservation laws. To allow for random variations in the hyperbolic conservation law (1.1) we recapitulate in Section 2.1 the necessary probabilistic framework to model uncertainties as random fields. We then introduce random hyperbolic conservation laws and give a definition of the notion of solutions for random conservation laws, which are called random entropy admissible weak solutions. An important aspect for random conservation laws is the existence and uniqueness of random entropy admissible weak solutions. Well-posedness of random scalar conservation laws is discussed in Section 2.2. In Section 2.3 we establish existence and uniqueness of random entropy admissible weak solutions for one-dimensional systems of hyperbolic conservation laws with initial data with sufficiently small total variation. Section 2.4 is concerned with the description and extension of the relative entropy method to random conservation laws. We also present a stability estimate for the difference between weak and strong random entropy admissible solutions, which will be the foundation of the subsequent a posteriori error analysis framework.

## 2.1 Preliminaries and problem of interest

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, where $\Omega$ is the set of all elementary events $\omega \in \Omega$, $\mathscr{F}$ is a $\sigma$-algebra and $\mathbb{P}$ is a probability measure on $\Omega$. We parametrize the uncertainty with a random vector $\xi : \Omega \to \Xi \subset \mathbb{R}^N$ with $N \in \mathbb{N}$ independent, absolutely continuous components, i.e. $\xi(\omega) = \big(\xi_1(\omega), \dots, \xi_N(\omega)\big) : \Omega \to \Xi \subset \mathbb{R}^N$. This means that for every random variable $\xi_j$ there exists a density function $w_j : \mathbb{R} \to \mathbb{R}_+$, such that $\int_{\mathbb{R}} w_j(y)\,\mathrm{d}y = 1$ and $\mathbb{P}[\xi_j \in A] = \int_A w_j(y)\,\mathrm{d}y$, for any $A \in \mathscr{B}(\mathbb{R})$, for all $j = 1, \dots, N$. Here, $\mathscr{B}(\mathbb{R})$ is the Borel $\sigma$-algebra on $\mathbb{R}$. Moreover, the joint density function $w$ of the random vector $\xi = (\xi_1, \dots, \xi_N)$ can be written as

$$w(y) = \prod_{j=1}^{N} w_j(y_j) \quad \forall\, y = (y_1, \dots, y_N)^\top \in \Xi.$$

The random vector induces a probability measure $\tilde{\mathbb{P}}(B) := \mathbb{P}(\xi^{-1}(B))$ for all $B \in \mathscr{B}(\Xi)$ on the measurable space $(\Xi, \mathscr{B}(\Xi))$. This measure is called the law of $\xi$ and for the remaining part of

this thesis we work on the image probability space $(\Xi, \mathscr{B}(\Xi), \tilde{\mathbb{P}})$.

For a second measurable space $(E, \mathscr{B}(E))$, we consider the weighted $L^r_w$-spaces defined as follows

$$L^r_w(\Xi;E) := \{f : \Xi \to E \mid f \text{ is measurable and } \|f\|_{L^r_w(\Xi;E)} < \infty\},$$

where

$$\|f\|_{L^r_w(\Xi;E)} := \begin{cases} \Big(\int\limits_\Xi \|f(y)\|^r_E \, w(y)\mathrm{d}y\Big)^{1/r} = \mathbb{E}\big(\|f\|^r_E\big)^{1/r}, & 1 \le r < \infty \\ \operatorname{ess\,sup}_{y\in\Xi} \|f(y)\|_E, & r = \infty. \end{cases}$$

Let us now formulate (1.1) in terms of an uncertain initial condition and random flux function. Our problem of interest is the following random initial value problem

$$\begin{cases} \partial_t u(t,x,y) + \sum\limits_{i=1}^d \partial_{x_i} F_i(u(t,x,y),y) = 0, & (t,x,y) \in (0,T)\times\mathbb{R}^d\times\Xi, \\ u(0,x,y) = u^0(x,y), & (x,y) \in \mathbb{R}^d\times\Xi. \end{cases} \tag{2.1}$$

Here, $u(t,x,y) \in \mathscr{U} \subset \mathbb{R}^m$ is the vector of conserved random quantities and $F_i(\cdot,y) \in C^1(\mathscr{U};\mathbb{R}^m)$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ is the random flux function in dimension $i = 1,\dots,d$. This means that

$$\tilde{\mathbb{P}}(\{y \in \Xi \mid F_i(\cdot,y) \in C^1(\mathscr{U};\mathbb{R}^m)\}) = 1.$$

Moreover, $\mathscr{U} \subset \mathbb{R}^m$ is the state space, which is assumed to be an open set and $T \in (0,\infty)$. Following [26, Def. 3.1.1.] we first define the notion of hyperbolicity of (2.1), which essentially coincides with the deterministic definition of hyperbolicity path-wise in $\Xi$.

**Definition 2.1** (Hyperbolicity and characteristic speeds)**.**
*The system* (2.1) *is called hyperbolic if for any $u \in \mathscr{U}$, the $m \times m$ matrix*

$$\sum_{i=1}^d \alpha_i \,\mathrm{D}\, F_i(u,y) \tag{2.2}$$

*has $m$ real eigenvalues, denoted by $\{\lambda_i(\alpha;u,y)\}_{i=1}^m$, and $m$ linearly independent eigenvectors, denoted by $\{r_i(\alpha;u,y)\}_{i=1}^m$, for all $\alpha = (\alpha_1,\dots,\alpha_d) \in S^{d+1} := \{\alpha \in \mathbb{R}^d \mid \sum\limits_{i=1}^d \alpha_i^2 = 1\}$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. It is called strictly hyperbolic if* (2.2) *has $m$ distinct, real eigenvalues. We call the eigenvalues $\{\lambda_i(\alpha;u,y)\}_{i=1}^m$ of* (2.2) *characteristic speeds.*

The following definition of linear degeneracy and genuine nonlinearity from [26, Def. 7.5.1.] will later be required for the analysis of one-dimensional random conservation laws.

**Definition 2.2** (Linear degeneracy and genuine nonlinearity)**.**
*We call a state $u \in \mathscr{U}$ a state of genuine nonlinearity of the i-th characteristic family if*

$$\mathrm{D}\lambda_i(\alpha;u,y)r_i(\alpha;u,y) \neq 0 \tag{2.3}$$

*for all $\alpha \in S^{d+1}$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. We call $u \in \mathscr{U}$ a state of linear degeneracy of the i-th characteristic family if*

$$\mathrm{D}\lambda_i(\alpha;u,y)r_i(\alpha;u,y) = 0, \tag{2.4}$$

*for all $\alpha \in S^{d+1}$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$.*

To establish the notion of entropy weak solutions of (2.1) we first have to introduce a random entropy/entropy flux pair.

**Definition 2.3** (Random entropy/entropy flux pair)**.**
*We say that $\eta \in L^1_w(\Xi;C^2(\mathscr{U};\mathbb{R}))$, $q_i \in L^1_w(\Xi;C^2(\mathscr{U};\mathbb{R}))$, for $i = 1,\dots,d$, form a random entropy/entropy-flux pair if $\eta(\cdot,y)$ is strictly convex $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ and if $\eta$ and $q$ satisfy*

$$\mathrm{D}\eta(\cdot,y)\,\mathrm{D}F_i(\cdot,y) = \mathrm{D}q_i(\cdot,y),$$

*$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$, for all $i = 1,\dots,d$.*

We assume that the random conservation law (2.1) is equipped with at least one random entropy/ entropy-flux pair.

**Remark 2.4.**
*Most hyperbolic systems of conservation laws arising from physics are equipped with an entropy/entropy flux pair, where the entropy function corresponds to the physical entropy resp. energy of the system. In most cases this is the only non-trivial entropy.*

Using Definition 2.3 we can now define random entropy admissible weak solutions of (2.1).

**Definition 2.5** (Random entropy admissible solution)**.**
*The function $u \in L^1_w(\Xi;L^1((0,T)\times\mathbb{R};\mathscr{U}))$ is called a random entropy admissible weak solution of* (2.1) *if it satisfies.*

1. *It is a weak solution:*

$$\int_0^T \int_{\mathbb{R}} \Big( u(t,x,y)\cdot\partial_t\phi(t,x) + \sum_{i=1}^{d} F_i(u(t,x,y),y)\cdot\partial_{x_i}\phi(t,x)\Big)\,\mathrm{d}x\,\mathrm{d}t = -\int_{\mathbb{R}} u^0(x,y)\cdot\phi(0,x)\,\mathrm{d}x, \tag{2.5}$$

*$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ and for all $\phi \in C_c^\infty([0,T)\times\mathbb{R}^d;\mathbb{R}^m)$.*

2. *For any random entropy/entropy flux pair from Definition 2.3 it satisfies the weak entropy inequality:*

$$\int_0^T \int_{\mathbb{R}} \left( \eta(u(t,x,y),y)\partial_t \Phi(t,x) + \sum_{i=1}^{d} q_i(u(t,x,y),y)\partial_{x_i}\Phi(t,x) \right) \mathrm{d}x\,\mathrm{d}t \geq -\int_{\mathbb{R}} \eta(u^0(t,x),y)\Phi(0,x)\,\mathrm{d}x, \tag{2.6}$$

$\tilde{\mathbb{P}}$*-a.s.* $y \in \Xi$ *and for all* $\Phi \in C_c^\infty([0,T)\times\mathbb{R}^d;\mathbb{R}_+)$.

The well-posedness of the integrals in (2.5)-(2.6) is implicitly assumed and follows for scalar conservation laws from the $L^\infty$-bound of the random entropy admissible solution (cf. Theorem 2.7) and for one-dimensional systems from the TV bound of the random entropy admissible solution (Theorem 2.9), resp. the assumption that the random entropy admissible solution only takes values in a compact set.

We are now going to discuss in which cases random entropy admissible weak solutions of (2.1) exist and if they are unique. We first review the theory for scalar conservation laws and then turn to one-dimensional systems of hyperbolic conservation laws.

## 2.2 Well-posedness of scalar random conservation laws

We first consider multi-dimensional scalar random conservation laws, i.e. we restrict ourselves to (2.1) with $m = 1$. Thanks to the fundamental work of Kružkov [72] and the stability estimates [64, Thm. 2.14, Thm. 4.3], there exists a firm theory [78, 79, 89] for the well-posedness of multi-dimensional random scalar conservation laws with uncertain initial condition and uncertain flux function. Before we review the well-posedness result for scalar random conservation laws we present the main stability result of [72] for the deterministic case.

### 2.2.1 Deterministic scalar conservation laws

Let us first consider a deterministic version of (2.1) with $m = 1$ and let us make the following assumptions on initial data and flux function. We keep the same notation for initial condition, flux function and solution as in (2.1).

(D1) The initial condition satisfies $u^0 \in L^1(\mathbb{R}^d;\mathbb{R}) \cap L^\infty(\mathbb{R}^d;\mathbb{R})$.

(D2) The flux function $F_i$ satisfies $F_i \in C^1(\mathbb{R};\mathbb{R})$, for each $i = 1,\ldots,d$.

We then have the following classical result, cf. [70, 72].

**Theorem 2.6.**
*Suppose assumptions* (D1)-(D2) *hold. Then the scalar conservation law* (2.1) *admits a unique random entropy admissible solution. Moreover, for every $t > 0$, let $\mathscr{S}(t)$ be the nonlinear data-to-solution map which is given by $u(t,\cdot) = \mathscr{S}(t)u^0(\cdot)$. It satisfies in particular:*

*(i)* $\mathscr{S}(t) : L^1(\mathbb{R}^d;\mathbb{R}) \to L^1(\mathbb{R}^d;\mathbb{R})$ *is a contraction, i.e.*

$$\|\mathscr{S}(t)u^0 - \mathscr{S}(t)v^0\|_{L^1(\mathbb{R}^d;\mathbb{R})} \leq \|u^0 - v^0\|_{L^1(\mathbb{R}^d;\mathbb{R})},$$

*for almost all $t > 0$.*

*(ii) It is $L^\infty$-bounded, i.e.*

$$\|\mathscr{S}(t)u^0\|_{L^\infty(\mathbb{R}^d;\mathbb{R})} \leq \|u^0\|_{L^\infty(\mathbb{R}^d;\mathbb{R})},$$

*for almost all $t > 0$.*

### 2.2.2 Random scalar conservation laws

We now consider the scalar random conservation law (2.1) with random flux function and random initial data. We have to make the following assumptions on initial data and flux function. Note that they are the probabilistic counterparts to Assumptions (D1)-(D2).

(R1) The initial condition satisfies $u^0 \in L^1_w(\Xi;L^1(\mathbb{R}^d;\mathbb{R}))$ and $\|u^0(\cdot,y)\|_{L^\infty(\mathbb{R}^d)} < \infty$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$.

(R2) The random flux function $F_i$ satisfies $F_i \in L^1_w(\Xi;C^1(\mathbb{R};\mathbb{R}))$ for each $i = 1,\dots,d$ and there exists a constant $0 < c < \infty$, such that $\sum_{i=1,\dots,d} \|F_i(\cdot,y)\|_{C^1(\mathbb{R};\mathbb{R})} \leq c$, holds $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$.

A path-wise application of Kružkov's theorem yields the existence and uniqueness of a path-wise entropy admissible solution. The mapping properties of the data-to-solution operator from Theorem 2.6 are needed to prove the measurability of the path-wise entropy admissible solution.

**Theorem 2.7** ([78, 89])**.**
*Suppose assumptions* (R1)-(R2) *hold. Then the random scalar conservation law* (2.1) *admits a unique random entropy admissible solution, which satisfies*

$$\|u(t,\cdot,y)\|_{L^\infty(\mathbb{R}^d)} \leq \|u^0(\cdot,y)\|_{L^\infty(\mathbb{R}^d)},$$

*for a.e. $t \in (0,T)$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$.*
*Moreover, if $u^0 \in L^k_w(\Xi;L^1(\mathbb{R}^d))$ for any $k \in [1,\infty)$, then*

$$\|u\|_{L^k_w(\Xi;C^0((0,T);L^1(\mathbb{R}^d)))} \leq \|u^0\|_{L^k_w(\Xi;L^1(\mathbb{R}^d))}.$$

From an interpolation inequality in $L^r(\mathbb{R}^d)$-spaces and Theorem 2.7 it follows that $u(\cdot,\cdot,y) \in L^r((0,T)\times\mathbb{R}^d;\mathbb{R})$ for any $1 \le r < \infty$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. Under the assumption that the mapping $y \mapsto u^0(\cdot,y) \in L^r(\mathbb{R}^d;\mathbb{R})$ is measurable one can prove that the random entropy admissible weak solution $u$ is measurable w.r.t. $L^r((0,T)\times\mathbb{R}^d;\mathbb{R})$.

**Lemma 2.8** ([49], Lemma 2.4)**.**
*Let assumptions* (R1)-(R2) *hold and assume that* $u^0 \in L^k_w(\Xi;L^r(\mathbb{R}^d))$ *for some* $k,r \in [1,\infty)$. *Then the mapping*

$$\Big(\Xi,\mathscr{B}(\Xi)\Big) \ni y \mapsto u(\cdot,\cdot,y) \in \Big(L^r((0,T)\times\mathbb{R}^d;\mathbb{R}),\mathscr{B}(L^r((0,T)\times\mathbb{R}^d;\mathbb{R}))\Big)$$

*is measurable.*

## 2.3 Well-posedness of one-dimensional systems of random conservation laws

In contrast to scalar conservation laws, where the entropy admissibility criterion played a central role for uniqueness, for nonlinear multi-dimensional systems of conservation laws the correct notion of solution is still an open problem. Therefore, general existence and uniqueness results are not available. However, in the case of one-dimensional systems, i.e. (2.1) with $d = 1$, equipped with initial data with small total variation, Glimm was able to provide a proof for the global existence of entropy admissible solutions [53]. Later, Bressan and coauthors [15, 16, 17] developed a semi-group theory, which allowed the authors to prove that the entropy admissible solutions constructed by the Glimm scheme (or equivalently by wave-front tracking) are unique in the sense that they are the only entropy admissible solutions satisfying additional stability properties such as certain bounds on the growth of their total variation. Based on these deterministic results, in combination with a stability result of Bianchini and Colombo [14] we prove the existence and uniqueness of random entropy admissible solutions for one-dimensional systems of random hyperbolic conservation laws with uncertain initial data and uncertain flux function. Before we present our main theorem we shortly review some classical deterministic results for hyperbolic conservation laws in one spatial dimension, on which our proof is based.

### 2.3.1 One-dimensional deterministic hyperbolic conservation laws

In this section we first consider a deterministic version of (2.1), while keeping the same notation for initial condition, flux function and solution as in (2.1). The Cauchy problem for the one-

dimensional system of $m$ conservation laws reads as follows.

$$\begin{cases} \partial_t u(t,x) + \partial_x F(u(t,x)) = 0, & (t,x) \in (0,T) \times \mathbb{R}, \\ u(0,x) = u^0(x), & x \in \mathbb{R}. \end{cases} \tag{2.7}$$

We assume that (2.7) is equipped with a strictly convex entropy/entropy flux function. Furthermore, we make the following assumptions on the initial condition and flux function.

(D1) The initial condition satisfies $u^0 \in L^1(\mathbb{R};\mathscr{U})$.

(D2) The flux function satisfies $F \in C^1(\mathscr{U};\mathbb{R}^m)$ and the Jacobian $\mathrm{D}F$ has $m$ distinct real eigenvalues, with each characteristic field being either genuinely nonlinear or linearly degenerate, cf. Definition 2.1 and Definition 2.2.

**Theorem 2.9** ([16], Theorem 2)**.**
*Provided* (D1) *and* (D2) *hold, there exists a non-empty closed domain $\mathscr{D} \subset L^1(\mathbb{R};\mathscr{U})$ of integrable functions with small total variation and a semi-group $\mathscr{S}(t) : [0,\infty) \times \mathscr{D} \to \mathscr{D}$, called Standard Riemann Semigroup (SRS), that is unique (up to its domain) and which has in particular the following properties:*

*(i) There exists a constant $L > 0$, such that*

$$\|\mathscr{S}(s)\overline{u} - \mathscr{S}(t)\overline{v}\|_{L^1(\mathbb{R};\mathscr{U})} \leq L\Big(|s-t| + \|\overline{u}-\overline{v}\|_{L^1(\mathbb{R};\mathscr{U})}\Big),$$

*for all $\overline{u},\overline{v} \in \mathscr{D}$ and for all $s,t \geq 0$.*

*(ii) For $\overline{u} \in \mathscr{D}$ the function $u(t,x) := (\mathscr{S}(t)\overline{u})(x)$ is an entropy admissible solution of* (2.7)*. It is the unique entropy admissible solution that is obtained as $L^1$-limit of the wave-front tracking algorithm.*

**Remark 2.10** (Uniqueness)**.**
*While* (2.7) *may have several entropy admissible solutions there is one and only one entropy admissible solution induced by the SRS; in this sense entropy admissible solutions induced by SRS are unique. It was proven in [16] that the SRS-induced entropy admissible solution is the only entropy admissible solution satisfying certain additional stability properties, cf. [16, (A2),(A3)].*

Additionally, we will use the following result on the stability of the SRS. In particular, we can quantify how much the SRS-induced entropy admissible solution varies if the flux is changed. This result will be necessary to prove measurability of the random entropy admissible solution when considering uncertain flux functions.

**Theorem 2.11** ([14], Corollary 2.5)**.**
*Let the flux function $F$ satisfy* (D2) *and assume*

$$\mathscr{D}(F) \subseteq \{u \in L^1(\mathbb{R};\mathscr{C}) \mid TV(u) \leq M\}, \tag{2.8}$$

*for some suitable positive $M \in \mathbb{R}$ and some compact set $\mathscr{C} \subset \mathbb{R}^m$. For $t > 0$ we denote by $\mathscr{S}(t,F)$ the SRS from Theorem 2.9, associated with the flux function $F$.*

*Then there exists a constant $C > 0$, such that for any flux function $\tilde{F}$, satisfying* (D2) *and $\mathscr{D}(\tilde{F}) \subseteq \mathscr{D}(F)$, it holds that*

$$\|\mathscr{S}(t,F)u - \mathscr{S}(t,\tilde{F})u\|_{L^1(\mathbb{R};\mathscr{U})} \leq Ct\|\mathrm{D}F - \mathrm{D}\tilde{F}\|_{C^0(\mathscr{U})}, \tag{2.9}$$

*for all $u \in \mathscr{D}(\tilde{F})$.*

**Remark 2.12** (Domain of SRS)**.**
*The domain of the SRS is discussed in [16, equation (1.3)]. Note that it can always be replaced by a smaller set in order to make sure that additional conditions (such as* (2.8)*) hold.*

### 2.3.2 One-dimensional random hyperbolic conservation laws

Based on the deterministic results from the previous section, we are now able to prove existence and uniqueness of random entropy admissible weak solutions of (2.1). We assume that (2.1) admits an random entropy/entropy flux pair as defined in Definition 2.3. and make the following assumptions on the random initial condition and on the random flux function. Note that the first two assumptions are the probabilistic versions of assumptions (D1) and (D2).

(R1) The uncertain initial condition satisfies $u^0 \in L^1_w(\Xi;L^1(\mathbb{R};\mathscr{U}))$.

(R2) For almost every realization $y \in \Xi$ we have $F(\cdot,y) \in C^1(\mathscr{U};\mathbb{R}^m)$ and the Jacobian $\mathrm{D}F(\cdot,y)$ has $m$ distinct real eigenvalues, and each characteristic field is either linearly degenerate or genuinely nonlinear. Moreover, we assume that $F \in L^1_w(\Xi;C^1(\mathscr{U};\mathbb{R}^m))$.

(R3) We define $\mathscr{D} := \bigcap_{y\in\Xi} \mathscr{D}(F(\cdot,y))$, where $\mathscr{D}(F(\cdot,y))$ is the domain of the SRS from (2.8) in Theorem 2.11. We assume that $\mathscr{D} \neq \emptyset$ and $u^0(\cdot,y) \in \mathscr{D}$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$.

(R4) There exists a compact and convex set $\mathscr{C} \subset \mathscr{U}$, s.t. $\mathscr{S}(t,F(\cdot,y))u^0(\cdot,y)(x) \in \mathscr{C}$, a.e. $(t,x,y) \in (0,T)\times\mathbb{R}\times\Xi$ and $u^0(x,y) \in \mathscr{C}$, a.e. $(x,y) \in \mathbb{R}\times\Xi$.

**Theorem 2.13.**
*Let the assumptions* (R1)-(R4) *hold. For $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ we define $u(t,x,y) := \mathscr{S}(t,F(\cdot,y))u^0(\cdot,y)(x)$, where $\{\mathscr{S}(t,F(\cdot,y))\}_{t\geq 0}$ is the SRS from Theorem 2.9 associated with the flux-function $F(\cdot,y)$.*

*Then $u$ is a random entropy admissible solution of* (2.1). *It is unique in the sense that it is the only random entropy admissible solution which path-wise coincides with the SRS-induced entropy admissible solution of the deterministic version of* (2.1).

*Proof.* The function $u$ is path-wise the unique SRS-induced entropy solution of (2.1) by construction. Note that we have assumed $u^0(\cdot,y) \in \mathscr{D} \subset \mathscr{D}(F(\cdot,y))$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. It remains to show, that $u$ is a random variable, i.e.

$$\Big(\Xi, \mathscr{B}(\Xi)\Big) \ni y \mapsto u(\cdot,\cdot,y) \in \Big(L^1((0,T)\times\mathbb{R};\mathscr{U}), \mathscr{B}((L^1((0,T)\times\mathbb{R};\mathscr{U})))\Big)$$

is a measurable map. To this end we define the vector space

$$E_1 := L^1(\mathbb{R};\mathscr{U}) \times C^1(\mathscr{U};\mathbb{R}^m)$$

equipped with the norm

$$\|(\overline{u},\overline{F})\|_{E_1} := \|\overline{u}\|_{L^1(\mathbb{R};\mathscr{U})} + \|\overline{F}\|_{C^1(\mathscr{U};\mathbb{R}^m)}.$$

Using Theorem 2.9 (i) and Theorem 2.11, which we can apply due to assumptions (R3) and (R4), we deduce

$$\begin{aligned}
&\|\mathscr{S}(t,F(\cdot,y))u^0(\cdot,y) - \mathscr{S}(t,F(\cdot,\tilde{y}))u^0(\cdot,\tilde{y})\|_{L^1(\mathbb{R};\mathscr{U})} \\
&\quad\le \|\mathscr{S}(t,F(\cdot,y))u^0(\cdot,y) - \mathscr{S}(t,F(\cdot,\tilde{y}))u^0(\cdot,y)\|_{L^1(\mathbb{R};\mathscr{U})} \\
&\qquad + \|\mathscr{S}(t,F(\cdot,\tilde{y}))u^0(\cdot,y) - \mathscr{S}(t,F(\cdot,\tilde{y}))u^0(\cdot,\tilde{y})\|_{L^1(\mathbb{R};\mathscr{U})} \\
&\quad\le Ct\|\mathrm{D}F(\cdot,y) - \mathrm{D}F(\cdot,\tilde{y})\|_{C^0(\mathscr{U})} + L\|u^0(\cdot,y) - u^0(\cdot,\tilde{y})\|_{L^1(\mathbb{R};\mathscr{U})} \\
&\quad\le Ct\|F(\cdot,y) - F(\cdot,\tilde{y})\|_{C^1(\mathscr{U};\mathbb{R}^m)} + L\|u^0(\cdot,y) - u^0(\cdot,\tilde{y})\|_{L^1(\mathbb{R};\mathscr{U})},
\end{aligned}$$

for $\tilde{\mathbb{P}}$-a.s. $y,\tilde{y} \in \Xi$.

Hence, the mapping $\mathrm{S}(t) : (\overline{u},\overline{F}) \ni E_1 \to L^1(\mathbb{R};\mathscr{U})$, $\mathrm{S}(t)(\overline{u},\overline{F}) := \mathscr{S}(t,\overline{F})\overline{u}(\cdot)$ is continuous for almost all $t > 0$. Due to the finite time horizon we immediately deduce that the mapping

$$\mathrm{S} : E_1 \to L^1((0,T)\times\mathbb{R};\mathscr{U}),\ \mathrm{S}(\overline{u},\overline{F}) := \mathscr{S}(\cdot,\overline{F})\overline{u}(\cdot)$$

is also continuous. Finally, it follows from assumptions (R1) and (R2) that the mapping

$$\mathrm{S}_0 : \Big(\Xi, \mathscr{B}(\Xi)\Big) \to \Big(E_1, \mathscr{B}(E_1)\Big),\ \mathrm{S}_0(y) := (u^0(\cdot,y), F(\cdot,y))$$

is measurable. Thus, $u(\cdot,\cdot,y) = \mathscr{S}(\cdot,F(\cdot,y))u^0(\cdot,y) = \mathrm{S}\circ\mathrm{S}_0(y)$ is a composition of measurable mappings and hence measurable itself. □

In light of the previous existence and uniqueness results for random conservation laws the remaining part of this thesis is dedicated to the derivation of different suitable numerical schemes to approximate the random entropy admissible weak solution numerically. In addition to approximating the entropy admissible weak solution another major goal of this thesis is to derive computable error bounds for the numerical approximation error. We derive error bounds based on a probabilistic version of the relative entropy stability framework from [26]. To this end we extend the relative entropy framework to random conservation laws and generalize the stability result from [26, Thm. 5.2.1 ] to random entropy admissible weak solutions. This is the content of the next section.

## 2.4 Relative entropy method and stability estimate

The aim of this chapter is to derive a general stability framework to estimate the distance between the random entropy admissible weak solution and a strong solution of a perturbed version of (2.1), in terms of the relative entropy. Our main result (Theorem 2.15) is a stability estimate for the difference between the random entropy admissible weak solution and a Lipschitz continuous (in space and time) function, which satisfies a perturbed version of (2.1). For the stability estimate we restrict ourselves to a one-dimensional spatial domain and consider periodic boundary conditions. Without loss of generality we set $D = [0,1]_{per}$. Moreover, we assume that $F(\cdot,y) \in C^2(\mathscr{U};\mathbb{R}^m)$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ and that the problem (2.1) admits a random entropy/entropy flux pair and a unique random entropy admissible weak solution which we denote by $u$. Additionally, we consider a function $\bar{v}$ which has to satisfy the following assumptions.

(A1) $\bar{v}(\cdot,\cdot,y) \in W^1_\infty((0,T)\times D;\mathscr{U})$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$

(A2) $\bar{v}(t,x,y) \in \mathscr{C}$ a.e. $(t,x,y) \in (0,T)\times D\times\Xi$, where $\mathscr{C} \subset \mathscr{U}$ is a compact and convex set

We denote the initial data of $\bar{v}$ by $\bar{v}^0(\cdot,y) := \bar{v}(0,\cdot,y)$, $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. Plugging $\bar{v}$ into (2.1), we may view $\bar{v}$ as solution of a perturbed version of (2.1), satisfying

$$\mathscr{R}(\cdot,\cdot,y) := \partial_t \bar{v}(\cdot,\cdot,y) + \partial_x F(\bar{v}(\cdot,\cdot,y),y), \tag{2.10}$$

$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. We call the function $\mathscr{R}(\cdot,\cdot,y)$ defined by (2.10) residual. With the help of the relative entropy stability framework we can bound the difference between $u$ and $\bar{v}$ in terms of the residual $\mathscr{R}$. Let us first extend the relative entropy/entropy flux pair to what we call random relative entropy/entropy flux pairs.

**Definition 2.14** (Random relative entropy/entropy flux pair)**.**
*Let* $(\eta,q)$ *be a random entropy/entropy flux pair of* (2.1) *as defined in Definition 2.3. For*

*$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ and for any $v, w \in \mathscr{U}$, we call the functions*

$$\eta(v|w,y) := \eta(v,y) - \eta(w,y) - \mathrm{D}\,\eta(w,y)(v-w), \tag{2.11}$$

$$q(v|w,y) := q(v,y) - q(w,y) - \mathrm{D}\,\eta(w,y)(F(v,y) - F(w,y)). \tag{2.12}$$

*random relative entropy and random relative entropy flux. Here, $\mathrm{D}\,\eta(w,y) := \mathrm{D}_w\,\eta(w,y)$ is the matrix which consists of the partial derivatives of $\eta$ w.r.t w but not w.r.t. y.*

Before stating the main estimate, we establish bounds on the derivatives of the flux function and the entropy, as they enter the upper bounds in the main estimate. Due to Assumption (A2) and the compactness of $\mathscr{C}$, there exist $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ constants $0 < C_{\overline{F}}(y) < \infty$ and $0 < C_{\underline{\eta}}(y) < C_{\overline{\eta}}(y) < \infty$, such that,

$$|w^\top \mathrm{H}\,F(v,y)w| \le C_{\overline{F}}(y)|w|^2, \quad C_{\underline{\eta}}(y)|w|^2 \le w^\top \mathrm{H}\,\eta(v,y)w \le C_{\overline{\eta}}(y)|w|^2, \quad \forall w \in \mathbb{R}^m, \forall v \in \mathscr{C}. \tag{2.13}$$

Here, for a generic function $f$, $\mathrm{H}\,f := \mathrm{H}_v f$ denotes its Hessian matrix which contains all second order derivatives with respect to $v$. We are now able to state the main stability estimate of this section.

**Theorem 2.15** (Stability estimate)**.**
*Let $u$ be the random entropy admissible weak solution of* (2.1) *and let $\bar{v}$ be a Lipschitz solution of* (2.10) *satisfying Assumptions* (A1)-(A2)*. Then the difference between $u$ and $\bar{v}$ satisfies*

$$\begin{aligned}
&\|u(t,\cdot,\cdot) - \bar{v}(t,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \\
&\qquad \le \int_{\tilde{\Xi}} \left[ C_{\underline{\eta}}(y)^{-1} \left( \|\mathscr{R}(\cdot,\cdot,y)\|^2_{L^2((0,T)\times D)} + C_{\overline{\eta}}(y)\|u^0(\cdot,y) - \bar{v}^0(\cdot,y)\|^2_{L^2(D)} \right) \right. \\
&\qquad\qquad \left. \times \exp\left( \int_0^t \frac{C_{\overline{\eta}}(y)C_{\overline{F}}(y)\|\partial_x \bar{v}(s,\cdot,y)\|_{L^\infty(D)} + C_{\overline{\eta}}(y)^2}{C_{\underline{\eta}}(y)} \mathrm{d}s \right) \right] w(y)\mathrm{d}y,
\end{aligned} \tag{2.14}$$

*for a.e. $t \in (0,T)$ and any $\tilde{\mathbb{P}}$-measurable set $\tilde{\Xi} \subset \Xi$.*

*Proof.* Recalling the argument in [26, Sec. 4.5], which says that the distribution

$$\partial_t \eta(u,y) + \partial_x q(u,y) \le 0$$

has a sign and is therefore a measure $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$, we may replace the smooth test function in (2.6) by Lipschitz continuous ones. Thus, because $u$ satisfies the entropy inequality (2.6)

$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$, we have for almost any realization $y \in \Xi$ and for any nonnegative Lipschitz continuous test function $\phi$ the inequality

$$0 \leq \int_0^T \int_D \eta(u,y)\partial_t\phi(\cdot,\cdot) + q(u,y)\partial_x\phi(\cdot,\cdot)\,\mathrm{d}x\mathrm{d}t + \int_D \eta(u^0,y)\phi(0,\cdot)\mathrm{d}x. \tag{2.15}$$

Next, we multiply (2.10) by $\mathrm{D}\eta(\bar{v},y)$. Upon using the chain rule for Lipschitz continuous functions and the relationship $\mathrm{D}q(\cdot,y) = \mathrm{D}\eta(\cdot,y)\mathrm{D}F(\cdot,y)$ we derive the following relation

$$\mathrm{D}\eta(\bar{v}(\cdot,\cdot,y),y)\mathscr{R}(\cdot,\cdot,y) = \partial_t\eta(\bar{v}(\cdot,\cdot,y),y) + \partial_x q(\bar{v}(\cdot,\cdot,y),y), \tag{2.16}$$

$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. Let us consider the weak form of (2.16) and subtract it from (2.15) to obtain

$$\begin{aligned} 0 \leq & \int_0^T \int_D (\eta(u,y) - \eta(\bar{v},y))\partial_t\phi + (q(u,y) - q(\bar{v},y))\partial_x\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D \mathscr{R}(\cdot,\cdot,y)\mathrm{D}\eta(\bar{v},y)\phi \,\mathrm{d}x\mathrm{d}t + \int_D (\eta(u^0,y) - \eta(\bar{v}^0,y))\phi(0,x)\,\mathrm{d}x, \end{aligned}$$

$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. Recalling Definition 2.14 we write

$$\begin{aligned} 0 \leq & \int_0^T \int_D (\eta(u|\bar{v},y) + \mathrm{D}\eta(\bar{v},y)(u-\bar{v})\partial_t\phi \,\mathrm{d}x\mathrm{d}t \\ & + \int_0^T \int_D (q(u|\bar{v},y) + \mathrm{D}\eta(\bar{v},y)(F(u,y) - F(\bar{v},y)))\partial_x\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D \mathscr{R}(\cdot,\cdot,y)\mathrm{D}\eta(\bar{v},y)\phi \,\mathrm{d}x\mathrm{d}t + \int_D (\eta(u^0,y) - \eta(\bar{v}^0,y))\phi(0,x)\,\mathrm{d}x. \end{aligned} \tag{2.17}$$

Using the Lipschitz continuous (in space and time) test function $\mathrm{D}\eta(\bar{v},y)\phi$ in (2.5) and in the weak formulation of (2.10) we obtain

$$\begin{aligned} 0 = & \int_0^T \int_D (u-\bar{v})\partial_t(\mathrm{D}\eta(\bar{v},y)\phi) + (F(u,y) - F(\bar{v},y))\partial_x(\mathrm{D}\eta(\bar{v},y)\phi) \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D \mathscr{R}(\cdot,\cdot,y)\mathrm{D}\eta(\bar{v},y)\phi \,\mathrm{d}x\mathrm{d}t + \int_D (u^0 - \bar{v}^0)(\mathrm{D}\eta(\bar{v}^0,y)\phi(0,x))\,\mathrm{d}x. \end{aligned} \tag{2.18}$$

Using the product rule yields $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$

$$\begin{aligned} \partial_t(\mathrm{D}\eta(\bar{v},y)\phi) &= \partial_t\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)\phi + \partial_t\phi\,\mathrm{D}\eta(\bar{v},y), \\ \partial_x(\mathrm{D}\eta(\bar{v},y)\phi) &= \partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)\phi + \partial_x\phi\,\mathrm{D}\eta(\bar{v},y). \end{aligned}$$

Combining (2.18) with (2.17), we obtain

$$0 \leq \int_0^T \int_D \eta(u|\bar{v},y)\partial_t\phi + q(u|\bar{v},y)\partial_x\phi - \partial_t\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)(u-\bar{v})\phi \,\mathrm{d}x\mathrm{d}t$$
$$- \int_0^T \int_D (\partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)(F(u,y)-F(\bar{v},y))\phi \,\mathrm{d}x\mathrm{d}t + \int_D \eta(u^0|\bar{v}^0,y)\phi(0,x)\,\mathrm{d}x.$$

Rearranging (2.10) yields

$$\partial_t\bar{v}(\cdot,\cdot,y) = -\mathrm{D}F(\bar{v}(\cdot,\cdot,y),y)\partial_x\bar{v}(\cdot,\cdot,y) + \mathscr{R}(\cdot,\cdot,y),$$

$\tilde{\mathbb{P}}$-a.s. $y \in \Xi$. We conclude that

$$\begin{aligned} 0 \leq & \int_0^T \int_D \eta(u|\bar{v},y)\partial_t\phi + q(u|\bar{v},y)\partial_x\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D [-\mathrm{D}F(\bar{v},y)\partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y) + \mathscr{R}(\cdot,\cdot,y)](u-\bar{v})\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D (\partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)(F(u,y)-F(\bar{v},y))\phi \,\mathrm{d}x\mathrm{d}t \\ & + \int_D \eta(u^0|\bar{v}^0,y)\phi(0,x)\,\mathrm{d}x. \end{aligned}$$

Using the fact that

$$\mathrm{D}F(\bar{v},y)^\top \mathrm{H}\,(\bar{v},y) = \mathrm{H}\,(\bar{v},y)\mathrm{D}F(\bar{v},y)$$

holds $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ (cf. [26, (3.2.4.)]), the last inequality is reformulated as

$$\begin{aligned} 0 \leq & \int_0^T \int_D \eta(u|\bar{v},y)\partial_t\phi + q(u|\bar{v},y)\partial_x\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D \partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)[F(u,y)-F(\bar{v},y)-\mathrm{D}F(\bar{v},y)(u-\bar{v})]\phi \,\mathrm{d}x\mathrm{d}t \\ & - \int_0^T \int_D \mathscr{R}(\cdot,\cdot,y)(u-\bar{v})\phi \,\mathrm{d}x\mathrm{d}t + \int_D \eta(u^0|\bar{v}^0,y)\phi(0,x)\,\mathrm{d}x. \end{aligned} \tag{2.19}$$

Up to now the choice of $\phi$ was arbitrary. We fix $s > 0$ and $\varepsilon > 0$ and define $\phi$ as follows

$$\phi(\sigma,x) := \begin{cases} 1 & : \sigma < s, \\ 1 - \frac{\sigma - s}{\varepsilon} & : s < \sigma < s+\varepsilon, \\ 0 & : \sigma > s+\varepsilon. \end{cases}$$

With this particular choice we obtain

$$0 \leq -\frac{1}{\varepsilon}\int_s^{s+\varepsilon}\int_D \eta(u|\bar{v},y)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_D \partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)[F(u,y)-F(\bar{v},y)-\mathrm{D}F(\bar{v},y)(u-\bar{v})]\phi\,\mathrm{d}x\,\mathrm{d}t$$
$$-\int_0^T\int_D \mathscr{R}(\cdot,\cdot,y)(u-\bar{v})\phi\,\mathrm{d}x\,\mathrm{d}t + \int_D \eta(u^0|\bar{v}^0,y)\phi(0,x)\,\mathrm{d}x.$$

Sending $\varepsilon \to 0$ we find for all Lebesgue-points $s$ of $\eta(u(\sigma,\cdot,y),y)$ in $(0,T)$ that

$$\begin{aligned}0 \leq &-\int_D \eta(u(s,\cdot,y),y)|\bar{v}(s,\cdot,y),y)\,\mathrm{d}x\\ &-\int_0^s\int_D \partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)[F(u,y)-F(\bar{v},y)-\mathrm{D}F(\bar{v},y)(u-\bar{v})]\,\mathrm{d}x\,\mathrm{d}t\\ &-\int_0^s\int_D \mathrm{H}\,\eta(\bar{v},y)\mathscr{R}(\cdot,\cdot,y)(u-\bar{v})\,\mathrm{d}x\,\mathrm{d}t + \int_D \eta(u^0|\bar{v}^0,y)\,\mathrm{d}x.\end{aligned}$$

We then estimate

$$\int_0^s\int_D \mathrm{H}\,\eta(\bar{v},y)\mathscr{R}(\cdot,\cdot,y)(u-\bar{v})\,\mathrm{d}x\,\mathrm{d}t$$

by Young's inequality which in combination with (2.13) yields the constant $C_{\overline{\eta}}(y)^2$. The integral

$$\int_0^T\int_D \partial_x\bar{v}\,\mathrm{H}\,\eta(\bar{v},y)[F(u,y)-F(\bar{v},y)-\mathrm{D}F(\bar{v},y)(u-\bar{v})]\,\mathrm{d}x\,\mathrm{d}t$$

is estimated by Taylor's theorem and (2.13), which yields the constants $C_{\overline{F}}(y)$ and $C_{\overline{\eta}}(y)$. The remaining two terms involving the relative entropy are also estimated using Taylor's theorem and (2.13). Altogether we obtain $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$

$$\begin{aligned}C_{\underline{\eta}}(y)\int_D |u(s,\cdot,y)-\bar{v}(s,\cdot,y)|^2\,\mathrm{d}x&\\ \leq C_{\overline{F}}(y)C_{\overline{\eta}}(y)\int_0^s \Big(\|\partial_x\bar{v}(t,\cdot,y)\|_{L^\infty(D)}\|u(\cdot,\cdot,y)-\bar{v}(\cdot,\cdot,y)\|^2_{L^2(D)}\,\mathrm{d}x\Big)\,\mathrm{d}t&\\ +\|\mathscr{R}(\cdot,\cdot,y)\|^2_{L^2((0,s)\times D)} + C_{\overline{\eta}}(y)^2\int_0^s \|u(\cdot,\cdot,y)-\bar{v}(\cdot,\cdot,y)\|^2_{L^2(D)}\,\mathrm{d}t&\\ +C_{\overline{\eta}}(y)\|u^0(\cdot,y)-v^0(\cdot,y)\|^2_{L^2(D)}.&\end{aligned}$$

using Gronwall's inequality and integrating over any $\tilde{\mathbb{P}}$-measurable $\tilde{\Xi} \subset \Xi$ yields the assertion.

□

**Remark 2.16.**

*For a deterministic flux function $F = F(\cdot)$ the estimate* (2.14) *can be substantially simplified to*

$$\begin{aligned}
&\|u(t,\cdot,\cdot) - \bar{v}(t,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \\
&\qquad \leq C_{\underline{\eta}}^{-1}\left(\|\mathscr{R}(\cdot,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2((0,T)\times D))} + C_{\overline{\eta}}\|u^0(\cdot,\cdot) - \bar{v}^0(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))}\right) \\
&\qquad\quad \times \exp\left(\int_0^t \frac{C_{\overline{\eta}} C_{\overline{F}}\|\partial_x \bar{v}(s,\cdot,\cdot)\|_{L^\infty_w(\tilde{\Xi};L^\infty(D))} + C^2_{\overline{\eta}}}{C_{\underline{\eta}}}\right)\Bigg].
\end{aligned}$$

Based on Theorem 2.13 we are now able to derive different a posteriori error estimates for various numerical schemes which approximate the random entropy admissible solution numerically. The main idea is to modify the numerical solution to a function which is (at least) Lipschitz continuous in space and time. As the reconstruction process strongly depends on the UQ method, we derive the a posteriori error estimates after presenting the different UQ methods.

The next chapter is concerned with the space and time discretization of (2.1). In this thesis we rely on the Runge–Kutta Discontinuous Galerkin method. Its description and a review of existing deterministic a posteriori error estimates for Finite Volume and Discontinuous Galerkin schemes is the content of the next chapter.

# 3 Space and time discretization of conservation laws

All UQ methods that are used in this work rely on a spatio-temporal numerical scheme which solves deterministic equations resulting from stochastic discretization. For spatio-temporal discretizations of hyperbolic conservation laws there exist different numerical methods, for example Finite Differences (FD), Finite Volumes (FV) or Finite Elements (FE). In this thesis we rely on the Runge–Kutta Discontinuous Galerkin (RKDG) from [23], which combines the (local) high-order accuracy of FE with the discrete conservation property of FV. We give a detailed description of the RKDG method in the next section and for the sake of completeness we give a short description of the FV method, which can be regarded as a special case of the DG method. Beside the description of both methods we provide a small review of some classical a posteriori error estimates for the error between entropy admissible solutions and their numerical approximations obtained with FV or DG.

## 3.1 The Runge–Kutta Discontinuous Galerkin method

The RKDG method for systems of conservation laws has originally been introduced by Cockburn and Shu [23], see also [33, 62] for a detailed overview of the RKDG method. We explain the RKDG method by means of the deterministic conservation law (1.1), following closely the illustration of [33]. Let us consider a bounded $d$-dimensional spatial domain $D \subset \mathbb{R}^d$, which we assume to be a polyhedron. We partition $D$ into $N_s \in \mathbb{N}$ quadrilateral elements $Q_l$, $l = 1, \dots, N_s$ with $D = \bigcup_{l=1}^{N_s} Q_l$. Moreover, we assume that the mesh is conforming, i.e. it has no hanging nodes. We define the mesh parameters $h_{\max} := \max_{l=1,\dots,N_s} \bar{h}_l$, $h_{\min} := \max_{l=1,\dots,N_s} \underline{h}_l$, where $\bar{h}_l, \underline{h}_l$ is the longest, resp. shortest edge of $Q_l$. With $\mathscr{F}_h^I$ we denote the set of all $(d-1)$-dimensional (inner) faces (edges for $d = 2$), that are contained in $D$. For each $\Gamma \in \mathscr{F}_h^I$ the exist two elements $Q_\Gamma^-, Q_\Gamma^+$, such that $\Gamma \in Q_\Gamma^- \cap Q_\Gamma^+$. We assume that $Q_\Gamma^+$ lies in the direction of the normal vector

of $\Gamma$, denoted by $n_\Gamma$ (cf. Figure 3.1). With $\mathscr{F}_h^B$ we denote the set of all $(d-1)$-dimensional faces which lie on the boundary of $D$, i.e. for all $\Gamma \in \mathscr{F}_h^B$ it follows that $\Gamma \subset \partial D$. With this

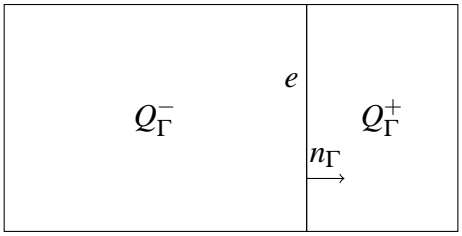



Figure 3.1: Edge $\Gamma \in \mathscr{F}_h^I$, adjacent elements and normal vector

convention we define the spatial traces on $\Gamma \in \mathscr{F}_h^I$ as

$$u_\Gamma^-(\xi) := \lim_{x \to \xi:\ (x-\xi)\cdot n_\Gamma > 0} u(x), \qquad u_\Gamma^+(\xi) := \lim_{x \to \xi:\ (x-\xi)\cdot n_\Gamma < 0} u(x), \tag{3.1}$$

for any $\xi \in \Gamma$. Using the same notation we define the jump and average function as follows:

$$[\![ u ]\!]_\Gamma(\xi) := (u_\Gamma^-(\xi) - u_\Gamma^+(\xi)), \qquad \{u\}_\Gamma(\xi) := \frac{1}{2}(u_\Gamma^-(\xi) + u_\Gamma^+(\xi)), \tag{3.2}$$

where for the remaining part of this thesis we suppress the argument $\xi \in \Gamma$.

Let us now introduce the space of piecewise DG polynomial ansatz and test functions:

$$\mathscr{V}_h^q := \{u : D \to \mathbb{R}^m \mid u\big|_{Q_l} \in \mathbb{P}_q(Q_l;\mathbb{R}^m),\ \text{for } 1 \le l \le N_s\},$$

where $\mathbb{P}_q(Q_l;\mathbb{R}^m)$ is the space of polynomials of degree $q \in \mathbb{N}_0$ on the element $Q_l$ and $u\big|_{Q_l}$ denotes $u$ restricted to $Q_l$. Two common choices for basis polynomials of $\mathscr{V}_h^q$ are either orthogonal polynomials which yield a so-called modal basis, or Lagrange polynomials associated with a set of interpolation nodes, yielding a nodal basis, cf. [62, p.20-21]. In this thesis we use the nodal basis, i.e. on each element $Q_l$, $l = 1,\dots,N_s$ we use tensor products of local one-dimensional Lagrange interpolation polynomials of degree $q \in \mathbb{N}_0$, where the interpolation nodes are chosen to be Gauß–Legendre or Gauß–Lobatto nodes, cf. [68].

The DG solution $u_h$ of (1.1) is sought in $\mathscr{V}_h^q$, i.e. $u_h(t,\cdot) \in \mathscr{V}_h^q$, a.e. $t \in (0,T)$. To derive the DG spatial discretization we consider the weak form of (1.1) using the same ansatz and test functions.

$$\sum_{l=1}^{N_s} \int_{Q_l} \left(\partial_t u_h \cdot v_h - \sum_{i=1}^d F_i(u_h)\cdot \partial_{x_i} v_h\right) \mathrm{d}x + \sum_{\Gamma \in \mathscr{F}_h^I} \int_\Gamma \sum_{i=1}^d F_i(u_h) n_{\Gamma,i} \cdot [\![ v_h ]\!]_\Gamma \ \mathrm{d}s$$
$$+ \sum_{\Gamma \in \mathscr{F}_h^B} \int_\Gamma \sum_{i=1}^d F_i(u_h) n_{\Gamma,i} \cdot v_h \ \mathrm{d}s = 0, \tag{3.3}$$

for all $v_h \in \mathscr{V}_h^q$. Here, $n_\Gamma = (n_{\Gamma,1}, \dots, n_{\Gamma,d})$ denotes the outward unit normal to the faces $\Gamma$ of $Q_l$. The fluxes across the cell interfaces are now replaced by numerical fluxes $\hat{F}_i : \mathscr{U} \times \mathscr{U} \to \mathbb{R}^m$, $i = 1, \dots, d$, satisfying the following conditions:

1. $\hat{F}_i(\cdot,\cdot)$ is locally Lipschitz continuous
2. $\hat{F}_i$ is consistent: $\hat{F}_i(u,u) = F_i(u)$, for all $u \in \mathscr{U}$
3. $\hat{F}_i$ is conservative: $\hat{F}_i(u,v) = -\hat{F}_i(v,u)$, for all $u, v \in \mathscr{U}$

In this thesis we consider the upwind numerical flux [111], the Lax-Friedrichs numerical flux [23], the Lax-Wendroff numerical flux [97] and the HLLE numerical flux [36]. We provide a detailed explanation of the numerical fluxes when we describe the numerical experiment in which they are used.

**Remark 3.1.**
*For simplicity we do not discuss the evaluation of the numerical flux at the boundary, i.e. for states $u_\Gamma^-$, resp. $u_\Gamma^+$ for any $\Gamma \in \mathscr{F}_h^B$. A detailed discussion for different boundary conditions and their numerical treatment can be found in [33, Section 8.3].*

For notational convenience we denote by $u_\Gamma^-, u_\Gamma^+$ the spatial traces of $u_h$ as defined in (3.1). With the help of the numerical flux we rewrite (3.3) to

$$\sum_{l=1}^{N_s} \int_{Q_l} \left( \partial_t u_h \cdot v_h - \sum_{i=1}^{d} F_i(u_h) \cdot \partial_{x_i} v_h \right) \mathrm{d}x + \sum_{\Gamma \in \mathscr{F}_h^I} \int_\Gamma \sum_{i=1}^{d} \hat{F}_i(u_\Gamma^-, u_\Gamma^+) n_{\Gamma,i} \cdot [\![ v_h ]\!]_\Gamma + \sum_{\Gamma \in \mathscr{F}_h^B} \int_\Gamma \sum_{i=1}^{d} \hat{F}_i(u_\Gamma^-, u_\Gamma^+) n_{\Gamma,i} \cdot v_h \, \mathrm{d}s = 0, \tag{3.4}$$

for all $v_h \in \mathscr{V}_h^q$.

Letting $\mathscr{L}_{\mathscr{V}_h^q}$ denote the $L^2$-projection into the DG space $\mathscr{V}_h^q$, the semi-discrete initial-value problem of (1.1) reads as follows: find $u_h \in C^1([0,T); \mathscr{V}_h^q)$, satisfying

$$\begin{cases} \sum_{l=1}^{N_s} \int_{Q_l} \partial_t u_h \cdot v_h \, \mathrm{d}x &= \sum_{l=1}^{N_s} \int_{Q_l} \sum_{i=1}^{d} F_i(u_h) \cdot \partial_{x_i} v_h \, \mathrm{d}x \\ &\quad - \sum_{\Gamma \in \mathscr{F}_h^I} \int_\Gamma \sum_{i=1}^{d} \hat{F}_i(u_\Gamma^-, u_\Gamma^+) n_{\Gamma,i} \cdot [\![ v_h ]\!]_\Gamma \, \mathrm{d}s \\ &\quad - \sum_{\Gamma \in \mathscr{F}_h^B} \int_\Gamma \sum_{i=1}^{d} \hat{F}_i(u_\Gamma^-, u_\Gamma^+) n_{\Gamma,i} \cdot v_h \, \mathrm{d}s \\ u_h(t=0) &= \mathscr{L}_{\mathscr{V}_h^q} u^0, \end{cases} \tag{DG}$$

for all $v_h \in \mathscr{V}_h^q$. The volume and surface integrals in (DG) are approximated using numerical quadrature. Since the interpolation nodes are chosen to be Gaussian quadrature points, the

discrete orthogonality of property of the Lagrange polynomials yields an efficient evaluation of the corresponding integrals, cf. [63].

The initial-value problem (DG) is advanced in time by a $R$-th order strong stability preserving Runge–Kutta (SSP RK) method [65, 94]. To this end we let $0 = t_0 < t_1 < \ldots < t_{N_t} = T$ be a (non-equidistant) temporal decomposition of $[0,T]$. To ensure stability, the explicit time-stepping scheme has to obey the following CFL-type condition

$$\Delta t \leq C \frac{h_{\min}}{\lambda_{\max}(2q+1)}, \tag{3.5}$$

where $\lambda_{\max}$ is the maximal characteristic speed (cf. Definition 2.1) and $C \in (0,1]$. Furthermore, we let $\Lambda\Pi_h : \mathbb{R}^m \to \mathbb{R}^m$ be a suitable limiter, where we use either the TVBM minmod slope limiter from [24] or a Finite-Volume subcell limiter from [95]. In writing down the method we denote by $L_h(u_h(t,\cdot))$ the right-hand side of (DG), with the operator $L_h : \mathscr{V}_h^q \to \mathscr{V}_h^q$ being defined appropriately. The complete $S$-stage time-marching algorithm for given $n$-th time-iterate $u_h^n := u_h(t_n,\cdot) \in \mathscr{V}_h^q$ can then be written as follows:

**Algorithm 1** `Runge-Kutta time-step of (DG)`

1: Set $u_h^{(0)} = u_h^n(t_n)$.
2: **for** $s = 1,\ldots,S$ **do**
3: Compute: $u_h^{(s)} = \Lambda\Pi_h\Big(\sum_{l=0}^{s-1} \alpha_{sl} w_h^{sl}\Big), \quad w_h^{sl} = u_h^{(l)} + \frac{\beta_{sl}}{\alpha_{sl}} \Delta t_n L_h(u_h^{(l)})$.
4: **end for**
5: Set $u_h^{(n+1)}(t_{n+1}) = u_h^{(S)}$.

The initial condition $u_h^{(0)}$ also has to be limited by $\Lambda\Pi_h$. The parameters $\alpha_{sl}$ satisfy the conditions $\alpha_{sl} \geq 0$, $\sum_{l=0}^{s-1} \alpha_{sl} = 1$ , and if $\beta_{sl} \neq 0$, then $\alpha_{sl} \neq 0$ for all $s = 1,\ldots,S$, $l = 0,\ldots,s$.

## 3.2 The Finite Volume method

Based on the semi-discrete formulation (DG) we immediately derive the FV method by considering constant ansatz functions, i.e. $u_h(t,\cdot) \in \mathscr{V}_h^0$, a.e. $t \in (0,T)$. Setting $v_h = \chi_{Q_l}$, for each $l = 1,\ldots,N_s$ yields (cf. [33, Remark 8.8])

$$\begin{cases} \partial_t \bar{u}_l & = -\frac{1}{|Q_l|} \sum\limits_{\Gamma \in \partial Q_l} |\Gamma| \sum\limits_{i=1}^{d} \hat{F}_i(u_\Gamma^-, u_\Gamma^+) n_{\Gamma,i}, \\ \bar{u}_l(t=0) & = \bar{u}_l^0, \end{cases} \tag{FV}$$

where

$$\bar{u}_l := \frac{1}{|Q_l|} \int_{Q_l} u_h \, \mathrm{d}x$$

and $|\Gamma|$ denotes the $(d-1)$-Lebesgue measure of face $\Gamma \in \partial Q_l$ and $|Q_l|$ is the $d$-dimensional measure of the element $Q_l$. Similar to the RKDG method the initial-value problem (FV) is advanced in time by a $R$-th order strong stability preserving Runge–Kutta (SSP RK) method [65, 94].

## 3.3 A posteriori error estimates for deterministic hyperbolic conservation laws

In this section we shortly review some existing a posteriori error estimates for numerical approximations of the entropy admissible solution using FV or DG. One approach to derive a posteriori error estimates for hyperbolic conservation laws is to consider the dual problem of the conservation law. With the help of the solution of the discrete dual problem, it is possible to derive a posteriori error bounds for linear output functionals, see for example [60, 96] in the context of FV and RKDG. We do not pursue this approach in this thesis. A different paradigm to derive a posteriori error bounds, on which we focus in this thesis, are residual-based methods which we directly apply to the solution of the original conservation law. We view the numerical approximation as a solution of a perturbed problem of the underlying conservation law, where the perturbation is given by a computable residual. Using an appropriate stability analysis for the problem at hand one can bound the difference between the exact and numerical approximation in terms of the residual. For scalar conservation laws based on Kružkov's theory the $L^1$-setting is the suitable stability framework to derive a posteriori error estimates for FV schemes or modified RKDG schemes, cf. [29, 54, 71, 84]. We shortly recall the main theorem from [71, Thm. 2.11, resp. Cor. 2.14] for a FV discretization for scalar conservation laws, where for simplicity we restrict ourselves to one spatial dimension.

For the following definitions we assume that the constants $R, \omega, T$ are given and we specify their values in Theorem 3.2 below. We define the following sets

$$\begin{aligned}
I_0 &:= \Big\{ n \mid 0 \le t_n \le \min\{\frac{R+1}{\omega}, T\} \Big\}, \\
D_{R+1} &:= \Big\{ (x,t) \mid |x - x_0| + \omega t < R+1 \Big\}, \\
M(t) &:= \Big\{ j \mid \text{ there exists } x \in Q_j \text{ such that } (x,t) \in D_{R+1} \Big\}.
\end{aligned}$$

**Theorem 3.2** (A posteriori error estimate for deterministic scalar conservation laws)**.**
*Let $u$ be an entropy admissible weak solution of* (1.1) *and let $u_h$ be its numerical approximation using the FV method. Furthermore, let $\hat{F}$ be a consistent, conservative and Lipschitz continuous numerical flux with Lipschitz constant $L$. Moreover, assume that the numerical flux is monotone, i.e. $\partial_u \hat{F}(u,v) \geq 0$, $\partial_v \hat{F}(u,v) \leq 0$, and let $u^0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$ with constants $A \leq u^0(x) \leq B$ a.e. $x \in \mathbb{R}$.*

*Let $K \subset\subset \mathbb{R} \times \mathbb{R}_+$, $\omega = \sup_{A \leq u \leq B} F'(u)$ and choose $R, T$ and $x_0$ such that $T \in (0, \frac{R}{\omega})$ and*

$$K \subset \bigcup_{0 \leq t \leq T} B_{R-\omega t}(x_0) \times \{t\}.$$

*Then there exists computable constants $a, a_0, b$, such that the following estimate holds*

$$\int_K |u - u_h| \, \mathrm{d}x \, \mathrm{d}t \leq T \Big[ a_0 \int_{|x-x_0|<R+1} |u^0(x) - u_h(0,x)| \, \mathrm{d}x + aQ + 2\sqrt{bQ} \Big],$$

*where*

$$Q := \sum_{n \in I_0} \sum_{j \in M(t_n)} \Delta t_n h_j |u_j^{n+1} - u_j^n| + 4L \sum_{n \in I_0} \sum_{j,l \in E(t_n)} \Delta t_n (\Delta t_n + h_{jl}) |u_j^n - u_l^n|)$$

*and $E(t_n)$ is the set of all edges which lie in $M(t_n)$. In the sum over $E(t_n)$ the indices $j, l$ refer to $Q_j, Q_l$ such that $Q_j \cap Q_l$ is the corresponding edge and $h_{jl} = \max\{h_j, h_l\}$.*

The proof of Theorem 3.2 relies heavily on the doubling of variables argument of Kružkov and is therefore not generalizable to systems of conservation laws, where most systems are equipped with only one entropy/entropy flux pair. Up to our knowledge the only available stability framework which allows for the derivation of a residual-based a posteriori error estimate for systems of hyperbolic conservation laws (endowed with a strictly convex entropy) is the relative entropy framework of Dafermos and DiPerna [26], which we have presented in Section 2.4 in the context of random conservation laws. The authors of [28, 46] used the relative entropy method to derive an a posteriori error estimate for the difference between the entropy admissible solution and a reconstruction of the numerical solution obtained with the RKDG method. Because the relative entropy is equivalent to the $L^2$-norm on compact and convex subsets, the error between the exact solution and the reconstruction is bounded in the $L^2$-norm. For the sake of completeness we write down the main theorem from [28, Thm. 25] but we do not elaborate upon the reconstruction process. We explain the temporal and spatial reconstruction process in detail in Section 5.1.1. Let us assume that we have a reconstructed numerical approximation

$$\hat{u}^{st} \in W_\infty^1((0,T); \mathscr{V}_h^{q+1} \cap C^0(D; \mathscr{U}))$$

of the numerical approximation $\{u_h^n\}_{n=0}^{N_t} \subset \mathscr{V}_h^q$ at hand. This allows us to define the following residual

$$\mathscr{R}^{st} := \partial_t \hat{u}^{st} + \partial_x F(\hat{u}^{st}).$$

We can then derive the following a posteriori error estimate.

**Theorem 3.3** (A posteriori error estimate for deterministic systems of conservation laws)**.** *Let the reconstruction $\hat{u}^{st}$ of the numerical approximation $\{u_h^n\}_{n=0}^{N_t}$ of the entropy admissible weak solution only take values in a compact and convex set $\mathscr{C} \subset \mathscr{U}$. Then we have the following estimate for all $n = 0, \dots, N_t$.*

$$\begin{aligned}\|u(t_n,\cdot) - u_h^n(\cdot)\|_{L^2(D)}^2 \leq\ & 2\|\hat{u}^{st}(t_n,\cdot) - u_h^n(\cdot)\|_{L^2(D)}^2 \\ & + 2C_{\underline{\eta}}^{-1}\left(\|\mathscr{R}^{st}\|_{L^2((0,t_n)\times D)}^2 + \|u^0(\cdot,\cdot) - \hat{u}^{st}(0,\cdot,\cdot)\|_{L^2(D)}^2\right) \\ & \times \exp\left(C_{\underline{\eta}}^{-1}\int_0^{t_n}\left(C_{\overline{\eta}}C_{\overline{F}}\|\partial_x\hat{u}^{st}(t,\cdot)\|_{L^\infty(D)} + C_{\overline{\eta}}^2\right)\,\partial_t\right),\end{aligned}$$

*$C_{\underline{\eta}}, C_{\overline{\eta}}, C_{\overline{F}} > 0$ are constants depending on the flux function $F$ and the entropy $\eta$.*

The a posteriori error analysis that we derive in Chapter 5 can be regarded as generalization of Theorem 3.3 to random conservation laws.

Now that we have the necessary spatio-temporal discretization at hand, we can introduce different UQ method for discretization of the random space $\Xi$. Their description, discussion of different issues, resp. their advantages and disadvantages is part of the upcoming chapters.

# 4 Non-statistical Uncertainty Quantification methods

This chapter is devoted to the discussion of common non-statistical, i.e. polynomial-based, UQ methods. On the one hand we consider the Stochastic Galerkin (SG) method and on the other hand, we review the Non-Intrusive Spectral Projection (NISP) and Stochastic Collocation (SC) method. Furthermore, we introduce the Multi-Element method from [87, 101], which decomposes the random space into multiple elements (so called Multi-Elements) and the approximation is only required to be polynomial on each element, which helps to deal with Gibbs oscillations arising from interpolating or projecting discontinuous data.

A major drawback of the SG method when applied to nonlinear systems of hyperbolic conservation laws, is that the resulting SG flux Jacobian may have complex eigenvalues and therefore the SG system loses its hyperbolicity [87]. To ensure that the underlying SG system remains hyperbolic, we present in Section 4.2 a novel hyperbolicity-preserving numerical scheme for a SG-DG discretization of the random conservation law. We apply our numerical scheme to different challenging one- and two-dimensional Riemann problems for which the plain SG method fails.

We end this chapter with a comparison of the efficiency of SG and NISP methods by means of a smooth manufactured solution. The comparison demonstrates that the SG method is only competitive up to two random dimensions.

## 4.1 Review of the employed non-statistical UQ methods

Let us start with an overview of the non-statistical UQ methods which we use in this thesis. When we talk about non-statistical UQ methods we refer to UQ methods which rely on a polynomial representation of the random field. A major class of polynomial-based methods relies on the Polynomial Chaos Expansion (PCE), whose foundations have been laid in [103] and can

be described as a polynomial approximation of Gaussian random variables to represent random processes. This approach has been generalized to a broader class of distributions by Xiu and Karniadakis in [109] by considering polynomials from the so-called Askey-scheme, where the approximating polynomials are orthonormal polynomials with respect to an inner product induced by the probability density function of the chosen probability distribution. The intrusive spectral projection, also known as Stochastic Galerkin (SG) approach, considers a weak formulation of (2.1) with respect to the random variables and uses the corresponding orthonormal polynomials as ansatz and test functions to derive a highly coupled deterministic system of conservation laws. Solving the deterministic system for the unknown modes yields an approximation of the underlying random field. The method is called intrusive because one has to rewrite existing numerical solvers to approximate the modified conservation law.

Another approach to compute the deterministic modes in the PCE expansion is called Non-Intrusive Spectral projection (NISP). It solves deterministic versions of (2.1) at so-called collocation points in random space and and computes the modes of the solution via a pseudo-spectral projection, i.e. via numerical quadrature, cf. [77, 107]. Apart from approximating the modes via numerical quadrature it is also possible to compute the modes using a least square fit, as described in [76, Section 3.5]. However, we do not consider this approach in this thesis. Stochastic numerical schemes which require to solve the random conservation law at prescribed collocation points are in general called non-intrusive methods, because any suitable existing deterministic numerical solver can be used as "black-box" solver to approximate the collocated deterministic problems. A major advantage of the NISP method compared to SG is that it can effectively dampen the curse of dimensionality, which occurs for a high-dimensional random space. In contrast to SG where evaluating the modified modified flux function of the SG system becomes prohibitively expensive for high-dimensional random inputs, non-intrusive method can use collocation points along sparse grids [19] to drastically reduce the number of evaluations of (2.1).

While the NISP and SG method are based on the PCE expansion of the underlying random field, the SC method approximates the random field by a series of Lagrange polynomials. Unlike NISP and SG, the SC method does not aim at computing the corresponding modes via a projection approach, but rather uses interpolation to approximate the underlying random field. The polynomial interpolation is chosen in such a way, that it satisfies (2.1) at a prescribed set of collocation points, cf. [107, 108], resulting in a set of collocated deterministic problems.

### 4.1.1 The Stochastic Galerkin Method

Under the assumption that the solution of (2.1) has a finite second moment we expand $u$ into a generalized Fourier series using a suitable orthonormal basis.

**Remark 4.1.**
*For scalar conservation laws Theorem 2.7 implies that if $u^0 \in L^2_w(\Xi; L^1(D))$ it follows that $u \in L^2_w(\Xi; L^1((0,T)\times D))$ hence, u has a finite second moment. For nonlinear systems of conservation laws we need to assume that the solution of* (2.1) *has a finite second moment.*

For a multi-dimensional random space $\Xi \subset \mathbb{R}^N$, we define $\Xi_j := \xi_j(\Omega)$ and let $\{\Psi^j_k(\cdot)\}_{k\in\mathbb{N}_0} : \Xi_j \to \mathbb{R}$ be a $L^2_{w_j}(\Xi_j)$-orthonormal basis with respect to the one-dimensional density function $w_j$, i.e. for $k,l \in \mathbb{N}_0$ we have

$$\left\langle \Psi^j_k, \Psi^j_l \right\rangle := \mathbb{E}\big(\Psi^j_k \Psi^j_l\big) = \int_{\mathbb{R}} \Psi^j_k(y)\Psi^j_l(y)\; w_j(y)\mathrm{d}(y) = \delta_{k,l}, \tag{4.1}$$

for $j = 1,\dots,N$.

**Remark 4.2.**
*For a uniformly distributed random variable we use Legendre orthogonal polynomial and for a normally distributed random variable we consider Hermite polynomials , cf. [109].*

For any multi-index $\mathrm{k} = (\mathrm{k}_1,\dots,\mathrm{k}_N)^\top \in \mathbb{N}_0^N$ we define the following multivariate polynomials

$$\Psi_{\mathrm{k}}(y) := \Psi^1_{\mathrm{k}_1}(y_1)\cdot\ldots\cdot\Psi^N_{\mathrm{k}_N}(y_N).$$

Following [37, 109], the random entropy solution $u$ of (2.1) can be written as

$$u(t,x,y) = \sum_{\mathrm{k}\in\mathbb{N}_0^N} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y). \tag{4.2}$$

The deterministic Fourier modes $u_{\mathrm{k}} = u_{\mathrm{k}}(t,x)$ in (4.2) are defined by

$$u_{\mathrm{k}}(t,x) = \mathbb{E}\big(u(t,x,\cdot)\Psi_{\mathrm{k}}(\cdot)\big) \qquad \forall\, \mathrm{k} \in \mathbb{N}_0^N. \tag{4.3}$$

From the Fourier modes (4.3) we immediately extract expectation and variance of $u$ via

$$\mathbb{E}\big(u(t,x,\cdot)\big) = u_0(t,x) \text{ and } \mathrm{Var}\big(u(t,x,\cdot)\big) = \sum_{\mathrm{k}\in\mathbb{N}^N\setminus(0,\dots,0)}^{\infty} u_{\mathrm{k}}(t,x)^2.$$

To derive the SG system of (2.1) we first truncate the infinite series in (4.2). As a finite-dimensional basis we consider the complete polynomial space of degree $K \in \mathbb{N}_0$, described

by the following index set

$$\mathscr{K} := \{\mathrm{k} = (\mathrm{k}_1, \dots \mathrm{k}_N)^\top \in \mathbb{N}_0^N \mid \sum_{j=1}^{N} \mathrm{k}_j \leq K\}, \tag{4.4}$$

cf. [108]. The corresponding polynomial approximation space is defined as

$$\mathscr{W}_K(\Xi) := \bigotimes_{\mathrm{k}\in\mathscr{K}} \mathscr{P}_{\mathrm{k}_j}(\Xi_j), \quad \mathscr{P}_{\mathrm{k}_j}(\Xi_j), := \{p : \Xi_j \to \mathbb{R}^m \mid p \text{ is a polynomial of degree } \mathrm{k}_j\}.$$

Hence, the SG approximation reads as follows

$$u(t,x,y) \approx \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \tag{4.5}$$

Plugging the ansatz (4.5) into the random conservation law (2.1) and testing against the same orthonormal basis functions yields the following SG system.

$$\int_\Xi \partial_t \Big( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \Big) \Psi_{\mathrm{l}}(y)\, w(y)\mathrm{d}y + \int_\Xi \sum_{i=1}^{d} \partial_{x_i} F_i \Big( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \Big) \Psi_{\mathrm{l}}(y)\, w(y)\mathrm{d}y = 0, \tag{4.6}$$

for all $\mathrm{l} \in \mathscr{K}$. Using the orthogonality relation (4.1) yields

$$\partial_t \underbrace{\begin{pmatrix} u_0 \\ \vdots \\ u_{N_K} \end{pmatrix}}_{=:\mathbf{u}} + \sum_{i=1}^{d} \partial_{x_i} \underbrace{\begin{pmatrix} \int_\Xi F_i \Big( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \Big) \Psi_0\, w(y)\mathrm{d}y \\ \vdots \\ \int_\Xi F_i \Big( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \Big) \Psi_{N_K}\, w(y)\mathrm{d}y \end{pmatrix}}_{=:\mathbf{F}^i(\mathbf{u})} = 0, \tag{SG}$$

where (with a slight abuse of notation) we define $\mathbf{u},\ \mathbf{F}^i(\mathbf{u}) \in \mathbb{R}^{m\cdot N_K}$, where $N_K := \binom{N+K}{N} - 1 = \mathrm{card}(\mathscr{K})$ is the number of basis polynomials of $\mathscr{W}_K(\Xi)$. For existing numerical approximation schemes for the system (SG) and variants thereof, we refer to [20, 31, 75, 87, 91, 99, 106].

Approximating the deterministic system of conservation laws (SG) by the RKDG method from Section 3.1 yields a sequence of numerical modes (to distinguish the numerically computed modes from the modes of the exact solution we add additional hats) $\{\hat{u}_{\mathrm{k}}^n\}_{n=0}^{N_t}$ at points $\{t_n\}_{n=0}^{N_t}$ in time, for all $\mathrm{k} \in \mathscr{K}$. The numerical approximation of the random entropy solution of (2.1) at time $t = t_n$ reads as follows

$$u(t_n,x,y) \approx u_h^n(x,y) := \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}_{\mathrm{k}}^n(x)\Psi_{\mathrm{k}}(y). \tag{4.7}$$

For a one-dimensional spatial domain $D \subset \mathbb{R}$, it has been shown in [87] that for the Euler equations the flux Jacobian

$$\frac{\partial \mathbf{F}}{\partial \mathbf{u}} = \begin{pmatrix} \hat{\mathbf{F}}_{00} & \cdots & \hat{\mathbf{F}}_{0N_K} \\ \vdots & & \vdots \\ \hat{\mathbf{F}}_{N_K 0} & \cdots & \hat{\mathbf{F}}_{N_K N_K} \end{pmatrix} \in \mathbb{R}^{(m\cdot N_K)\times(m\cdot N_K)}, \tag{4.8}$$

where

$$\hat{\mathbf{F}}_{lm} = \int_{\Xi} \frac{\partial F}{\partial u} \left( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \right) \Psi_l(y)\Psi_m(y)\, w(y)\mathrm{d}y, \quad l,m = 0,\ldots,N_K,$$

may have complex eigenvalues and therefore the system (SG) is no longer hyperbolic. This issue is the main motivation for the development of our hyperbolicity-preserving numerical scheme in Section 4.2. Furthermore, due to the polynomial approximation of discontinuous data, Gibbs oscillations may arise and the plain SG approach is also prone to fail. To dampen the Gibbs oscillations we introduce in the following section the Multi-Element approach as for example in [98, 101, 102]. This approach corresponds to a domain decomposition of the random space, where instead of a global interpolation in $\Xi$, we use local interpolations on every random element.

### 4.1.2 The Multi-Element Stochastic Galerkin method

A major drawback of any global approximation approach in $\Xi$ for hyperbolic conservation laws is that discontinuities in physical space also propagate into random space. Due to the Gibbs phenomenon the approximation may therefore oscillate, resulting in a slow convergence, or even failure of convergence, cf. [87, 101]. To overcome this issue, we employ the Multi-Element (ME) approach as presented in [98, 101, 102], i.e. we subdivide $\Xi$ into disjoint elements and consider a local approximation of (2.1) on every random element.

For the ease of presentation we assume that $\Xi = [0,1]^N$, and let $0 = d_1 < d_2 < \ldots < d_{N_\Xi+1} = 1$ be a decomposition of $[0,1]$.

**Remark 4.3.**

*For random variables with unbounded image, for example for a normal distribution $\xi \sim \mathscr{N}(0,1)$, the authors of [102] proposed a strategy to deal with the unbounded support of $\xi$. The idea is to subdivide $\mathbb{R}$ into three elements $(-\infty,-a),[-a,a],(a,\infty)$ and choose $a \in \mathbb{R}$ such that the tail probability satisfies $\mathbb{P}(X \geq a) \leq \varepsilon$ for some small $\varepsilon > 0$. Due to the small probability of the tail elements, one performs the ME method only on $[-a,a]$.*

We define $D_m := [d_m, d_{m+1})$, for $m = 1, \dots, N_\Xi - 1$, and $D_{N_\Xi} := [d_{N_\Xi}, d_{N_\Xi+1}]$. Introducing the tensor-product index-set $\mathscr{M} := \{\mathrm{m} = (m_1, \dots, m_N)^\top \in \mathbb{N}_0^N : m_j \leq N_\Xi,\ j = 1, \dots, N\}$ allows us to define for $\mathrm{m} \in \mathscr{M}$, the Multi-Element $\mathrm{D_m} := \bigtimes_{j=1}^{N} D_{m_j}$. We consider a new local random variable $\xi^{\mathrm{m}} : \xi^{-1}(\mathrm{D_m}) \to \mathrm{D_m}$ defined on the local probability space

$$\Big(\xi^{-1}(\mathrm{D_m}), \mathscr{F} \cap \xi^{-1}(\mathrm{D_m}), \mathbb{P}(\cdot|\xi^{-1}(\mathrm{D_m}))\Big).$$

Using Bayes' rule we compute the local probability density function of $\xi^{\mathrm{m}}$ via

$$w_{\mathrm{D_m}}(y^{\mathrm{m}}) := w(y^{\mathrm{m}}|\xi^{-1}(\mathrm{D_m})) = \frac{w(y^{\mathrm{m}})}{\mathbb{P}(\xi^{-1}(\mathrm{D_m}))}, \qquad y^{\mathrm{m}} \in \mathrm{D_m}, \tag{4.9}$$

where $\mathbb{P}(\xi^{-1}(\mathrm{D_m})) > 0$ for $\mathrm{m} \in \mathscr{M}$ can be assumed w.l.o.g., due to the independence of the corresponding random variables.

**Remark 4.4.**
*When considering uniform distributions, the local probability density function* (4.9) *remains a density function of a uniformly distributed random variable. Hence, shifted Legendre polynomials can readily be used. Other distributions, for example normal distributions, require to numerically compute a set of polynomials which are orthogonal w.r.t.* (4.9)*, cf. [102]. For simplicity we consider only uniform distributions.*

We parametrize the uncertain input in (2.1) using the local random variable $\xi^{\mathrm{m}}$ and consider a local approximation on every $\mathrm{D_m}$. If we let $\{\Psi_{\mathrm{k}}^{\mathrm{m}}(y^{\mathrm{m}})\}_{\mathrm{k} \in \mathbb{N}_0^N}$ be the orthonormal polynomials with respect to the conditional probability density function (4.9), we may consider the local gPC approximation in element $\mathrm{D_m}$,

$$u_{\mathrm{m}}(t,x,y^{\mathrm{m}}) = \sum_{\mathrm{k} \in \mathbb{N}_0^N} u_{\mathrm{k,m}}(t,x)\Psi_{\mathrm{k}}^{\mathrm{m}}(y^{\mathrm{m}}) \approx \sum_{\mathrm{k} \in \mathscr{K}} u_{\mathrm{k,m}}(t,x)\Psi_{\mathrm{k}}^{\mathrm{m}}(y^{\mathrm{m}}), \quad (t,x,y^{\mathrm{m}}) \in (0,T) \times D \times \mathrm{D_m}, \tag{4.10}$$

for all $\mathrm{m} \in \mathscr{M}$. The global approximation can be written as

$$u(t,x,y) = \sum_{\mathrm{m} \in \mathscr{M}} u_{\mathrm{m}}(t,x,y)\chi_{\mathrm{D_m}}(y) \approx \sum_{\mathrm{m} \in \mathscr{M}} \sum_{\mathrm{k} \in \mathscr{K}} u_{\mathrm{k,m}}(t,x)\Psi_{\mathrm{k}}^{\mathrm{m}}(y)\chi_{\mathrm{D_m}}(y), \quad (t,x,y) \in (0,T) \times D \times \Xi, \tag{4.11}$$

where the local approximation (4.11) converges to the global solution in $L^2(\Xi)$ as $N_\Xi, K \to \infty$, cf [2, 101].

**Remark 4.5.**
*Due to the disjoint decomposition of the random space, we can now apply the SG method on every random element* $\mathrm{D_m}$ *for all* $\mathrm{m} \in \mathscr{M}$ *in parallel. This yields a trivial parallelization strategy in random space.*

The expected value of $u$ is given by its moment of zeroth order. We assume that $\Psi_0^{\mathrm{m}}(y^{\mathrm{m}}) = \prod_{j=1}^{N} \Psi_0^{\mathrm{m}}(y_j^{\mathrm{m}}) = 1$ in $\mathrm{D_m}$ and obtain the expected value and variance as weighted sum of the local expected values and variances.

$$\begin{aligned}\mathbb{E}\big(u\big) &\approx \int_{\Xi} \sum_{\mathrm{m}\in\mathscr{M}} \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k,m}} \Psi_{\mathrm{k}}^{\mathrm{m}} \chi_{\mathrm{D_m}}(y)\, w(y)\mathrm{d}y \\ &= \sum_{\mathrm{m}\in\mathscr{M}} \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k,m}} \int_{\mathrm{D_m}} \Psi_{\mathrm{k}}^{\mathrm{m}} \Psi_0^{\mathrm{m}}\, \mathbb{P}(\xi^{-1}(\mathrm{D_m})) w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}} \\ &= \sum_{\mathrm{m}\in\mathscr{M}} \mathbb{P}(\xi^{-1}(\mathrm{D_m}))\, u_{0,\mathrm{m}}, \end{aligned} \tag{4.12}$$

$$\mathrm{Var}\big(u\big) \approx \sum_{\mathrm{m}\in\mathscr{M}} \Big(\mathrm{Var}\big(u_{\mathrm{m}}\big) + \big(u_{0,\mathrm{m}} - \mathbb{E}\big(u\big)\big)^2\Big)\, \mathbb{P}(\xi^{-1}(\mathrm{D_m})), \tag{4.13}$$

where the local variance $\mathrm{Var}\big(u_{\mathrm{m}}\big)$ is given by

$$\mathrm{Var}\big(u_{\mathrm{m}}\big) \approx \int_{\mathrm{D_m}} \Big( \sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k,m}} \Psi_{\mathrm{k}}^{\mathrm{m}} \Big)^2 w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}} - u_{0,\mathrm{m}}^2 = \sum_{\mathrm{k}\in\mathscr{K}\setminus(0,\ldots,0)} u_{\mathrm{k,m}}^2.$$

### 4.1.3 The Non-intrusive Spectral Projection Method

Instead of computing the Fourier modes in (4.3) via an intrusive Galerkin projection, it is possible to approximate the modes using a non-intrusive discrete orthogonal projection. For the derivation of the NISP method we consider for $Q_\Xi \in \mathbb{N}_0$ the following tensor-product multi-index set,

$$\mathscr{Q} := \{\mathrm{q} = (\mathrm{q}_1, \ldots, \mathrm{q}_N) \in \mathbb{N}_0^N \mid \mathrm{q}_j \leq Q_\Xi,\ j = 1, \ldots, N\}$$

and define the multivariate quadrature points and weights as follows

$$y_{\mathrm{q}} := (y_{\mathrm{q}_1}, \ldots, y_{\mathrm{q}_N}), \qquad w_{\mathrm{q}} := w_{\mathrm{q}_1} \cdot \ldots \cdot w_{\mathrm{q}_N}, \tag{4.14}$$

for all $\mathrm{q} \in \mathscr{Q}$. Here $\{y_{\mathrm{q}_j}\}_{\mathrm{q}_j=0}^{Q_\Xi}$, $\{w_{\mathrm{q}_j}\}_{\mathrm{q}_j=0}^{Q_\Xi}$ are corresponding one-dimensional quadrature rules. The choice of the quadrature points and weights is an important aspect of the NISP method and depending on the distribution of the random variable $\xi_j$ we choose the quadrature points as zeros of the corresponding chaos polynomials, cf. [109]. For example, if $\xi_j \sim \mathscr{U}(a,b)$ is uniformly distributed, we choose $\{y_{\mathrm{q}_j}\}_{\mathrm{q}_j=0}^{Q_\Xi}$ to be the zeros of the $(Q_\Xi+1)$-th Legendre polynomial. For a Gaussian distribution we use the zeros of the Hermite polynomials accordingly.

**Remark 4.6.** *1. For a high-dimensional random space $\Xi \subset \mathbb{R}^N$, $N >> 1$ tensor-product structures become computationally infeasible due to the curse of dimensionality. It is therefore necessary to consider sparse grids [19].*

2. *Apart from using Gaussian quadrature points as collocation points, other common choices are quadrature points associated with for example Kronrod-Patterson, Fejér or Clenshaw-Curtis quadrature rules. Especially Kronrod-Patterson and Clenshaw-Curtis quadrature points are very efficient in combination with sparse grids, as they decrease the number of quadrature points significantly due to their nested structure. For an overview of different nodes and weights see for example [38].*

Using the quadrature points as inputs in (2.1) we obtain $\mathrm{card}(\mathscr{Q}) = (Q_\Xi + 1)^N$ uncoupled (collocated) deterministic initial value problems.

$$\begin{cases} \partial_t u(t,x,y_\mathrm{q}) + \sum_{i=1}^{d} \partial_{x_i} F_i(u(t,x,y_\mathrm{q}),y_\mathrm{q}) = 0, & (t,x) \in (0,T)\times D, \\ u(0,x,y_\mathrm{q}) = u^0(x,y_\mathrm{q}), & x \in D. \end{cases} \tag{4.15}$$

The system (4.15) is a deterministic system of hyperbolic conservation laws which can be solved by any suitable deterministic numerical solver.

For every quadrature point $y_\mathrm{q} \in \Xi$ we denote the corresponding numerical approximation at time $t_n = t_n(y_\mathrm{q})$ by $u_h^n(\cdot, y_\mathrm{q})$. Let us assume that the time-partition $\{t_n\}_{n=0}^{N_t}$ is the same for every quadrature point $\{y_\mathrm{q}\}_{\mathrm{q}\in\mathscr{Q}}$. We approximate the Fourier modes in (4.3) at time $t = t_n$ and for $x \in D$ by

$$u_\mathrm{k}(t_n,x) = \mathbb{E}\big(u(t_n,x,\cdot)\Psi_\mathrm{k}(\cdot)\big) \approx \sum_{\mathrm{q}\in\mathscr{Q}} u_h^n(x,y_\mathrm{q})\Psi_\mathrm{k}(y_\mathrm{q})w_\mathrm{q} =: \hat{u}_\mathrm{k}^n(x), \tag{4.16}$$

for all $\mathrm{k} \in \mathscr{K}$, where $\mathscr{K}$ is the index set describing the complete polynomial space, cf. (4.4). The numerical approximation of the random entropy solution of (2.1) at time $t = t_n$ reads as follows

$$u(t_n,x,y) \approx u_h^n(x,y) := \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}_\mathrm{k}^n(x)\Psi_\mathrm{k}(y). \tag{4.17}$$

### 4.1.4 The Stochastic Collocation method

The main difference between SC and NISP is the choice of basis polynomials. While NISP uses orthogonal polynomials and computes the corresponding modes with a pseudo-spectral projection, SC uses Lagrange polynomials for an interpolation in the random space. For the SC method the polynomial interpolant of the random entropy admissible solution is supposed to satisfy (2.1) at prescribed nodal (collocation) points. In the same manner as for NISP, the collocation points are used as input parameters in (2.1) to derive a set of deterministic hyperbolic

conservation laws, which are discretized with a deterministic numerical solver for each given collocation point.

For a multi-dimensional random space $\Xi \subset \mathbb{R}^N$, the most common polynomial approximation spaces are tensor products of one-dimensional interpolating polynomials. We follow [3] to first define the space $\mathscr{P}_K(\Xi)$ of tensor product polynomials of maximal degree $K \in \mathbb{N}_0$ by

$$\mathscr{P}_K(\Xi) := \bigotimes_{j=1}^{N} \mathscr{P}_K(\Xi_j), \quad \mathscr{P}_K(\Xi_j) := \{p : \Xi_j \to \mathbb{R} \mid p \text{ is a polynomial of degree } K\}.$$

**Remark 4.7.**
*An alternative approach to performing interpolation on tensor-product collocation points is a Smolyak sparse grids interpolation cf. [19, 108]. The advantages and disadvantages of the sparse grid collocation are extensively discussed in [107, 108]. A rule of thumb is that tensor-product collocation points are computationally feasible up to five random dimensions [107]. In our numerical experiments we only consider up to three random dimensions and we therefore focus on tensor-product structures.*

For a consistent notation we shall use the same symbol $\mathscr{K}$ for the index set of the SC method as for the SG and NISP method. Hence, we let $\mathscr{K} := \{\mathrm{k} = (k_1, \ldots, k_N)^\top \in \mathbb{N}_0^N : k_j \leq K,\ j = 1 \ldots, N\}$ be the corresponding tensor-product multi-index set and define the collocation points $y_\mathrm{k} = (y_{k_1}, \ldots, y_{k_N})^\top \in \Xi$, $\mathrm{k} \in \mathscr{K}$. Similar to the NISP method we use as one-dimensional collocation points the zeros of the corresponding chaos polynomials.

**Remark 4.8.**
*In [25, Lemma 2] it has been shown that for Gaussian quadrature points the SC interpolant coincides with the NISP approximation as long as the number of terms in the orthogonal series is greater or equal than the number of points in the quadrature rule.*

As a basis of $\mathscr{P}_K(\Xi_j)$ we choose the Lagrange basis $\{l_{k_j}\}_{k_j=0}^K$ associated with the collocation points $\{y_{k_j}\}_{k_j=0}^K$, such that

$$l_{k_j}(y_{o_j}) = \delta_{k_j,o_j}, \qquad \forall\, k_j, o_j = 0, \ldots, K,$$

for all $j = 1, \ldots, N$. We then define the multivariate Lagrange polynomials as

$$l_\mathrm{k}(y_\mathrm{o}) := l_{k_1}(y_{o_1}) \cdot \ldots \cdot l_{k_N}(y_{o_N}), \quad \mathrm{k}, \mathrm{o} \in \mathscr{K}.$$

Using the collocation points $\{y_\mathrm{k}\}_{\mathrm{k} \in \mathscr{K}}$ as input parameters in (2.1) yields $\mathrm{card}(\mathscr{K}) = (K+1)^N$

(uncoupled) collocated initial value problems.

$$\begin{cases} \partial_t u(t,x,y_\mathrm{k}) + \sum_{i=1}^{d} \partial_{x_i} F_i(u(t,x,y_\mathrm{k}),y_\mathrm{k}) = 0, & (t,x) \in (0,T)\times D, \\ u(0,x,y_\mathrm{k}) = u^0(x,y_\mathrm{k}), & x \in D. \end{cases} \tag{4.18}$$

For every collocation point $\{y_\mathrm{k}\}_{\mathrm{k}\in\mathscr{K}}$ we denote the corresponding numerical approximation at time $t_n = t_n(y_\mathrm{k})$ by $u_h^n(\cdot,y_\mathrm{k})$. Let us assume that the time-partition $\{t_n\}_{n=0}^{N_t}$ is the same for every collocation point $\{y_\mathrm{k}\}_{\mathrm{k}\in\mathscr{K}}$. The numerical approximation of the random entropy admissible solution of (2.1) at time $t = t_n$ can then be written as

$$u(t_n,x,y) \approx u_h^n(x,y) := \sum_{\mathrm{k}\in\mathscr{K}} u_h^n(x,y_\mathrm{k}) l_\mathrm{k}(y). \tag{4.19}$$

## 4.2 Hyperbolicity-preserving limiter for SG

As discussed at the end of Section 4.1.1 the flux Jacobian of (SG) may have complex eigenvalues and thus, the SG system loses its hyperbolicity. Therefore, the plain SG approach is not applicable for nonlinear and non-symmetric systems of hyperbolic conservation laws. To remedy this problem we derive in this section a hyperbolicity-preserving limiter for a RKDG discretization of the system (SG) which ensures that the SG system remains hyperbolic after each iteration of our numerical scheme. Results from this section have been published in [34].

For the derivation of the hyperbolicity-preserving numerical scheme we restrict ourselves to a two-dimensional spatial domain $D \subset \mathbb{R}^2$ and consider random variations in initial data. Thus, (2.1) becomes

$$\begin{cases} \partial_t u(t,x,y) + \partial_{x_1} F_1(u(t,x,y)) + \partial_{x_2} F_2(u(t,x,y)) = 0, & (t,x,y) \in (0,T)\times D\times \Xi, \\ u(0,x,y) = u^0(x,y), & (x,y) \in D\times\Xi. \end{cases} \tag{4.20}$$

Appropriate boundary conditions for (4.20) will be specified in the corresponding numerical experiments in Section 4.2.2.

To sketch the main idea of the hyperbolicity-preserving numerical scheme we first derive a modified CFL condition, such that the (space-stochastic) cell-mean is preserved after one time-step of forward Euler. This is the statement of Theorem 4.16. Using a (space-stochastic) slope-limiter, which limits the numerical approximation at the new time-step towards the admissible cell-mean, we provide a pointwise admissible SG approximation. Let us first recall the definition of the hyperbolicity set for (4.20) which is similar to Definition 2.1.

**Definition 4.9** (Hyperbolicity set)**.**
*We call the set*

$$\mathscr{H} := \{u \in \mathbb{R}^m \;\Big|\; \alpha_1 \frac{\partial F_1(u)}{\partial u} + \alpha_2 \frac{\partial F_2(u)}{\partial u} \text{ has m real, distinct eigenvalues, for all } \alpha \in S^3\}$$

*the hyperbolicity set and every state $u \in \mathscr{H}$ is called admissible.*

**Assumption 4.10.**
*In the following we assume that the hyperbolicity set $\mathscr{H}$ is open and convex.*

Moreover, we recall the following theorem from [106], which shows that admissible states generate a hyperbolic SG system.

**Theorem 4.11** ([106], Theorem 2.1)**.**
*If the SG approximation $\sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y)$ of* (4.20) *is admissible, i.e.*

$$\sum_{\mathrm{k}\in\mathscr{K}} u_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y) \in \mathscr{H} \qquad \text{a.e. } (t,x,y) \in (0,T)\times D\times \Xi,$$

*then the system* (SG) *is hyperbolic.*

In a first step to derive the hyperbolicity-preserving numerical scheme we consider a weak formulation of (4.20) with respect to physical and random variables. Similar to the ME method from Section 4.1.2 we partition the spatial domain $D \subset \mathbb{R}^2$ into a uniform rectangular mesh with cells $C_{r,s} = [x_{1,r-\frac{1}{2}}, x_{1,r+\frac{1}{2}}] \times [x_{2,s-\frac{1}{2}}, x_{2,s+\frac{1}{2}}]$ and cell-widths $\Delta x_1 := (x_{1,r+\frac{1}{2}} - x_{1,r-\frac{1}{2}})$, $\Delta x_2 := (x_{2,s+\frac{1}{2}} - x_{2,s-\frac{1}{2}})$. We test (4.20) with test function $v(x_1,x_2,y)$, where $\mathrm{supp}\,(v) \subseteq C_{r,s}\times \mathrm{D_m}$. The weak formulation of (4.20) reads as

$$\partial_t \int_{C_{r,s}}\int_{\mathrm{D_m}} u\, v\, w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}}\mathrm{d}(x_1,x_2) = \int_{C_{r,s}}\int_{\mathrm{D_m}} \Big(F_1(u)\,\partial_{x_1} v + F_2(u)\,\partial_{x_2} v\Big)\, w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}}\mathrm{d}(x_1,x_2) \tag{4.21}$$

$$- \int_{x_{2,s-\frac{1}{2}}}^{x_{2,s+\frac{1}{2}}}\int_{\mathrm{D_m}} \Big[F_1(u)v\Big]_{x_{1,r-\frac{1}{2}}}^{x_{1,r+\frac{1}{2}}}\, w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}}\mathrm{d}x_2 \tag{4.22}$$

$$- \int_{x_{1,r-\frac{1}{2}}}^{x_{1,r+\frac{1}{2}}}\int_{\mathrm{D_m}} \Big[F_2(u)v\Big]_{x_{2,s-\frac{1}{2}}}^{x_{2,s+\frac{1}{2}}}\, w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}}\mathrm{d}x_1. \tag{4.23}$$

The numerical approximation $u_h$ of (4.21)-(4.23) is then sought in $\mathscr{W}_K(\Xi)\otimes\mathscr{V}_h^q$. The local approximation in space-stochastic cell $C_{r,s}\times\mathrm{D_m}$ reads as follows (we suppress the discretization

parameter $h$ for brevity)

$$u_{r,s,\mathrm{m}}(t,x_1,x_2,y^{\mathrm{m}}) := u\big|_{C_{r,s}\times \mathrm{D_m}}(t,x_1,x_2,y^{\mathrm{m}}) = \sum_{\kappa\in\mathscr{J}}\sum_{\mathrm{k}\in\mathscr{K}} u_{\kappa,\mathrm{k},r,s,\mathrm{m}}(t)\varphi_{\kappa}^{r,s}(x_1,x_2)\Psi_{\mathrm{k}}^{\mathrm{m}}(y^{\mathrm{m}}). \tag{4.24}$$

here $\{\varphi_{\kappa}^{r,s}\}_{\kappa\in\mathscr{J}}$ are DG polynomials on cell $C_{r,s}$ and $\mathscr{J} := \{\kappa = (\kappa_1,\kappa_2)^T \in \mathbb{N}_0^2 \mid \kappa_i \le q,\ i = 1,2\}$ denotes a two-dimensional tensor-product index set. For simplicity we do not insert (4.24) into (4.20) and write down the semi-discrete version of (4.21)-(4.23) but postpone it to the proof of Theorem 4.16.

An important aspect of the hyperbolicity-preserving numerical scheme is the approximation of the integrals in (4.21)-(4.23). We use either Gauß–Legendre or Gauß–Lobatto quadrature rules. In this section we denote Gauß–Legendre quadrature rules with Latin letters and Gauß–Lobatto quadrature rules with Greek letters and add additional hats to the points and weights. To approximate the integral (4.21) on the spatial cell $C_{r,s}$ we use a tensor-product of one-dimensional Gauß–Lobatto quadratures with $Q_D+1 = \lceil\frac{q+1}{2}\rceil+1$ points and weights, denoted by $(\hat{x}_{1,\alpha},\hat{w}_\alpha)$ and $(\hat{x}_{2,\beta},\hat{w}_\beta)$ respectively, for $\alpha,\beta = 0,\dots,Q_D$.

**Remark 4.12.**
*The Gauß–Lobatto quadrature rule includes the endpoints, i.e. cell interfaces. This property will be used in the proof of Theorem 4.16.*

To approximate the (spatially) one-dimensional integrals in (4.22) and (4.23), we use Gauß–Legendre quadratures with sufficient accuracy. Let us assume that we employ Gauß–Legendre quadratures with $Q_L+1 \in \mathbb{N}$ points and weights denoted by $(x_{1,p},w_p)$ and $(x_{2,q},w_q)$ respectively, for $p,q = 0,\dots,Q_L$.

For a uniformly-distributed random variable and Legendre basis functions, we apply a Gauß–Lobatto quadrature rule on $\mathrm{D_m}$ with $Q_\Xi+1 = \lceil\frac{K+1}{2}\rceil+1$ points and weights $(\hat{y}_{\mathrm{q}},\hat{\omega}_{\mathrm{q}})$, $\mathrm{q} = 0,\dots,Q_\Xi$. For a multi-dimensional random space $\Xi\subset\mathbb{R}^N$ we also use tensor-product quadrature rules and introduce the multi-index set

$$\mathrm{q}\in\mathscr{Q} := \{\mathrm{q} = (\mathrm{q}_1,\dots,\mathrm{q}_N)^T \in \mathbb{N}_0^N \mid \mathrm{q}_j \le Q_\Xi,\ j = 1,\dots,N\}. \tag{4.25}$$

For other probability distributions we use the corresponding Gauß quadrature based on the orthogonal basis polynomials and weighted by the probability density function $w_{\mathrm{D_m}}$. We scale the quadrature weights such that

$$\sum_{\alpha=0}^{Q_D}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\hat{w}_\alpha w_q\hat{\omega}_{\mathrm{q}} = 1,\quad \sum_{p=0}^{Q_L}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}} w_p\hat{w}_\beta\hat{\omega}_{\mathrm{q}} = 1,\quad \int_{\mathrm{D_m}} u\, w_{\mathrm{D_m}}(y^{\mathrm{m}})\mathrm{d}y^{\mathrm{m}} \approx \sum_{\mathrm{q}\in\mathscr{Q}} u(\hat{y}_{\mathrm{q}})\hat{\omega}_{\mathrm{q}}. \tag{4.26}$$

Since the numerical approximation $u_{r,s,\mathrm{m}}(t,x_1,x_2,y^{\mathrm{m}})$ is discontinuous across the spatial cell interfaces $x_{1,r\pm\frac{1}{2}}$, $x_{2,s\pm\frac{1}{2}}$, we replace the evaluation of the flux components $F_1$, $F_2$ at these points with numerical flux functions $\hat{F}_1(\cdot,\cdot),\hat{F}_2(\cdot,\cdot)$, obtained by (approximately) solving the corresponding Riemann problem at the interface. To this end we denote the spatial traces by

$$
\begin{aligned}
u_{r,s,\mathrm{m}}(x^-_{1,r+\frac{1}{2}},x_2,y^{\mathrm{m}}) &:= \lim_{x_1\uparrow x_{1,r+\frac{1}{2}}} u|_{C_{r,s}\times \mathrm{D_m}}(x_1,x_2,y^{\mathrm{m}}),\\
u_{r,s,\mathrm{m}}(x^+_{1,r+\frac{1}{2}},x_2,y^{\mathrm{m}}) &:= \lim_{x_1\downarrow x_{1,r+\frac{1}{2}}} u|_{C_{r,s}\times \mathrm{D_m}}(x_1,x_2,y^{\mathrm{m}}),\\
u_{r,s,\mathrm{m}}(x_1,x^-_{2,s+\frac{1}{2}},y^{\mathrm{m}}) &:= \lim_{x_2\uparrow x_{2,s+\frac{1}{2}}} u|_{C_{r,s}\times \mathrm{D_m}}(x_1,x_2,y^{\mathrm{m}}),\\
u_{r,s,\mathrm{m}}(x_1,x^+_{2,s+\frac{1}{2}},y^{\mathrm{m}}) &:= \lim_{x_2\downarrow x_{2,s+\frac{1}{2}}} u|_{C_{r,s}\times \mathrm{D_m}}(x_1,x_2,y^{\mathrm{m}}).
\end{aligned}
$$

A main ingredient of the hyperbolicity-preserving numerical scheme and the proof of Theorem 4.16 are positivity-preserving numerical fluxes as described in [112]. For a simple definition of a positivity-preserving numerical flux let us first assume that $\bar{u}^n_r$ is the approximation of the cell average of an exact solution $u(t,x)$ in a one-dimensional cell $[x_{r-\frac{1}{2}},x_{r+\frac{1}{2}}]$ at time $t=t_n$. After one time-step of forward Euler we obtain at time $t=t_{n+1}$ the updated cell-mean

$$
\bar{u}^{n+1}_r = \bar{u}^n_r - \frac{\Delta t}{\Delta x_1}\Big(\hat{F}_1(\bar{u}^n_r,\bar{u}^n_{r+1}) - \hat{F}_1(\bar{u}^n_{r-1},\bar{u}^n_r)\Big). \tag{4.27}
$$

**Definition 4.13** (Positivity-preserving numerical flux)**.**
*The scheme* (4.27) *and the numerical flux $\hat{F}_1$ are called positivity-preserving, if $\bar{u}^n_r \in \mathscr{H}$ for all $r$ implies that $\bar{u}^{n+1}_r \in \mathscr{H}$.*

The positivity-preserving property is in general achieved under a suitable CFL condition, see also [112]. In the following we assume that the numerical fluxes $\hat{F}_1,\hat{F}_2$ are positivity-preserving.

**Assumption 4.14.**
*The numerical fluxes $\hat{F}_1,\hat{F}_2$, are positivity-preserving under the following CFL condition*

$$
\lambda^1_{\max}\frac{\Delta t}{\Delta x_1} + \lambda^2_{\max}\frac{\Delta t}{\Delta x_2} \le C,
$$

*where $\lambda^i_{\max}$ is the maximum eigenvalue of the flux Jacobian $\mathrm{D}F_i$, for $i=1,2$ and $C\in(0,1]$.*

**Remark 4.15.**
*The positivity-preserving property from Assumption (4.14) is one ingredient of our hyperbolicity-preserving numerical scheme and the proof of Theorem 4.16. We want to mention two numerical fluxes, used in our numerical experiments, which satisfy Assumption (4.14). The Lax-Friedrichs flux is positivity-preserving with $C=1$ (see [112, Remark 2.4], where the property is shown for*

*the deterministic Euler equations). The HLLE flux also fulfills this property for the deterministic Euler equations (cf. [36]), whereas we use $C = 0.5$ for our numerical results in Section 4.2.2.*

We can now state the main theorem of this section.

**Theorem 4.16** (Hyperbolicity-preservation under a modified CFL condition)**.**
*Let Assumption 4.14 hold and assume that at time $t_n$ all point values satisfy*

$$u_{r,s,\mathrm{m}}(t_n, \hat{x}_{1,\alpha}, x_{2,q}, \hat{y}_\mathrm{q}) \in \mathscr{H} \text{ and } u_{r,s,\mathrm{m}}(t_n, x_{1,p}, \hat{x}_{2,\beta}, \hat{y}_\mathrm{q}) \in \mathscr{H}.$$

*Then the space-stochastic cell-mean*

$$\overline{u}_{r,s,\mathrm{m}} := \frac{1}{\int\limits_{C_{r,s}}\int\limits_{\mathrm{D_m}} 1 \; w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}(x_1,x_2)} \int\limits_{C_{r,s}}\int\limits_{\mathrm{D_m}} u_{r,s,\mathrm{m}}(t,x_1,x_2,y^\mathrm{m}) \; w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}(x_1,x_2) \tag{4.28}$$

*is admissible after one Euler forward time-step under the modified CFL condition*

$$\lambda^1_{\max}\frac{\Delta t}{\Delta x_1} + \lambda^2_{\max}\frac{\Delta t}{\Delta x_2} \leq C\hat{w}_0, \tag{4.29}$$

*where $\hat{w}_0$ is the first weight of the $(Q_D+1)$-point Gauß–Lobatto quadrature rule.*

*Proof.* In the weak formulation (4.21), we use $u = u_{r,s,\mathrm{m}}$ and choose the test function to be $v = \frac{1}{\Delta x_1 \Delta x_2} = \frac{1}{\Delta x_1 \Delta x_2 \Delta y}$, where $\Delta y = \int\limits_{\mathrm{D_m}} w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m} = 1$, because $w_{\mathrm{D_m}}$ is a probability density function on $\mathrm{D_m}$. We note that for well-posedness of the surface integrals in (4.30) it is necessary to use the numerical flux functions, however, to keep notation slim and more clear, we do not introduce the numerical fluxes until (4.31). We formally obtain

$$\begin{aligned}
&\partial_t \int\limits_{C_{r,s}}\int\limits_{\mathrm{D_m}} u_{r,s,\mathrm{m}} \frac{1}{\Delta x_1 \Delta x_2 \Delta y} \; w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}(x_1,x_2) \\
&= \int\limits_{C_{r,s}}\int\limits_{\mathrm{D_m}} \left( F_1(u_{r,s,\mathrm{m}})\frac{1}{\Delta y}\partial_{x_1}\frac{1}{\Delta x_1 \Delta x_2} + F_2(u_{r,s,\mathrm{m}})\frac{1}{\Delta y}\partial_{x_2}\frac{1}{\Delta x_1 \Delta x_2} \right) w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}(x_1,x_2) \\
&- \int\limits_{x_{2,s-\frac{1}{2}}}^{x_{2,s+\frac{1}{2}}}\int\limits_{\mathrm{D_m}} \frac{1}{\Delta x_1 \Delta x_2 \Delta y}\Big[F_1(u)\Big]_{x_{1,r-\frac{1}{2}}}^{x_{1,r+\frac{1}{2}}} \; w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}x_2 \\
&- \int\limits_{x_{1,r-\frac{1}{2}}}^{x_{1,r+\frac{1}{2}}}\int\limits_{\mathrm{D_m}} \frac{1}{\Delta x_1 \Delta x_2 \Delta y}\Big[F_2(u)\Big]_{x_{2,s-\frac{1}{2}}}^{x_{2,s+\frac{1}{2}}} \; w_{\mathrm{D_m}}(y^\mathrm{m})\mathrm{d}y^\mathrm{m}\mathrm{d}x_1.
\end{aligned} \tag{4.30}$$

The definition of the cell-mean, the properties $\partial_{x_1} v = 0$, $\partial_{x_2} v = 0$ and using numerical quadrature

yield

$$\partial_t \overline{u}_{r,s,\mathrm{m}} = -\frac{1}{\Delta x_1}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big(F_1\big(u_{r,s,\mathrm{m}}(t,x_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)-F_1\big(u_{r,s,\mathrm{m}}(t,x_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)\Big)\,w_q\hat{\omega}_\mathrm{q}$$
$$-\frac{1}{\Delta x_2}\sum_{p=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big(F_2\big(u_{r,s,\mathrm{m}}(t,x_{1,p},x_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q})\big)-F_2\big(u_{r,s,\mathrm{m}}(t,x_{1,p},x_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q})\big)\Big)\,w_p\hat{\omega}_\mathrm{q}.$$

We now replace the flux components $F_i$ by the corresponding numerical fluxes $\hat{F}_i$, $i=1,2$, yielding

$$\begin{aligned}\partial_t \overline{u}_{r,s,\mathrm{m}} = &-\frac{1}{\Delta x_1}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big[\hat{F}_1\big(u_{r,s,\mathrm{m}}(t,x^-_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t,x^+_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)\\
&-\hat{F}_1\big(u_{r,s,\mathrm{m}}(t,x^-_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t,x^+_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)\Big]\,w_q\hat{\omega}_\mathrm{q}\\
&-\frac{1}{\Delta x_2}\sum_{p=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big[\hat{F}_2\big(u_{r,s,\mathrm{m}}(t,x_{1,p},x^-_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t,x_{1,p},x^+_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q})\big)\\
&-\hat{F}_2\big(u_{r,s,\mathrm{m}}(t,x_{1,p},x^-_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t,x_{1,p},x^+_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q})\big)\Big]\,w_p\hat{\omega}_\mathrm{q}.\end{aligned}\tag{4.31}$$

The cell-mean evaluated at time-step $t_n$ can either be written as

$$\overline{u}^{(n)}_{r,s,\mathrm{m}} = \sum_{p=0}^{Q_L}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,x_{1,p},\hat{x}_{2,\beta},\hat{y}_\mathrm{q})w_p\hat{w}_\beta\hat{\omega}_\mathrm{q},$$

or as

$$\overline{u}^{(n)}_{r,s,\mathrm{m}} = \sum_{\alpha=0}^{Q_D}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,\hat{x}_{1,\alpha},x_{2,q},\hat{y}_\mathrm{q})\hat{w}_\alpha w_q\hat{\omega}_\mathrm{q}.$$

Let us define $\delta_1 := \lambda^1_{\max}\frac{\Delta t}{\Delta x_1}$, $\delta_2 := \lambda^2_{\max}\frac{\Delta t}{\Delta x_2}$ and $\mu := \delta_1+\delta_2$. This allows us to write $\overline{u}^{(n)}_{r,s,\mathrm{m}}$ as the following convex combination:

$$\begin{aligned}
\overline{u}^{(n)}_{r,s,\mathrm{m}} =& \frac{\delta_1}{\mu}\overline{u}^{(n)}_{r,s,\mathrm{m}} + \frac{\delta_2}{\mu}\overline{u}^{(n)}_{r,s,\mathrm{m}} \qquad (4.32)\\
=& \frac{\delta_1}{\mu}\sum_{\alpha=0}^{Q_D}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,\hat{x}_{1,\alpha},x_{2,q},\hat{y}_\mathrm{q})\hat{w}_\alpha w_q\hat{\omega}_\mathrm{q}\\
&+\frac{\delta_2}{\mu}\sum_{p=0}^{Q_L}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,x_{1,p},\hat{x}_{2,\beta},\hat{y}_\mathrm{q})w_p\hat{w}_\beta\hat{\omega}_\mathrm{q}\\
=& \frac{\delta_1}{\mu}\sum_{\alpha=1}^{Q_D-1}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,\hat{x}_{1,\alpha},x_{2,q},\hat{y}_\mathrm{q})\hat{w}_\alpha w_q\hat{\omega}_\mathrm{q}\\
&+\frac{\delta_1}{\mu}\hat{w}_0\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big(u_{r,s,\mathrm{m}}(t_n,x^+_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})+u_{r,s,\mathrm{m}}(t_n,x^-_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\Big)w_q\hat{\omega}_\mathrm{q},\\
&+\frac{\delta_2}{\mu}\sum_{p=0}^{Q_L}\sum_{\beta=1}^{Q_D-1}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n,x_{1,p},\hat{x}_{2,\beta},\hat{y}_\mathrm{q})w_p\hat{w}_\beta\hat{\omega}_\mathrm{q}\\
&+\frac{\delta_2}{\mu}\hat{w}_0\sum_{p=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big(u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^+_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q})+u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^-_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q})\Big)w_p\hat{\omega}_\mathrm{q}, \qquad (4.33)
\end{aligned}$$

where we used the fact that

$$\hat{w}_0=\hat{w}_{Q_D},\ x^+_{1,r-\frac{1}{2}}=\hat{x}_{1,0}\big|_{C_{r,s}},\ x^-_{1,r+\frac{1}{2}}=\hat{x}_{1,Q_D}\big|_{C_{r,s}},\ x^+_{2,s-\frac{1}{2}}=\hat{x}_{2,0}\big|_{C_{r,s}},\ x^-_{2,s+\frac{1}{2}}=\hat{x}_{2,Q_D}\big|_{C_{r,s}}.$$

One Euler forward time-step of (4.31) reads as follows

$$\begin{aligned}
\overline{u}^{(n+1)}_{r,s,\mathrm{m}} = \overline{u}^{(n)}_{r,s,\mathrm{m}} &- \frac{\Delta t}{\Delta x_1}\sum_{q=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big[\hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n,x^-_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t_n,x^+_{1,r+\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)\\
&\qquad -\hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n,x^-_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t_n,x^+_{1,r-\frac{1}{2}},x_{2,q},\hat{y}_\mathrm{q})\big)\Big]w_q\hat{\omega}_\mathrm{q}\\
&- \frac{\Delta t}{\Delta x_2}\sum_{p=0}^{Q_L}\sum_{\mathrm{q}\in\mathscr{Q}}\Big[\hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^-_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^+_{2,s+\frac{1}{2}},\hat{y}_\mathrm{q})\big)\\
&\qquad -\hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^-_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q}),u_{r,s,\mathrm{m}}(t_n,x_{1,p},x^+_{2,s-\frac{1}{2}},\hat{y}_\mathrm{q})\big)\Big]w_p\hat{\omega}_\mathrm{q}.
\end{aligned}$$

Inserting (4.32) gives

$$
\begin{aligned}
\overline{u}_{r,s,\mathrm{m}}^{(n+1)} = {} & \frac{\delta_1}{\mu} \sum_{\alpha=1}^{Q_D-1} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, \hat{x}_{1,\alpha}, x_{2,q}, \hat{y}_\mathrm{q}) \hat{w}_\alpha w_q \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_1}{\mu} \hat{w}_0 \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big( u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) + u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) \Big) w_q \hat{\omega}_\mathrm{q}, \\
& + \frac{\delta_2}{\mu} \sum_{p=0}^{Q_L} \sum_{\beta=1}^{Q_D-1} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, x_{1,p}, \hat{x}_{2,\beta}, \hat{y}_\mathrm{q}) w_p \hat{w}_\beta \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_2}{\mu} \hat{w}_0 \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big( u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}) + u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}) \Big) w_p \hat{\omega}_\mathrm{q} \\
& - \frac{\delta_1}{\lambda^1_{\max}} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& \qquad\qquad - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \Big] w_q \hat{\omega}_\mathrm{q} \qquad (4.34) \\
& - \frac{\delta_2}{\lambda^2_{\max}} \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& \qquad\qquad - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \Big] w_p \hat{\omega}_\mathrm{q}. \qquad (4.35)
\end{aligned}
$$

Addition and subtraction of $\hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big)$ in (4.34) and of $\hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big)$ in (4.35) yields

$$
\begin{aligned}
\overline{u}_{r,s,\mathrm{m}}^{(n+1)} = {} & \frac{\delta_1}{\mu} \sum_{\alpha=1}^{Q_D-1} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, \hat{x}_{1,\alpha}, x_{2,q}, \hat{y}_\mathrm{q}) \hat{w}_\alpha w_q \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_2}{\mu} \sum_{p=0}^{Q_L} \sum_{\beta=1}^{Q_D-1} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, x_{1,p}, \hat{x}_{2,\beta}, \hat{y}_\mathrm{q}) w_p \hat{w}_\beta \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_1}{\mu} \hat{w}_0 \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big( u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) + u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) \Big) w_q \hat{\omega}_\mathrm{q}, \\
& + \frac{\delta_2}{\mu} \hat{w}_0 \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big( u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}) + u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}) \Big) w_p \hat{\omega}_\mathrm{q} \\
& - \frac{\delta_1}{\lambda^1_{\max}} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& \qquad - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \Big] w_q \hat{\omega}_\mathrm{q} \\
& - \frac{\delta_1}{\lambda^1_{\max}} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& \qquad - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \Big] w_q \hat{\omega}_\mathrm{q} \\
& - \frac{\delta_2}{\lambda^2_{\max}} \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& \qquad - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \Big] w_p \hat{\omega}_\mathrm{q} \\
& - \frac{\delta_2}{\lambda^2_{\max}} \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} \Big[ \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& \qquad - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \Big] w_p \hat{\omega}_\mathrm{q}.
\end{aligned}
\tag{4.36}
$$

Rearranging the previous equation (4.36) yields

$$
\begin{aligned}
\overline{u}_{r,s,\mathrm{m}}^{(n+1)} = {} & \frac{\delta_1}{\mu} \sum_{\alpha=1}^{Q_D-1} \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, \hat{x}_{1,\alpha}, x_{2,q}, \hat{y}_\mathrm{q}) \hat{w}_\alpha w_q \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_2}{\mu} \sum_{p=0}^{Q_L} \sum_{\beta=1}^{Q_D-1} \sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}(t_n, x_{1,p}, \hat{x}_{2,\beta}, \hat{y}_\mathrm{q}) w_p \hat{w}_\beta \hat{\omega}_\mathrm{q} \\
& + \frac{\delta_1}{\mu} \hat{w}_0 \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} w_q \hat{\omega}_\mathrm{q} \Big( u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) \\
& \qquad - \frac{\mu}{\lambda^1_{\max} \hat{w}_0} \big( \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& \qquad\quad - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big)\big) \Big) \\
& + \frac{\delta_1}{\mu} \hat{w}_0 \sum_{q=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} w_q \hat{\omega}_\mathrm{q} \Big( u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) \\
& \qquad - \frac{\mu}{\lambda^1_{\max} \hat{w}_0} \big( \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& \qquad\quad - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big)\big) \Big) \\
& + \frac{\delta_2}{\mu} \hat{w}_0 \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} w_p \hat{\omega}_\mathrm{q} \Big( u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}) \\
& \qquad - \frac{\mu}{\lambda^2_{\max} \hat{w}_0} \big( \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& \qquad\quad - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q})\big)\big) \Big) \\
& + \frac{\delta_2}{\mu} \hat{w}_0 \sum_{p=0}^{Q_L} \sum_{\mathrm{q}\in\mathscr{Q}} w_p \hat{\omega}_\mathrm{q} \Big( u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}) \\
& \qquad - \frac{\mu}{\lambda^2_{\max} \hat{w}_0} \big( \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& \qquad\quad - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big)\big) \Big).
\end{aligned}
$$

Every term of the form

$$
\begin{aligned}
u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}) - \frac{\mu}{\lambda^1_{\max} \hat{w}_0} \Big[ & \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^-_{1,r+\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \\
& - \hat{F}_1\big(u_{r,s,\mathrm{m}}(t_n, x^-_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x^+_{1,r-\frac{1}{2}}, x_{2,q}, \hat{y}_\mathrm{q})\big) \Big]
\end{aligned}
$$

and

$$
\begin{aligned}
u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}) - \frac{\mu}{\lambda^2_{\max} \hat{w}_0} \Big[ & \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s+\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \\
& - \hat{F}_2\big(u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^-_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q}), u_{r,s,\mathrm{m}}(t_n, x_{1,p}, x^+_{2,s-\frac{1}{2}}, \hat{y}_\mathrm{q})\big) \Big]
\end{aligned}
$$

is admissible under the modified CFL condition

$$\lambda^1_{\max}\frac{\Delta t}{\Delta x_1}+\lambda^2_{\max}\frac{\Delta t}{\Delta x_2}\leq C\hat{w}_0,$$

due to Assumption (4.14). Hence, $\overline{u}^{(n+1)}_{r,s,\mathrm{m}}$ is a convex combination of admissible quantities in $\mathscr{H}$ and therefore $\overline{u}^{(n+1)}_{r,s,\mathrm{m}}\in\mathscr{H}$. □

**Remark 4.17.**
*Because SSP RK methods are convex combinations of Euler forward time-steps (see Algorithm 3.1) and $\mathscr{H}$ is a convex set, SSP RK time-stepping methods are also positivity-preserving.*

To ensure that at time $t_n$ all point values satisfy

$$u_{r,s,\mathrm{m}}\big(t_n,\hat{x}_{1,\alpha},x_{2,q},\hat{y}_{\mathrm{q}}\big),u_{r,s,\mathrm{m}}(t_n,x_{1,p},\hat{x}_{2,\beta},\hat{y}_{\mathrm{q}})\in\mathscr{H}$$

for each cell $C_{r,s}\times\mathrm{D_m}$, we define the slope-limited polynomial in $C_{r,s}\times\mathrm{D_m}$ as

$$\Lambda\Pi^{\theta}\left(u_{r,s,\mathrm{m}}\right)(t,x_1,x_2,y^{\mathrm{m}}):=\theta\,\overline{u}_{r,s,\mathrm{m}}+(1-\theta)u_{r,s,\mathrm{m}}(t,x_1,x_2,y^{\mathrm{m}}).\tag{4.37}$$

The variable $\theta$ limits the polynomial towards the (assumed to be) hyperbolic cell-mean. The case $\theta=0$ coincides with the unlimited solution and for $\theta=1$ we have

$$\Lambda\Pi^{\theta=1}(u_{r,s,\mathrm{m}})(t,x_1,x_2,y^{\mathrm{m}})=\overline{u}^{(n)}_{r,s,\mathrm{m}},$$

which is supposed to be hyperbolic. Because of this property and since $\mathscr{H}$ is convex, we can choose

$$\begin{aligned}\hat{\theta}_{r,s,\mathrm{m}}(t_n):=\inf\Big\{\tilde{\theta}\in[0,1]\;\Big|\;&\Lambda\Pi^{\tilde{\theta}}(u_{r,s,\mathrm{m}})(t_n,\hat{x}_{1,\alpha},x_{2,q},\hat{y}_{\mathrm{q}})\in\mathscr{H}\\&\wedge\Lambda\Pi^{\tilde{\theta}}(u_{r,s,\mathrm{m}})(t_n,x_{1,p},\hat{x}_{2,\beta},\hat{y}_{\mathrm{q}})\in\mathscr{H}\\&\forall\,\alpha,\beta=0,\dots,Q_D,\ p,q=0,\dots,Q_L,\ \mathrm{q}=0,\dots,Q_\Xi\Big\}.\end{aligned}$$

Due to the assumption that $\mathscr{H}$ is an open set, we need to modify $\theta$ slightly in order to avoid placing the solution onto the boundary (if the limiter was active). Therefore, we use

$$\theta=\begin{cases}\hat{\theta}, & \text{if } \hat{\theta}=0,\\ \min(\hat{\theta}+\varepsilon,1), & \text{if } \hat{\theta}>0,\end{cases}$$

where $0<\varepsilon$ should be chosen small enough to ensure that the approximation quality is not influenced significantly. For our numerical experiments we have $\varepsilon$ chosen to be $10^{-8}$. Note that

the cell-mean is preserved since

$$\begin{aligned}
&\sum_{\alpha=0}^{Q_D}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}} \Lambda\Pi^{\theta}(u_{r,s,\mathrm{m}})(t_n,\hat{x}_{1,\alpha},\hat{x}_{2,\beta},\hat{y}_{\mathrm{q}})\hat{w}_\alpha\hat{w}_\beta\hat{\omega}_{\mathrm{q}} \\
&\qquad = \theta\,\overline{u}^{(n)}_{r,s,\mathrm{m}}\sum_{\alpha=0}^{Q_D}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}}\hat{w}_\alpha\hat{w}_\beta\hat{\omega}_{\mathrm{q}} + (1-\theta)\sum_{\alpha=0}^{Q_D}\sum_{\beta=0}^{Q_D}\sum_{\mathrm{q}\in\mathscr{Q}} u_{r,s,\mathrm{m}}\big(t_n,\hat{x}_{1,\alpha},\hat{x}_{2,\beta},\hat{y}_{\mathrm{q}}\big)\hat{w}_\alpha\hat{w}_\beta\hat{\omega}_{\mathrm{q}} \\
&\qquad = \overline{u}^{(n)}_{r,s,\mathrm{m}}.
\end{aligned}$$

**Remark 4.18.**

*In the following pseudo-algorithm we slightly abuse notation and write $\Lambda\Pi^{\theta}(u)$ instead of the local polynomials $u_{r,s,\mathrm{m}}$. By this we mean the application of the hyperbolicity limiter on each cell $C_{r,s}\times \mathrm{D}_{\mathrm{m}}$ separately where we obtain independent values of the limiter variable $\theta$.*

A complete Runge–Kutta time-step using the hyperbolicity-preserving limiter is shown in Algorithm 2.

**Algorithm 2** `Runge-Kutta time-step with hyperbolic limiter`

1: Set $v^{(0)} = u^{(n)}$. *# initialization time-step n*
2: **for** $s = 1,\dots,S$ **do** *# loop Runge-Kutta stages*
3: Compute $v^{(s)} = \Lambda\Pi_h\Big(\sum_{l=0}^{s-1}\alpha_{sl}w^{sl}\Big), \quad w^{sl} = v^{(l)} + \frac{\beta_{sl}}{\alpha_{sl}}\Delta t_n L_h(v^{(l)})$. *# spatial limiter*
4: Compute $v^{(s)} = \Lambda\Pi^{\theta}(v^{(s)})$. *# hyperbolic limiter*
5: **end for**
6: Set $u^{(n+1)} = v^{(S)}$ *# solution at time-step n+1*

**Remark 4.19.**

*The initial condition $u^0$ has to be limited by both limiters $\Lambda\Pi_h$ and $\Lambda\Pi^{\theta}$.*

### 4.2.1 Application to the two-dimensional Euler equations

In this section we compute the values of the limiter variable $\theta$ for the two-dimensional compressible Euler equations for the flow of an ideal gas. The Euler equations are given by

$$\left.\begin{aligned}
\partial_t\rho + \partial_{x_1}m_1 &+ \partial_{x_2}m_2 &&= 0,\\
\partial_t m_1 + \partial_{x_1}\left(\frac{m_1^2}{\rho}+p\right) &+ \partial_{x_2}\left(\frac{m_1m_2}{\rho}\right) &&= 0,\\
\partial_t m_2 + \partial_{x_1}\left(\frac{m_1m_2}{\rho}\right) &+ \partial_{x_2}\left(\frac{m_2^2}{\rho}+p\right) &&= 0,\\
\partial_t E + \partial_{x_1}\left((E+p)\frac{m_1}{\rho}\right) &+ \partial_{x_2}\left((E+p)\frac{m_2}{\rho}\right) &&= 0,
\end{aligned}\right\} \tag{4.38}$$

where $\rho$ describes the mass density, $m_1$ and $m_2$ the momentum in $x_1$- and $x_2$-direction and $E$ the energy of the gas. The four equations model the conservation of mass, momentum and energy. The pressure $p$ reads

$$p = (\gamma - 1)\left(E - \frac{1}{2}\frac{\left(m_1^2 + m_2^2\right)}{\rho}\right),$$

with the adiabatic constant $\gamma > 1$. The hyperbolic set is given by

$$\mathscr{H} = \left\{ u = \begin{pmatrix} \rho \\ m_1 \\ m_2 \\ E \end{pmatrix} \;\middle|\; \rho > 0,\ p = (\gamma-1)\left(E - \frac{1}{2}\frac{\left(m_1^2+m_2^2\right)}{\rho}\right) > 0 \right\}.$$

**Lemma 4.20.**
*Let $\widetilde{u} = (\widetilde{\rho}, \widetilde{m}_1, \widetilde{m}_2, \widetilde{E})^T \in \mathscr{H}$ and $u = (\rho, m_1, m_2, E)^T \in \mathbb{R}^4$ be arbitrary. Then the solution of the hyperbolicity limiter problem*

$$\begin{aligned} &\textit{Find } \min_{[0,1]} \theta \\ &\textit{s.t. } \theta\widetilde{u} + (1-\theta)u \in \mathscr{H} \end{aligned}$$

*for the two-dimensional Euler equations* (4.38) *has the solution*

$$\begin{aligned} \theta^\star &= \max\left(h(\theta_1), h(\theta_{2+}), h(\theta_{2-})\right), \\ h(x) &= x \cdot \chi_{[0,1]}(x), \\ \theta_1 &= \frac{\rho}{\rho - \widetilde{\rho}}, \\ \theta_{2\pm} &= \frac{\rho\widetilde{E} - 2\rho E + \widetilde{\rho}E - m_1\widetilde{m}_1 - m_2\widetilde{m}_2 \pm \sqrt{\widetilde{\theta}}}{m_1^2 - 2m_1\widetilde{m}_1 + m_2^2 - 2m_2\widetilde{m}_2 + \widetilde{m}_1^2 + \widetilde{m}_2^2 - 2\rho E + 2\rho\widetilde{E} + 2\widetilde{\rho}E - 2\widetilde{\rho}\widetilde{E}} \end{aligned} \tag{4.39}$$

*where*

$$\begin{aligned} \widetilde{\theta} = {} & \rho^2\widetilde{E}^2 - 2\rho\widetilde{\rho}E\widetilde{E} + 2\rho E\widetilde{m}_1^2 + 2\rho E\widetilde{m}_2^2 - 2\rho\widetilde{E}m_1\widetilde{m}_1 - 2\rho\widetilde{E}m_2\widetilde{m}_2 \\ & + \widetilde{\rho}^2E^2 - 2\widetilde{\rho}Em_1\widetilde{m}_1 - 2\widetilde{\rho}Em_2\widetilde{m}_2 + 2\widetilde{\rho}\widetilde{E}m_1^2 + 2\widetilde{\rho}\widetilde{E}m_2^2 - m_1^2\widetilde{m}_2^2 + 2m_1m_2\widetilde{m}_1\widetilde{m}_2 - m_2^2\widetilde{m}_1^2. \end{aligned}$$

*Proof.* See [34]. □

**Remark 4.21.**
*In the previous lemma, the quantity $\widetilde{u}$ plays the role of the cell-mean $\overline{u}_{r,s,\mathrm{m}}$ in Theorem 4.16, whereas $u$ is given by the point values $u_{r,s,\mathrm{m}}(t, \hat{x}_{1,\alpha}, x_{2,q}, \hat{y}_\mathrm{q})$ and $u_{r,s,\mathrm{m}}(t, x_{1,p}, \hat{x}_{2,\beta}, \hat{y}_\mathrm{q})$.*

### 4.2.2 Numerical experiments using the hyperbolic-preserving numerical scheme

In this section we apply the hyperbolicity-preserving discontinuous stochastic Galerkin scheme (hDSG) to the one- and two-dimensional Euler equations. We examine the convergence of our numerical scheme by means of a manufactured solution and employ our numerical scheme to different Riemann problems for which the classical SG method fails.

For our numerical experiments we choose either the Lax-Friedrichs numerical flux with

- $\hat{F}_1(u^-_{1,r+\frac{1}{2}}, u^+_{1,r+\frac{1}{2}}) := \frac{1}{2}\Big(F_1(u^-_{1,r+\frac{1}{2}}) + F_1(u^+_{1,r+\frac{1}{2}}) - \lambda^1_{\max}(u^+_{1,r+\frac{1}{2}} - u^-_{1,r+\frac{1}{2}})\Big)$,
- $\hat{F}_2(u^-_{2,s+\frac{1}{2}}, u^+_{2,s+\frac{1}{2}}) := \frac{1}{2}\Big(F_2(u^-_{2,s+\frac{1}{2}}) + F_1(u^+_{2,s+\frac{1}{2}}) - \lambda^2_{\max}(u^+_{2,s+\frac{1}{2}} - u^-_{2,s+\frac{1}{2}})\Big)$,

where the numerical viscosity constants $\lambda^1_{\max}, \lambda^2_{\max}$ are taken as the global estimate of the absolute value of the largest eigenvalue of $\frac{\partial F_1(u)}{\partial u}$ and $\frac{\partial F_2(u)}{\partial u}$. Alternatively, we choose the HLLE numerical flux as in [36]

- $\hat{F}_1(u^-_{1,r+\frac{1}{2}}, u^+_{1,r+\frac{1}{2}}) := \frac{b^+_{1,r+\frac{1}{2}} F_1(u^-_{1,r+\frac{1}{2}}) - b^-_{1,r+\frac{1}{2}} F_1(u^+_{1,r+\frac{1}{2}})}{b^+_{1,r+\frac{1}{2}} - b^-_{1,r+\frac{1}{2}}} + \frac{b^+_{1,r+\frac{1}{2}} b^-_{1,r+\frac{1}{2}}}{b^+_{1,r+\frac{1}{2}} - b^-_{1,r+\frac{1}{2}}}(u^+_{1,r+\frac{1}{2}} - u^-_{1,r+\frac{1}{2}})$,

with signal speed estimates

$$b^-_{1,r+\frac{1}{2}} := \{\lambda^1_{\min}(\bar{u}_{1,r+\frac{1}{2}}), \lambda^1_{\min}(u^-_{1,r+\frac{1}{2}}), 0\},$$
$$b^+_{1,r+\frac{1}{2}} := \{\lambda^1_{\max}(\bar{u}_{1,r+\frac{1}{2}}), \lambda^1_{\max}(u^+_{1,r+\frac{1}{2}}), 0\},$$

where $\lambda^1_{\min}(u), \lambda^1_{\max}(u)$ are the smallest and largest eigenvalues of $\frac{\partial F_1(u)}{\partial u}$ and $\bar{u}_{1,r+\frac{1}{2}}$ is the corresponding Roe mean value, cf. [36]. Analogously we define

- $\hat{F}_2(u^-_{2,s+\frac{1}{2}}, u^+_{2,s+\frac{1}{2}}) := \frac{b^+_{2,s+\frac{1}{2}} F_2(u^-_{2,s+\frac{1}{2}}) - b^-_{2,s+\frac{1}{2}} F_2(u^+_{2,s+\frac{1}{2}})}{b^+_{2,s+\frac{1}{2}} - b^-_{2,s+\frac{1}{2}}} + \frac{b^+_{2,s+\frac{1}{2}} b^-_{2,s+\frac{1}{2}}}{b^+_{2,s+\frac{1}{2}} - b^-_{2,s+\frac{1}{2}}}(u^+_{2,s+\frac{1}{2}} - u^-_{2,s+\frac{1}{2}})$,

$$b^-_{2,s+\frac{1}{2}} := \{\lambda^2_{\min}(\bar{u}_{1,r+\frac{1}{2}}), \lambda^2_{\min}(u^-_{1,r+\frac{1}{2}}), 0\},$$
$$b^+_{2,s+\frac{1}{2}} := \{\lambda^2_{\max}(\bar{u}_{2,s+\frac{1}{2}}), \lambda^2_{\max}(u_{2,s+\frac{1}{2}}), 0\},$$

where $\lambda^2_{\min}(u), \lambda^2_{\max}(u)$ are the smallest and largest eigenvalues of $\frac{\partial F_2(u)}{\partial u}$.

As numerical solver we use the Runge–Kutta discontinuous stochastic Galerkin solver SG-FLEXI [12] with a SSP RK time-discretization of order four. We set the CFL-number from Assumption (4.14) to $C = 0.5$. In the following numerical examples we measure the error in mean and variance at final computational time $t = T$ in the $L^1(D)$- or $L^2(D)$-norm which we

approximate with a tensor product Gauß quadrature rule with 15 points (in one dimension) in every physical cell and 20 points (in one dimension) in every ME. We denote the number of spatial cells by $N_s$ and the number of MEs by $N_\Xi$. When we employ the ME ansatz from Section 4.1.2, we call our method ME-hDSG. The experimental order of convergence (eoc) is computed by

$$\text{eoc} = \frac{\log\left(\frac{\text{error}(r)}{\text{error}(r+1)}\right)}{\log\left(\frac{\text{dof}(r+1)}{\text{dof}(r)}\right)}, \tag{4.40}$$

where the degrees of freedom $\text{dof}(r)$ corresponds to the number of spatial cells $N_s$ or Multi-Elements $N_\Xi$ and $r$ represents the level of refinement.

### Convergence test for smooth solutions

As a first benchmark example we consider a spatial refinement of the physical domain $D = [0,1]^2_{\text{per}}$ for a SG polynomial degree of $K = 10$, DG polynomial degrees of one and three and one Multi-Element, i.e. $N_\Xi = 1$. We construct a smooth manufactured solution by introducing an additional source term in (4.38). We choose the following analytical function

$$u(t,x_1,x_2,y) = \begin{pmatrix} \rho(t,x_1,x_2,y) \\ m_1(t,x_1,x_2,y) \\ m_2(t,x_1,x_2,y) \\ E(t,x_1,x_2,y) \end{pmatrix} = \begin{pmatrix} 2+0.1(2\pi(x_1-yt)) \\ 2+0.1\cos(2\pi(x_1-yt)) \\ 0 \\ (2+0.1\cos(2\pi(x_1-yt))^2 \end{pmatrix}, \tag{4.41}$$

where $y = \xi(\omega)$, $\xi \sim \mathscr{U}(0.1,1)$ is an uncertain frequency. We compute the numerical approximation of (4.41) up to $T = 1$. Table 4.1 displays the $L^2$-error in mean and variance of density. The error is clearly dominated by the spatial error as the resolution in the random space is sufficiently high. Hence, the numerical error decreases with the rate of the DG method, which is $(q+1)/2$ in two spatial dimensions.

As second benchmark example we consider a stochastic refinement, where we increase the SG polynomial degree $K$. We consider the same analytical function (4.41) from the previous numerical experiment. In Table 4.2 we show the $L^2$-error in mean and variance for density for a $K$-refinement for one random element, i.e. $N_\Xi = 1$. Here, the physical mesh consists of $N_s = 20 \times 20 = 400$ cells and a DG polynomial degree of six. For the smooth solution (4.41) we observe that the error exhibits the expected spectral convergence when we increase the SG polynomial degree.

| hDSG, $q = 1$ | | | | |
|---|---|---|---|---|
| $N_s$ | $L^2(D)$-Mean | eoc | $L^2(D)$-Variance | eoc |
| 4 | 2.7033e-02 | - | 5.9515e-02 | - |
| 16 | 6.7904e-03 | 1.00 | 1.0958e-02 | 1.22 |
| 64 | 1.1669e-03 | 1.27 | 2.466e-03 | 1.08 |
| 256 | 1.6904e-04 | 1.39 | 3.8439e-04 | 1.34 |
| 1024 | 2.4801e-05 | 1.38 | 6.967e-05 | 1.23 |
| 4096 | 4.2716e-06 | 1.27 | 1.5161e-05 | 1.10 |
| hDSG, $q = 3$ | | | | |
| $N_s$ | $L^2(D)$-Mean | eoc | $L^2(D)$-Variance | eoc |
| 4 | 9.6072e-04 | - | 3.4636e-03 | - |
| 16 | 4.1815e-05 | 2.26 | 1.0036e-04 | 2.55 |
| 64 | 1.7619e-06 | 2.28 | 4.3666e-06 | 2.26 |
| 256 | 7.1412e-08 | 2.31 | 3.9645e-07 | 1.73 |
| 1024 | 2.9046e-09 | 2.31 | 1.7748e-08 | 2.24 |
| 4096 | 1.6708e-10 | 2.06 | 9.6501e-10 | 2.10 |

Table 4.1: $L^2(D)$-errors and experimental order of convergence (eoc) for the Euler equations (density) with $K = 10$, $N_\Xi = 1$, for DG polynomial degrees $q = 1, 3$. Example (4.41).

| hDSG | | |
|---|---|---|
| $K$ | $L_2$-Mean | $L_2$-Variance |
| 1 | 6.7406e-06 | 1.3562e-05 |
| 2 | 3.1228e-06 | 8.5896e-06 |
| 3 | 1.6532e-06 | 4.5845e-06 |
| 4 | 5.5879e-07 | 1.6006e-06 |
| 5 | 1.2829e-07 | 4.3977e-07 |
| 6 | 2.1531e-08 | 8.7639e-08 |
| 7 | 2.7923e-09 | 1.2051e-08 |
| 8 | 2.9392e-10 | 1.4562e-09 |
| 9 | 3.5925e-11 | 1.131e-10 |
| 10 | 2.4548e-11 | 4.7307e-11 |

Table 4.2: $L^2(D)$-errors for the Euler equations (density) with $N_\Xi = 1$, $N_s = 400$ for DG polynomial degree $q = 6$. Example (4.41).

**Uncertain Sod Shock test**

In this numerical test we study the behavior of the hyperbolicity limiter $\Lambda\Pi^{\theta}$ when it is applied to the one-dimensional uncertain Sod shock problem from [87, 91]. To this end we consider an uncertain position of the initial discontinuity, i.e., we consider the following set of initial conditions

$$\left.\begin{aligned} \rho(t=0,x,y) &= \begin{cases} 1, & x < 0.5+0.05y, \\ 0.125, & x \geq 0.5+0.05y, \end{cases} \\ m(t=0,x,y) &= 0, \\ E(t=0,x,y) &= \begin{cases} 2.5, & x < 0.5+0.05y, \\ 0.25, & x \geq 0.5+0.05y, \end{cases} \end{aligned}\right\} \tag{4.42}$$

where $\xi \sim \mathscr{U}(-1,1)$. The numerical solution is computed up to $T = 0.2$ and we define the spatial domain as $D = [0,1]$. At the boundary we prescribe exact boundary conditions. We choose the Lax-Friedrichs numerical flux and the TVBM minmod limiter from [24] as spatial limiter $\Lambda\Pi_h$. We divide the spatial domain into $N_s = 500$ cells, set the DG polynomial degree to three and consider the hDSG and ME-hDSG method. For the hDSG scheme we use a truncation order of $K = 10$ and for ME-hDSG we consider $N_{\Xi} = 10$ MEs and a linear approximation, i.e. $K = 1$. Both methods are compared to a Monte Carlo simulation obtained with an exact Riemann solver [4] with 200 000 samples.

In Figure 4.1 we compare mean and variance obtained with the hDSG and ME-hDSG methods against the reference solution given by the Monte Carlo sampling. The expected value in Figure 4.1(a) and Figure 4.1(b) indicates a good agreement between Monte Carlo, hDSG and ME-hDSG. However, for the hDSG method we can see in Figure 4.1(b) and Figure 4.1(d), especially around the shock at $x \approx 0.8$, that the hDSG solution exhibits ($K+1 = 11$) small shocks, which has also been observed for example in [31]. Because of the discontinuities in $y$, the plain hDSG approach suffers from Gibbs' oscillations and hence using a piecewise linear interpolation, as in the ME approach, yields a far better resolution of mean and variance compared to the plain hDSG approach.

Spurious oscillations for the hDSG method can also be observed in the $x-y$-diagram in Figure 4.2(a), especially in the vicinity of the shock-curve around $x \approx 0.8$. For the piecewise linear interpolation of the ME-hDSG method the oscillations have vanished, cf. Figure 4.2(b).

Furthermore, the influence of the $x-y$ discontinuities can be seen in Table 4.3, where we show the error in mean and variance between the hDSG-, ME-hDSG-approximation and the Monte

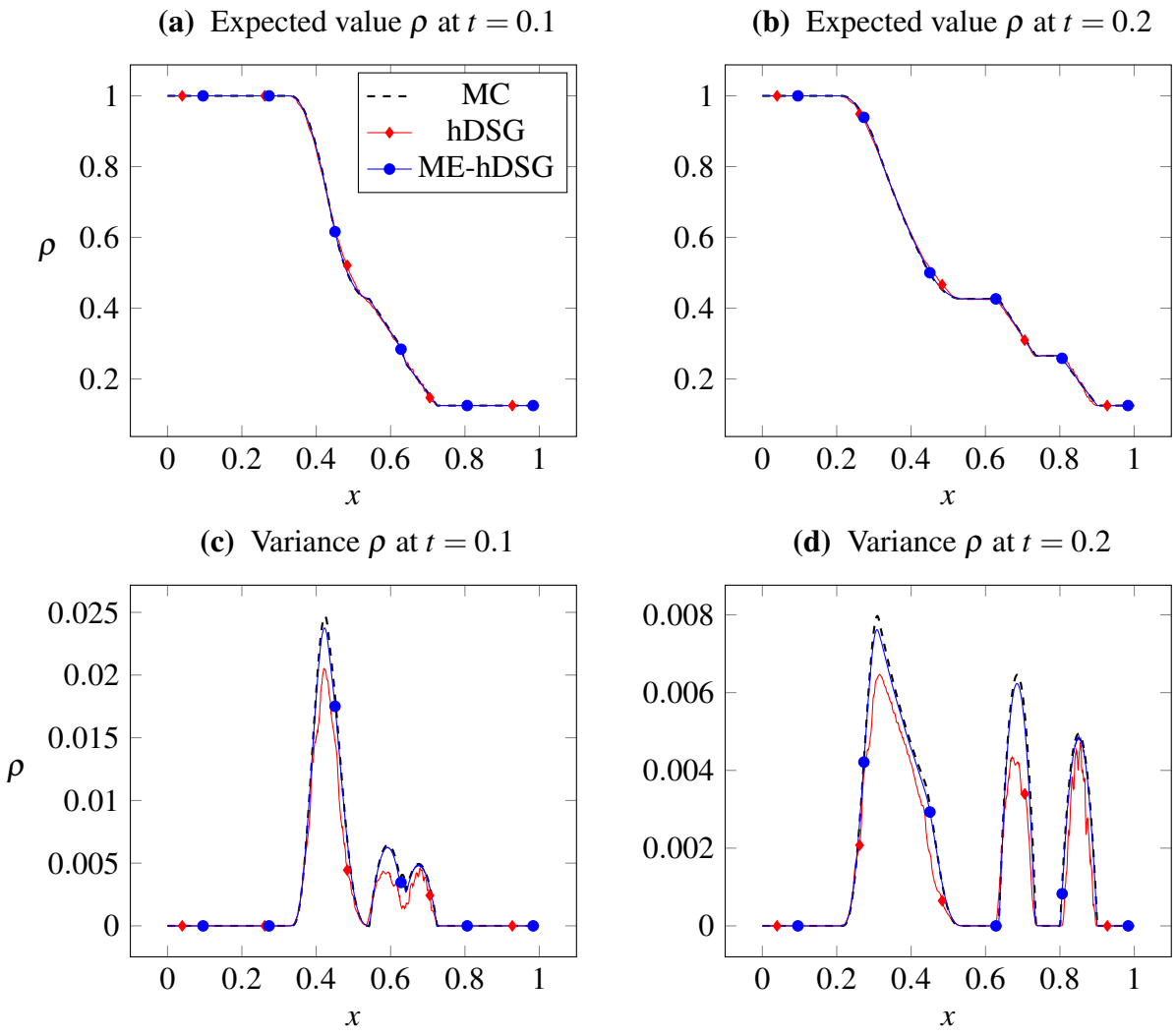



Figure 4.1: Density of Euler equations with initial state (4.42) at final time $T = 0.2$, $N_s = 500$ and DG polynomial degree $q = 3$. For hDSG, $K = 10$ and for ME-hDSG $K = 1$, $N_\Xi = 10$. Example (4.42).

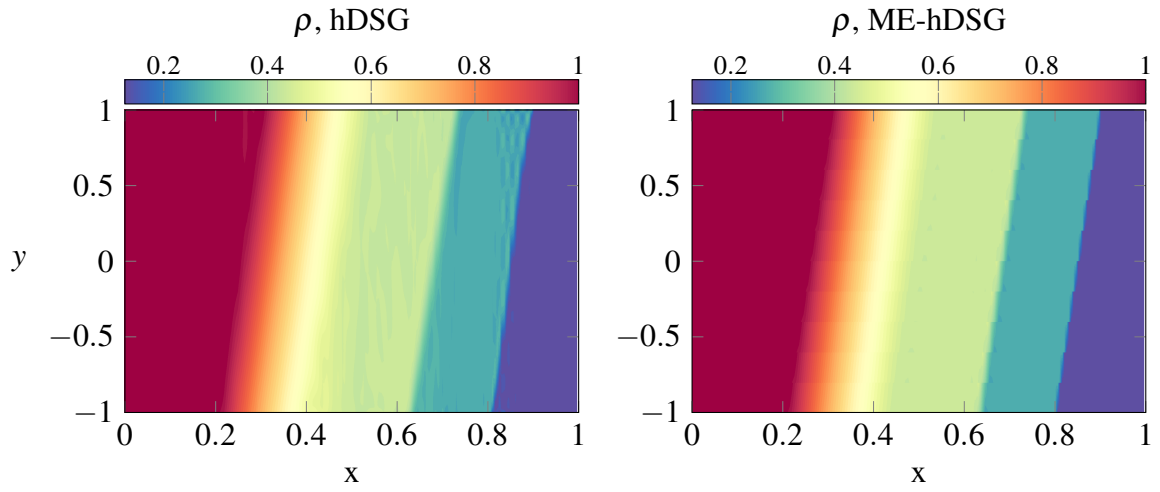



Figure 4.2: Space-stochastic surface plot for density of Euler equations with initial state (4.42), $N_s = 500$ and DG polynomial degree $q = 3$. For hDSG, $K = 10$ and for ME-hDSG, $K = 1$, $N_{\Xi} = 10$. Example (4.42).

Carlo reference solution. We deduce that the error for the hDSG-approach quickly starts to stagnate, whereas the error for the ME-hDSG approach is still decreasing. However, for all three methods the computed order of convergence in this example is smaller than one which is due to the discontinuities in $y$.

Finally, we display in Table 4.4 the percentage of limited cells for both methods compared to all space-time-stochastic cells. The percentage of limited cells for ME-hDSG is one order of magnitude lower than for the hDSG method, indicating that the numerical solution is more likely to leave the hyperbolicity set for the hDSG approach, than for the ME-hDSG method. This demonstrates the superiority of the ME approach when dealing with discontinuous data.

| hDSG | | | | |
|---|---|---|---|---|
| $K$ | $L^1(D)$-Mean | eoc | $L^1(D)$-Variance | eoc |
| 2 | 0.0075 | – | 7.3846e-04 | – |
| 4 | 0.0049 | 0.61 | 5.7026e-04 | 0.37 |
| 8 | 0.0038 | 0.37 | 4.4814e-04 | 0.35 |
| 16 | 0.0037 | 0.02 | 4.3632e-04 | 0.04 |
| ME-hDSG, $K = 0$ | | | | |
| $N_\Xi$ | $L^1(D)$-Mean | eoc | $L^1(D)$-Variance | eoc |
| 2 | 0.0086 | – | 8.8862e-04 | – |
| 4 | 0.0043 | 1.02 | 4.3359e-04 | 1.04 |
| 8 | 0.0021 | 0.99 | 2.2043e-04 | 0.98 |
| 16 | 0.0012 | 0.83 | 1.2997e-04 | 0.76 |
| ME-hDSG, $K = 1$ | | | | |
| $N_\Xi$ | $L^1(D)$-Mean | eoc | $L^1(D)$-Variance | eoc |
| 2 | 0.0054 | – | 5.5152e-04 | – |
| 4 | 0.0027 | 0.98 | 2.8091e-04 | 0.97 |
| 8 | 0.0016 | 0.75 | 1.6526e-04 | 0.77 |
| 16 | 0.0011 | 0.55 | 1.1586e-04 | 0.51 |

Table 4.3: $L^1(D)$-errors and experimental order of convergence (eoc) for the Euler equations (density) for $N_s = 500$ and DG polynomial degree $q = 3$. Example (4.42).

| $K/N_\Xi$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| hDSG [%] | 0.0146 | 0.1473 | 0.0721 | 0.0457 | 0.0376 | 0.0386 |
| ME-hDSG [%] | 0.0146 | 0.0028 | 0.0021 | 0.0034 | 0.0037 | 0.0022 |
| $K/N_\Xi$ | 7 | 8 | 9 | | | |
| hDSG [%] | 0.0335 | 0.0251 | 0.0199 | | | |
| ME-hDSG [%] | 0.0024 | 0.0021 | 0.0008 | | | |

Table 4.4: Percentage of limited cells over all time-steps for the Euler equations with $N_s = 500$ and DG polynomial degree $q = 3$. For ME-hDSG we use $K = 1$. Example (4.42).

**An uncertain Riemann problem**

In this numerical test we consider an uncertain Riemann problem in two spatial dimensions $D = [-0.5, 0.5]^2$. We define

$$Q_1 := [-0.5, 0] \times [-0.5, 0], \quad Q_2 := [0, 0.5] \times [-0.5, 0],$$
$$Q_3 := [-0.5, 0] \times [0, 0.5], \quad Q_4 := [0, 0.5] \times [0, 0.5],$$

and perturb the density in $Q_2$ and $Q_3$. Hence, the uncertain Riemann data in primitive variable read as follows

$$(\rho, u, v, p)(0, x_1, x_2, y) = \begin{cases} (0.138, 1.206, 1.206, 0.029), & (x_1, x_2) \in Q_1, \\ (0.5232 + 0.1y, 0, 1.206, 0.3), & (x_1, x_2) \in Q_2, \\ (0.5232 + 0.1y, 1.206, 0, 0.3), & (x_1, x_2) \in Q_3, \\ (1.5, 0, 0, 1.5), & (x_1, x_2) \in Q_4, \end{cases} \tag{4.43}$$

where $\xi \sim \mathscr{U}(-2, 2)$. This configuration ensures that all four constant states are separated by shock waves, cf. [92]. We set $T = 0.2$ and at all boundaries we use extrapolatory boundary conditions. As numerical flux we use the HLLE numerical flux and we choose the FV sub-cell limiter from [95] as spatial limiter $\Lambda\Pi_h$. To detect troubled cells we implement the modified JST indicator as described in [95].

For this example we use the ME-hDSG method with $N_\Xi = 4$, $K = 4$, DG polynomial degree $q = 4$ and $N_s = 200 \times 200$ cells. We compare the results obtained with the ME-hDSG method with a reference solution, which we obtained with the ME-SC method with $N_\Xi = 20$ MEs and a linear approximation, i.e. $K = 1$, using the Finite-Volume module of FLEXI [95] with a second order reconstruction, on a mesh with $N_s = 500 \times 500$ cells.

Figure 4.3 illustrates mean and standard deviation (std) of density at final time $T = 0.2$ obtained with both methods. In Figure 4.3(d) we also show a Schlieren plot of the mean of density obtained with ME-SC. We observe that in std the ME-SG approach is slightly more diffusive than ME-SC, but overall we see a very good agreement of both methods. To underline the strength of the ME approach we plot in Figure 4.4 mean and density obtained with a standard collocation approach using $21^2 = 441$ collocation points on one ME, i.e. $N_\Xi = 1$. While the mean flow coincides with the results of ME-hDSG and ME-SC, we observe that in regions of high uncertainty the std obtained with SC is considerably smaller than for ME-hDSG and ME-SC. This is presumably because of Gibbs oscillations due to the discontinuous initial condition.

Finally, in Figure 4.5 we plot the values of the limiter variable for each ME. The limiter is only

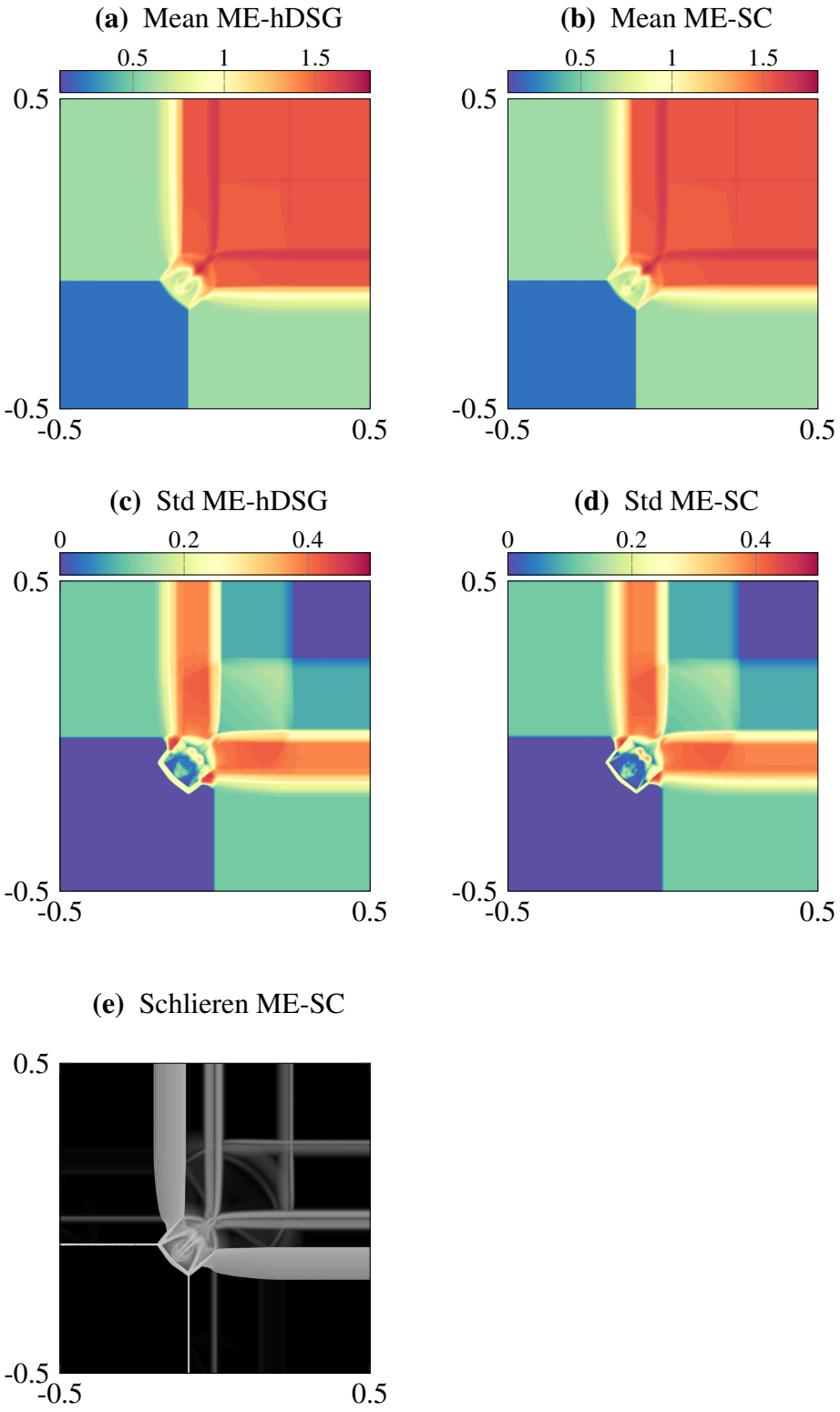



Figure 4.3: Mean, standard deviation and Schlieren plot of density at final time $T = 0.2$ obtained with ME-hDSG and ME-SC. Example (4.43).

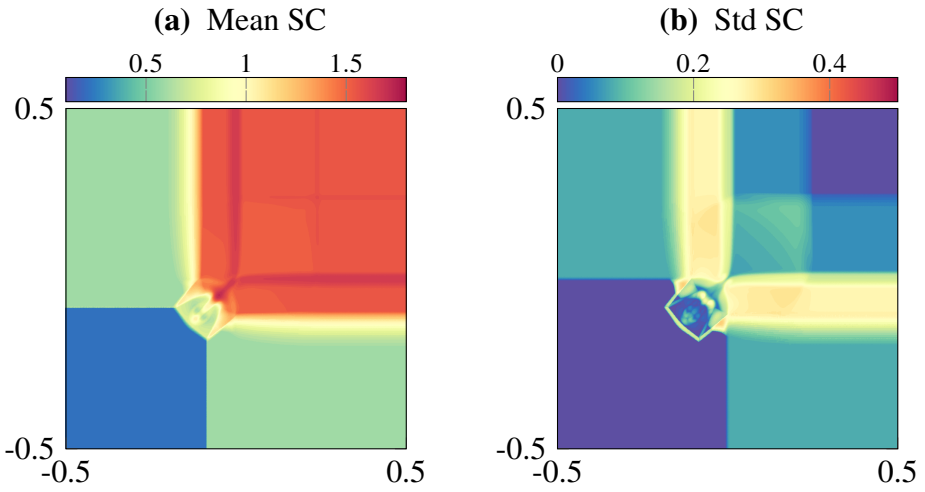



Figure 4.4: Mean and standard deviation obtained with standard SC. Example (4.43).

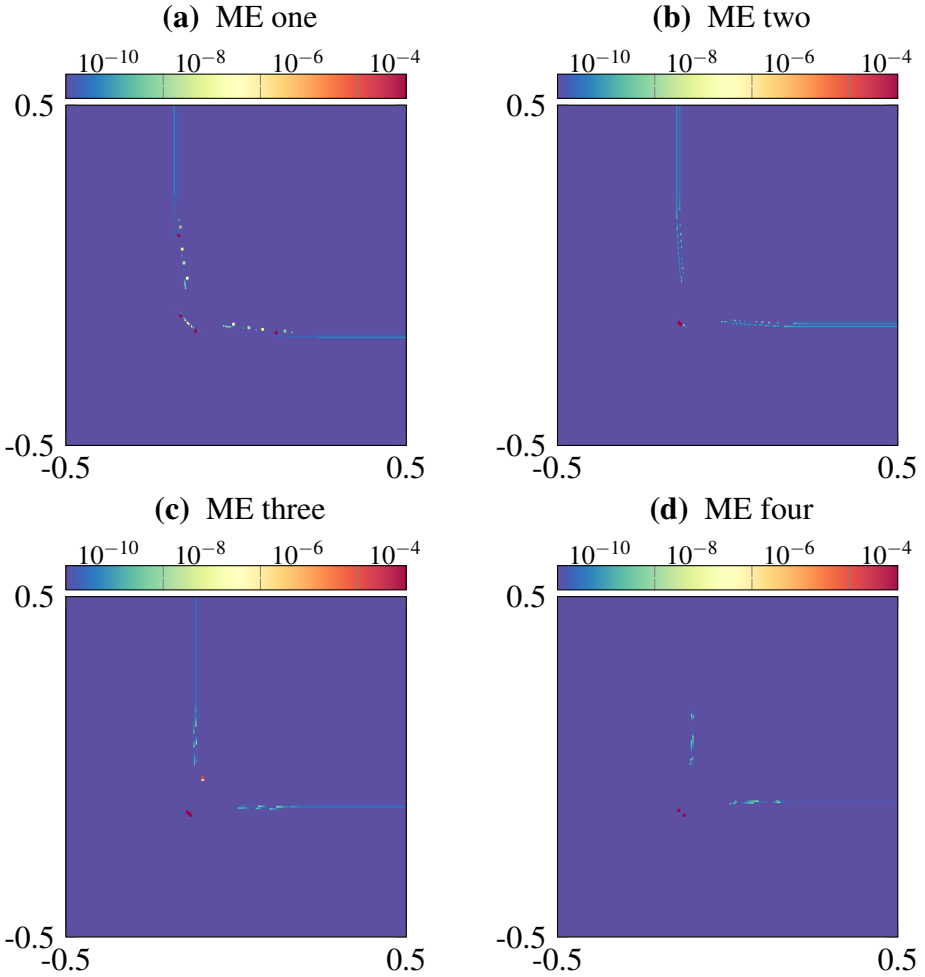



Figure 4.5: Plot of limiter variable $\theta$ in spatial domain for each ME at final time $T = 0.2$. Example (4.43).

active around the position of the shock, moreover it is only active between $Q_2$ and $Q_4$, resp. $Q_3$ and $Q_4$.

### Double Mach Reflection with uncertain angle

As a final numerical test for the hyperbolicity-preserving limiter, we consider the Double Mach Reflection test case suggested by Woodward and Colella [105]. It consists of a Mach 10 shock wave that hits a ramp which is inclined by 30 degrees. The Double Mach reflection poses a very challenging problem for the hDSG method because the solution is very likely to leave the hyperbolicity set due to the high jump in pressure. We choose the angle of the ramp uncertain, i.e., we let $\xi \sim \mathscr{U}(28^\circ, 32^\circ)$ and consider the following Riemann data in primitive variables

$$\begin{cases} \rho(t=0,x_1,x_2,y) & = \begin{cases} 8, & x < \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \\ 0.125, & x \geq \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \end{cases} \\ v_1(t=0,x_1,x_2,y) & = \begin{cases} 8.25\cos\left(\frac{y\pi}{180^\circ}\right) & x < \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \\ 0, & x \geq \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \end{cases} \\ v_2(t=0,x_1,x_2,y) & = \begin{cases} -8.25\cos\left(\frac{y\pi}{180^\circ}\right), & x < \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \\ 0, & x \geq \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \end{cases} \\ p(t=0,x_1,x_2,y) & = \begin{cases} 116.5, & x < \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \\ 1, & x \geq \bar{x} + \tan\left(\frac{y\pi}{180^\circ}\right) x_2, \end{cases} \end{cases} \tag{4.44}$$

where $\bar{x} = \frac{1}{6}$ is the start of the ramp. The computational domain is $D = [0,4] \times [0,1]$ and we set $T = 0.2$. At the bottom of the domain we employ reflective boundary conditions whereas we prescribe outflow boundary conditions at the right. At the remaining boundaries we apply Dirichlet boundary conditions, which correspond to the physical values. We use the HLLE numerical flux and we choose the FV sub-cell limiter as spatial limiter $\Lambda\Pi_h$. To detect troubled cells we use the modified JST indicator like for the uncertain Riemann problem.

For this numerical example we apply the ME-hDSG method with $N_\Xi = 8$ MEs, SG polynomial degree $K = 4$. The physical mesh consists of $N_s = 240 \times 40$ cells and we use a DG polynomial degree of $q = 4$. To compute a reference solution we use the ME-SC method with $N_\Xi = 20$ MEs and a linear approximation, i.e. $K = 1$, using the Finite-Volume module of FLEXI [95] on a mesh with $N_s = 1000 \times 450$ cells. In Figure 4.6 we plot mean and std of density at final time $T = 0.2$. We can see that the shock front is smeared out because of the variable angle. Thanks to the high-order resolution in physical and random space, small-scale flow features in mean and std of density are clearly visible. A high std can be identified around the position of

the shock wave. Also around $\bar{x} = \frac{1}{6}$, which corresponds to the start position of the ramp, we observe a high std. In summary we see a very good agreement between mean and std obtained with ME-hDSG and ME-SC.

Figure 4.7 shows the values of the limiter variable $\theta$ in ME one and eight at initial time $t = 0$. We see that the limiter is only active around the shock front. Since the values of $\theta$ are close to one, the solution is strongly limited towards the cell mean in this area. Furthermore, the limited cells vary with the uncertain angle. The plain SG and even the ME-SG approach without hyperbolicity-preserving limiter crash immediately after initialization of the initial condition.

In summary it can be said, that our proposed hyperbolicity-preserving scheme is a reliable and robust method to compute complex flow problems with a high resolution in space, time and stochasticity.

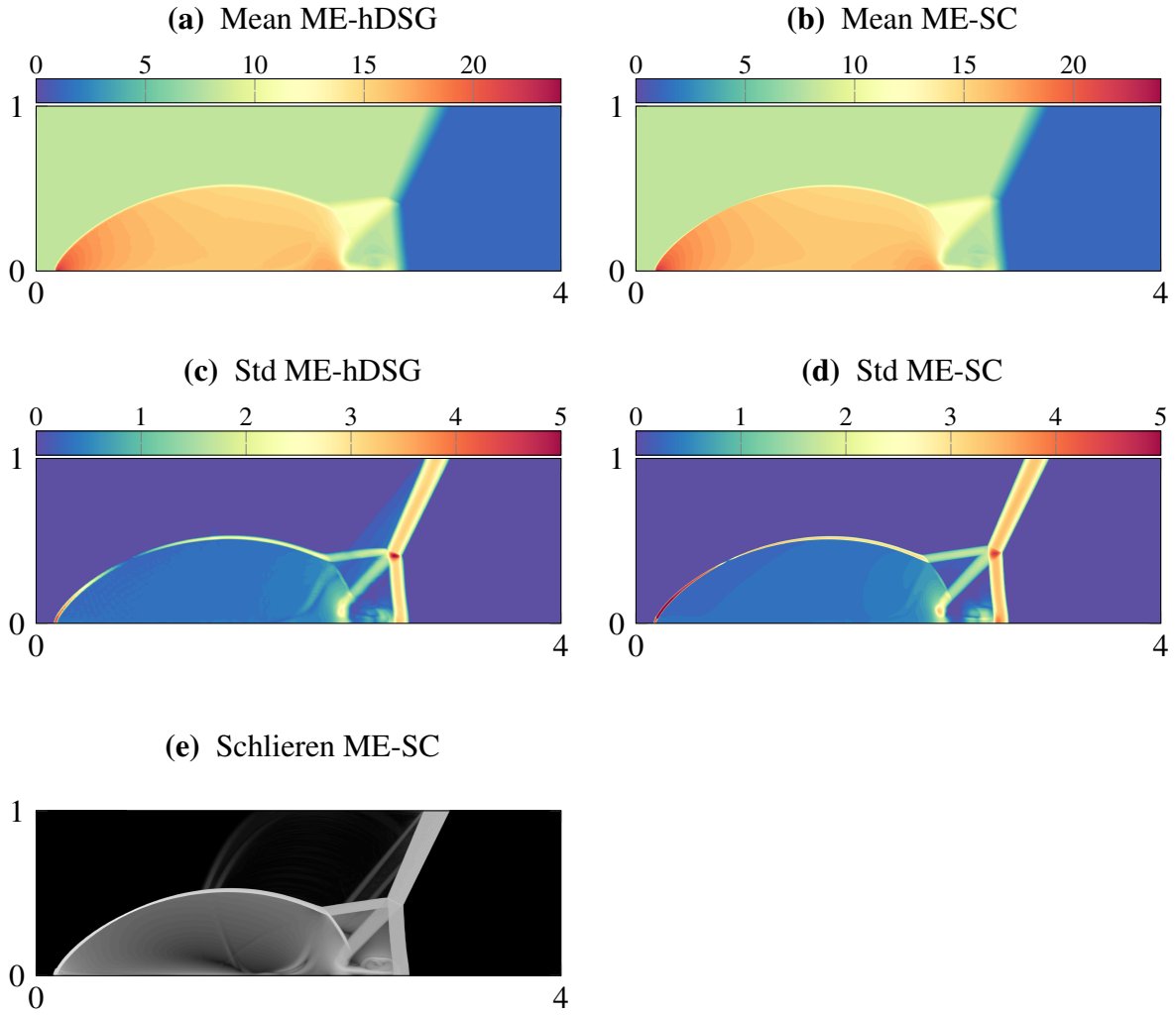



Figure 4.6: Mean, standard deviation and Schlieren plot of density at final time $T = 0.2$ obtained with ME-hDSG and ME-SC. Example (4.44).

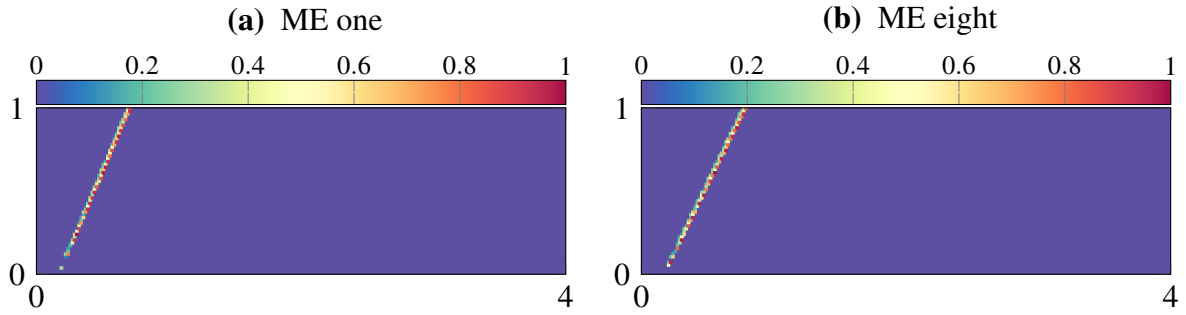



Figure 4.7: Plot of limiter variable $\theta$ in spatial domain for ME one and eight at initial time $t = 0$. Example (4.44).

## 4.3 Comparison between NISP and SG

In this section we compare the efficiency of SG with the efficiency of NISP by means of a smooth benchmark solution. To compare both methods we measure the total elapsed time (including post-processing for NISP) on a workstation equipped with an AMD Ryzen Thread-Ripper 2950x processor with sixteen kernels and 128 GB RAM. As deterministic solver for NISP we use the standard version of FLEXI [63] and for the SG method we use the modified FLEXI version SG-FLEXI from [12]. Both methods are parallelized using Open MPI and we measure the total elapsed time on sixteen kernels which we denote by cpu time. When we use NISP in combination with the ME method from Section 4.1.2, we call this method ME-NISP. For the collocation points in a multi-dimensional random space $\Xi$ we use tensor-products of one-dimensional Gauß–Legendre quadrature points. The number of collocation points in one dimension is always $K+1$.

As a benchmark solution we consider

$$u(t,x_1,x_2,y_1,y_2,y_3) = \begin{pmatrix} \rho(t,x_1,x_2,y_1,y_2,y_3) \\ m_1(t,x_1,x_2,y_1,y_2,y_3) \\ m_2(t,x_1,x_2,y_1,y_2,y_3) \\ E(t,x,y_1,y_2,y_3) \end{pmatrix} = \begin{pmatrix} y_3 + y_2\cos(2\pi(x_1 - y_1 t)) \\ y_3 + y_2\cos(2\pi(x_1 - y_1 t)) \\ 0 \\ (y_3 + y_2\cos(2\pi(x_1 - y_1 t))^2 \end{pmatrix}, \quad (4.45)$$

with three uniformly distributed random variables $\xi_1 \sim \mathscr{U}(0.1,1)$, $\xi_2 \sim \mathscr{U}(0.1,0.3)$ and $\xi_3 \sim \mathscr{U}(1.8,2.5)$. The exact solution (4.45) is computed up to $T = 1$. To test the efficiency of both methods for different random dimensions we successively increase the number of random variables starting from $\xi_1$ while fixing $\xi_2 = 0.1$, $\xi_3 = 2$ and so on. The number of spatial cells is always $N_s = 400$ and the DG polynomial degree is $q = 6$. As numerical flux we choose

Lax-Friedrichs and we compute the $L^2(D)$-error in mean of density at final computational time $T = 1$.

**Comparison: $K$-refinement**

In this example we consider a global $K$-refinement for one fixed ME, i.e. $N_\Xi = 1$. In Figures 4.8, 4.9 and Tables 4.5, 4.6, 4.7 we compare error vs. polynomial degree $K$ and error vs. cpu time for SG and NISP in one, two and three random dimensions. First we observe that both methods exhibit spectral convergence when we increase the number of polynomials. In almost all cases the SG method yields a smaller absolute error for the same polynomial degree, however, only in the one-dimensional case the SG scheme proves to be more efficient (in terms of error vs. cpu time) than the NISP method. Because the computation of SG is very expensive in three dimensions, we compute the error only up to SG polynomial degree of four.

| SG, $N = 1$ | | | | NISP, $N = 1$ | | |
|---|---|---|---|---|---|---|
| $K$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 6.7406e-06 | 1.3562e-05 | 6.0 | 0.25193 | 0.039658 | 2.0 |
| 2 | 3.1228e-06 | 8.5896e-06 | 6.0 | 0.048331 | 0.024901 | 6.0 |
| 3 | 1.6532e-06 | 4.5845e-06 | 8.0 | 0.0035809 | 0.010305 | 6.0 |
| 4 | 5.5879e-07 | 1.6006e-06 | 10.0 | 0.0001363 | 0.0018718 | 11.0 |
| 5 | 1.2829e-07 | 4.3977e-07 | 11.0 | 3.1597e-06 | 0.00019378 | 11.0 |
| 6 | 2.1531e-08 | 8.7639e-08 | 13.0 | 4.9265e-08 | 1.3059e-05 | 11.0 |
| 7 | 2.7923e-09 | 1.2051e-08 | 15.0 | 6.2033e-10 | 6.2083e-07 | 16.0 |
| 8 | 2.9392e-10 | 1.4562e-09 | 18.0 | 1.0251e-10 | 2.1989e-08 | 17.0 |
| 9 | 3.5925e-11 | 1.131e-10 | 21.0 | 1.0514e-10 | 5.9498e-10 | 17.0 |
| 10 | 2.4548e-11 | 4.7307e-11 | 25.0 | 1.0512e-10 | 3.1268e-11 | 18.0 |

Table 4.5: $L^2(D)$-errors and cpu time for the Euler equations (density) for SG and NISP in one random dimension. $K$-refinement.

**Comparison: ME-refinement**

In this numerical experiment we compare the ME-SG and ME-NISP method for an increasing number of MEs. For both methods we consider a linear interpolation, i.e. $K = 1$. Figures 4.10, 4.11 and Tables 4.8, 4.9, 4.10 show error vs. number of MEs $N_\Xi$ and error vs. cpu time for both

| SG, $N = 2$ | | | | NISP, $N = 2$ | | |
|---|---|---|---|---|---|---|
| $K$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 1.2374e-04 | 2.5015e-04 | 80.0 | 0.25193 | 0.043 | 5.0 |
| 2 | 6.041e-05 | 0.00018524 | 94.0 | 0.048331 | 0.024827 | 16.0 |
| 3 | 3.5254e-05 | 0.00012008 | 122.0 | 0.0035809 | 0.009956 | 20.0 |
| 4 | 1.4209e-05 | 5.7537e-05 | 173.0 | 0.0001363 | 0.0017333 | 31.0 |
| 5 | 5.2814e-06 | 2.1894e-05 | 221.0 | 3.1597e-06 | 0.00017395 | 46.0 |
| 6 | 1.5886e-06 | 6.3609e-06 | 336.0 | 4.9266e-08 | 1.1552e-05 | 63.0 |
| 7 | 3.5052e-07 | 1.4504e-06 | 467.0 | 6.1997e-10 | 5.5604e-07 | 85.0 |
| 8 | 5.8279e-08 | 2.4925e-07 | 700.0 | 1.0471e-10 | 2.0564e-08 | 109.0 |
| 9 | – | – | – | 1.0723e-10 | 5.9281e-10 | 138.0 |
| 10 | – | – | – | 1.0725e-10 | 2.7011e-11 | 170.0 |

Table 4.6: $L^2(D)$-errors and cpu time for the Euler equations (density) for SG and NISP in two random dimensions. $K$-refinement.

| SG, $N = 3$ | | | | NISP, $N = 3$ | | |
|---|---|---|---|---|---|---|
| $K$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 1.1315e-03 | 3.51281e-03 | 1844.0 | 0.25193 | 0.12457 | 81.0 |
| 2 | 2.192e-04 | 8.2764e-04 | 2722.0 | 0.048331 | 0.024827 | 97.0 |
| 3 | 5.3598e-05 | 2.4934e-04 | 5406.0 | 0.0035809 | 0.009956 | 122.0 |
| 4 | 1.2283e-05 | 5.4753e-05 | 14583.0 | 0.0001363 | 0.0017333 | 181.0 |
| 5 | – | – | – | 3.1597e-06 | 0.00017395 | 283.0 |
| 6 | – | – | – | 4.9273e-08 | 1.1552e-05 | 450.0 |
| 7 | – | – | – | 6.1245e-10 | 5.5605e-07 | 701.0 |
| 8 | – | – | – | 9.3374e-11 | 2.0562e-08 | 1080.0 |
| 9 | – | – | – | 9.5918e-11 | 5.9519e-10 | 1637.0 |
| 10 | – | – | – | 9.5906e-11 | 2.1529e-11 | 2448.0 |

Table 4.7: $L^2(D)$-errors and cpu time for the Euler equations (density) for SG and NISP in three random dimensions. $K$-refinement.

methods. Similar to the $K$-refinement in Section 4.3, the SG methods yields a smaller error for the same number of MEs. We observe that for a one-dimensional random space $\Xi$, ME-SG is clearly more efficient (in terms of cpu time) than ME-NISP. In the two-dimensional case it is only more efficient for one ME. However, for three random dimensions, ME-SG stands no chance against the ME-NISP method in terms of efficiency. This is exactly the behavior that

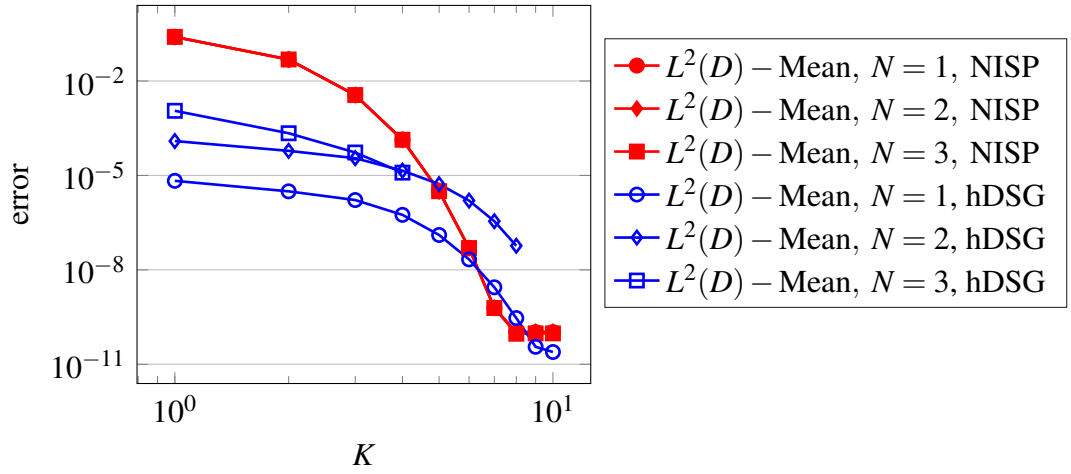



Figure 4.8: $L^2(D)$-errors vs. $K$ for the Euler equations (density) for SG and NISP. $K$-refinement.

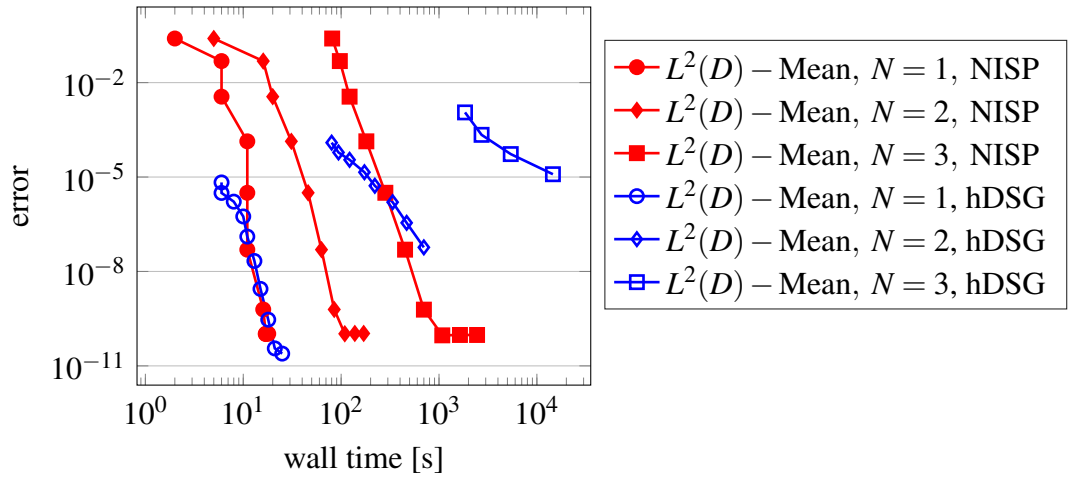



Figure 4.9: $L^2(D)$-errors vs. cpu time for the Euler equations (density) for SG and NISP. $K$-refinement.

we expect from SG and ME-SG, because computing the orthogonal projection of the modified fluxes in (4.5) becomes more and more expensive when we increase the number of random dimensions. Hence, for a one and maybe a two-dimensional random space, the SG method yields an efficiency gain compared to NISP but after that point it is not competitive against NISP. From our numerical experiments we deduce that SG is not suitable for UQ problems with a high-dimensional random space and should only be employed if one is interested in highly accurate numerical results for problems with one and maybe two uncertain parameters.

| ME-SG, $N = 1$ | | | | ME-NISP, $N = 1$ | | |
|---|---|---|---|---|---|---|
| $N_\Xi$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 6.7125e-06 | 1.2071e-05 | 4.0 | 0.048331 | 0.024901 | 7.0 |
| 2 | 1.3863e-06 | 4.3691e-06 | 7.0 | 0.00060436 | 0.0046097 | 11.0 |
| 3 | 3.5026e-07 | 1.162e-06 | 11.0 | 0.00010158 | 0.00023847 | 11.0 |
| 4 | 1.2083e-07 | 4.3255e-07 | 14.0 | 3.0497e-05 | 5.8152e-05 | 17.0 |
| 5 | 5.1526e-08 | 1.7694e-07 | 20.0 | 1.2199e-05 | 2.1431e-05 | 16.0 |
| 6 | 2.5395e-08 | 8.9625e-08 | 27.0 | 5.8089e-06 | 9.7896e-06 | 21.0 |
| 7 | 1.3884e-08 | 4.8357e-08 | 28.0 | 3.1117e-06 | 5.1192e-06 | 27.0 |
| 8 | 8.2032e-09 | 2.9729e-08 | 29.0 | 1.8151e-06 | 2.9407e-06 | 28.0 |
| 9 | 5.1462e-09 | 1.8097e-08 | 47.0 | 1.1293e-06 | 1.8109e-06 | 31.0 |
| 10 | 3.3861e-09 | 1.1955e-08 | 51.0 | 7.3914e-07 | 1.1766e-06 | 32.0 |

Table 4.8: $L^2(D)$-errors and cpu time for the Euler equations (density) for ME-SG and ME-NISP in one random dimension. ME-refinement.

| ME-SG, $N = 2$ | | | | ME-NISP, $N = 2$ | | |
|---|---|---|---|---|---|---|
| $N_\Xi$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 1.2374e-04 | 2.5016e-04 | 33.0 | 0.068057 | 0.027066 | 15.0 |
| 4 | 5.7001e-05 | 2.9659e-04 | 134.0 | 0.00060436 | 0.0047143 | 30.0 |
| 9 | 1.6892e-05 | 1.1411e-04 | 432.0 | 0.00010158 | 0.00026195 | 60.0 |
| 16 | 6.1867e-06 | 4.6866e-05 | 526.0 | 3.0497e-05 | 6.7483e-05 | 102.0 |
| 25 | – | – | – | 1.2199e-05 | 2.5668e-05 | 157.0 |
| 36 | – | – | – | 5.8089e-06 | 1.1948e-05 | 218.0 |
| 49 | – | – | – | 3.1117e-06 | 6.3225e-06 | 293.0 |
| 64 | – | – | – | 1.8151e-06 | 3.6608e-06 | 384.0 |
| 81 | – | – | – | 1.1293e-06 | 2.2669e-06 | 491.0 |
| 100 | – | – | – | 7.3915e-07 | 1.4788e-06 | 612.0 |

Table 4.9: $L^2(D)$-errors and cpu time for the Euler equations (density) for ME-SG and ME-NISP in two random dimensions. ME-refinement.

| ME-SG, $N = 3$ | | | | ME-NISP, $N = 3$ | | |
|---|---|---|---|---|---|---|
| $N_\Xi$ | $L^2$-Mean | $L^2$-Variance | cpu time [s] | $L^2$-Mean | $L^2$-Variance | cpu time [s] |
| 1 | 1.1316e-03 | 3.5136e-03 | 240.0 | 0.048331 | 0.024827 | 103.0 |
| 8 | 1.5013e-04 | 7.9673e-04 | 1799.0 | 0.00060436 | 0.0047143 | 182.0 |
| 27 | 3.5281e-05 | 2.3186e-04 | 10432.0 | 0.00010158 | 0.00026195 | 406.0 |
| 64 | – | – | – | 3.0497e-05 | 6.7483e-05 | 850.0 |
| 125 | – | – | – | 1.2199e-05 | 2.5668e-05 | 1583.0 |
| 216 | – | – | – | 5.8089e-06 | 1.1948e-05 | 2691.0 |
| 343 | – | – | – | 3.1117e-06 | 6.3225e-06 | 4321.0 |
| 512 | – | – | – | 1.8151e-06 | 3.6609e-06 | 6519.0 |
| 729 | – | – | – | 1.1293e-06 | 2.2669e-06 | 9335.0 |

Table 4.10: $L^2(D)$-errors and cpu time for the Euler equations (density) for ME-SG and ME-NISP in three random dimensions. ME-refinement.

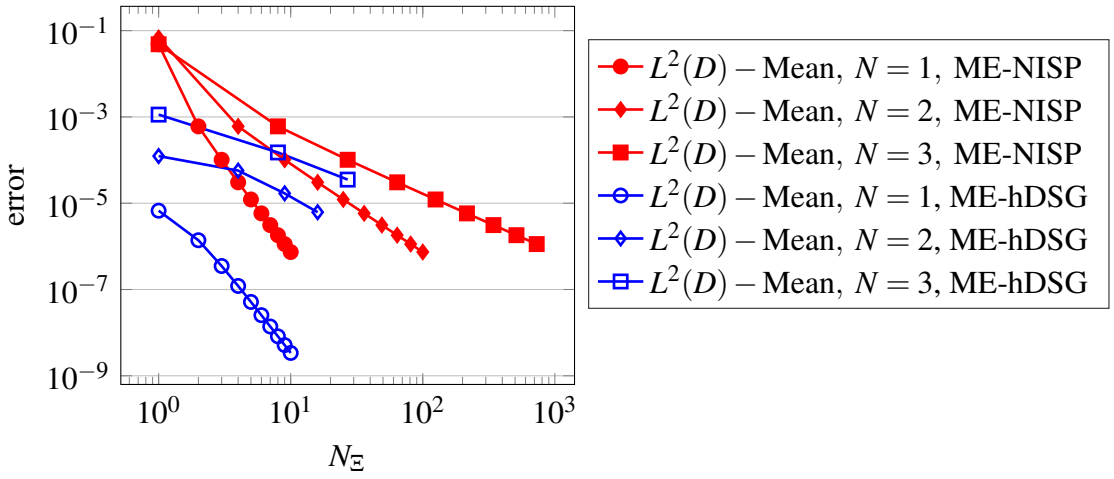



Figure 4.10: $L^2(D)$-errors vs. $N_\Xi$ for the Euler equations (density) for ME-SG and ME-NISP. ME-refinement.

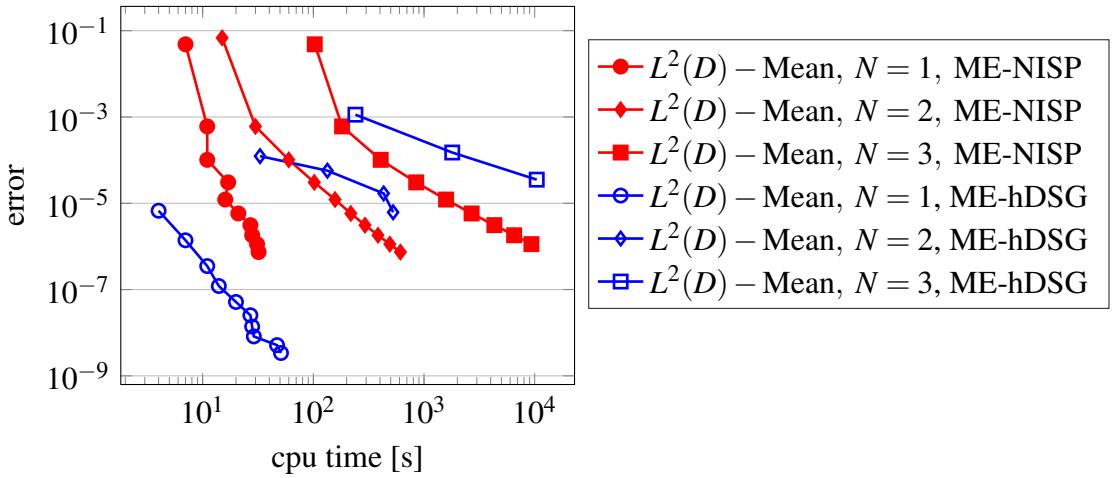



Figure 4.11: $L^2(D)$-errors and cpu time for the Euler equations (density) for ME-SG and ME-NISP method. ME-refinement.

# 5 A posteriori error analysis framework for non-statistical UQ methods

One of the most important tools to accelerate computations and increase efficiency of numerical schemes is to locally adapt the grid in physical and random space. The main ingredient for adaptive mesh refinements are indicators which mark spatial and stochastic cells for refinement or coarsening. From a heuristic point of view, indicators should capture important information of the solution like position of discontinuities or areas of steep gradients to ensure that the mesh is only refined in areas where mesh refinement is really necessary. On the other hand, from a numerical analysis point of view, the indicator should be linked to the true numerical error, in the sense that the indicator provides a computable bound for the error between the exact solution and its numerical approximation. These type of indicators are based on a posteriori error estimates and in this case it is possible to construct adaptive numerical schemes which guarantee that the true error is below a prescribed threshold (see for example [30, 71] in the case of nonlinear conservation laws). Compared to random elliptic and random parabolic conservation laws [13, 35, 52, 56, 90] the a posteriori error analysis for random hyperbolic conservation laws is still in its infancy. One of the main reasons for this deficiency is the general absence of a unified a posteriori error framework for the numerical approximation of deterministic hyperbolic conservation laws.

Despite the lack of an appropriate a posteriori error analysis, adaptive mesh refinements in physical and stochastic space based on heuristic indicators have already been successfully applied to random conservation laws. In [69, 99, 101] space-stochastic adaptive mesh refinements for the SG method are presented. Adaptive algorithms for the Stochastic Collocation method and for the Simplex Stochastic Collocation method have been considered in [18, 57, 104]. Except for [18], where the authors consider dual-based a posteriori error estimates for linear output functionals, all of the above mentioned papers use heuristic criteria, for example the relative change in the highest mode of the gPC approximation or the relative change in variance, to mark cells in the random space for refinement. In a similar manner, spatial refinement is based on heuristics indicators. In this thesis we make a first step towards an a posteriori error analysis

framework for random conservation laws and derive a novel residual-based a posteriori error estimate for space-time-stochastic discretizations which use the RKDG method in space and time, and polynomial-based UQ methods in random space.

The a posteriori error analysis framework is based on the relative entropy framework for random conservation laws, which we have presented in Section 2.4. To summarize the general idea, we view the numerical solution obtained with different UQ methods (or, more precisely, a reconstruction thereof) as the exact solution of a perturbed version of the original problem. The reconstructions which we use depend on the specific UQ method and will be explained in detail in the corresponding sections. The perturbation is given by a computable residual and by using the relative entropy method we can bound the difference between the exact and the numerical solution in terms of the residual. We prove that the residual admits a decomposition into a spatial and a stochastic part, which enables us to control the errors arising from spatio-temporal and stochastic discretization. Moreover, exploiting the residuals' structure we propose for the SC method a novel residual-based space-stochastic adaptive numerical scheme. Results from the following sections have led to the publications [47, 48, 49].

## 5.1 A posteriori error analysis based on the SG method

In this section we derive an a posteriori error estimator for a numerical approximation of the random entropy admissible solution of (2.1) using the SG method from Section 4.1.1. The presented results are published in [49].

For the following a posteriori error analysis we restrict ourselves to the one-dimensional torus $D = [0,1]_{per}$. More specifically, we consider the random conservation law (2.1) with uncertain initial data but deterministic flux function $F : \mathscr{U} \to \mathbb{R}^m$. For simplicity we consider a one-dimensional random space $\Xi \subset \mathbb{R}$. Our problem of interest is the following random initial-value problem.

$$\begin{cases} \partial_t u(t,x,y) + \partial_x F(u(t,x,y)) = 0, & (t,x,y) \in (0,T) \times D \times \Xi, \\ u(0,x,y) = u^0(x,y), & (x,y) \in D \times \Xi. \end{cases} \tag{5.1}$$

### 5.1.1 Discretization, reconstruction and residuals

Following the description in Section 4.1.1 we approximate the solution of (5.1) with the truncated generalized Fourier series $u(t,x,y) \approx \sum_{k=0}^{K} u_k(t,x)\Psi_k(y)$. According to Section 4.1.1, the

(SG)-system of (5.1) reads as follows:

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{F} = 0, & (t,x) \in (0,T) \times D \\ \mathbf{u}^0 = \left( \left\langle u^0, \Psi_k \right\rangle \right)_{k=0}^{K}, & x \in D, \end{cases} \tag{SG}$$

where we indicate the SG discretization with bold font. Here, $\mathbf{u} \in \mathbb{R}^{m\cdot(K+1)}$ and $\mathbf{F} : \boldsymbol{\mathcal{U}} \to \mathbb{R}^{m\cdot(K+1)}$, where $\boldsymbol{\mathcal{U}} \subset \mathbb{R}^{m\cdot(K+1)}$. We discretize the deterministic conservation law (SG) in space and time using the RKDG method from Section 3.1.

To this end we let $0 = x_0 < x_1 \ldots < x_{N_s} = 1$ be a quasi-uniform triangulation of $[0,1]$ and $0 = t_0 < t_1 < \ldots < t_{N_t} = T$ be a temporal decomposition of $[0,T]$ and identify $x_0 = x_{N_s}$ to account for the periodic boundary conditions. Performing the DG spatial discretization and introducing the numerical flux $\hat{\mathbf{F}} : \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{U}} \to \mathbb{R}^{m\cdot(K+1)}$, we end up with the following semi-discrete initial value problem.

$$\begin{cases} \sum_{l=0}^{N_s-1} \int_{x_l}^{x_{l+1}} \partial_t \mathbf{u}_h \cdot \mathbf{v_h} \, \mathrm{d}x = & \sum_{l=0}^{N_s-1} \int_{x_l}^{x_{l+1}} \mathbf{F}(\mathbf{u}_h) \cdot \partial_x \mathbf{v_h} \, \mathrm{d}x \\ & - \sum_{l=0}^{N_s-1} \hat{\mathbf{F}}(\mathbf{u}_h(x_l^-), \mathbf{u}_h(x_l^+)) \cdot [\![ \mathbf{v_h} ]\!]_l, \\ \mathbf{u}_h(t=0) = \mathscr{L}_{\mathscr{V}_h^q} \mathbf{u}^0, \end{cases} \tag{DG-SG}$$

for all $\mathbf{v_h} \in \mathscr{V}_h^q$. To account for the periodic boundary condition we set $\mathbf{u}_h(x_0^-) = \mathbf{u}_h(x_{N_s}^-)$, $\mathbf{u}_h(x_{N_s}^+) = \mathbf{u}_h(x_0^+)$. The initial-value problem (DG-SG) is solved in time using a SSP RK method, cf. Algorithm 3.1.

Using the sequence of approximate solutions of (DG-SG) at points $\{t_n\}_n^{N_t}$ in time, denoted by $\{\mathbf{u}_h^0, \ldots, \mathbf{u}_h^{N_t}\}$, we want to reconstruct the numerical approximation to a Lipschitz continuous function in space in time. With this function at hand we are able to apply Theorem 2.15 to derive an a posteriori error estimate for the difference between the entropy admissible solution and its numerical approximation, resp. the reconstruction thereof. We structure the reconstruction process into two parts.

**1. Temporal reconstruction $\hat{\mathbf{u}}^t$:**

We first compute the temporal reconstruction as proposed in [28]. For the reconstruction in time we define the spaces of piecewise polynomials in time of degree $r \in \mathbb{N}_0$ by

$$V_r^t((0,T); \mathscr{V}_h^q) := \{ \mathbf{u} : [0,T] \to \mathscr{V}_h^q \mid \mathbf{u}\big|_{(t_n, t_{n+1})} \in \mathbb{P}_r((t_n, t_{n+1}), \mathscr{V}_h^q) \}.$$

Using Hermite interpolation on each time interval $[t_n, t_{n+1}]$, we construct the temporal reconstruction $\hat{\mathbf{u}}^t \in V_r^t((0,T); \mathscr{V}_h^q)$. We note that the time derivative $\partial_t \mathbf{u}_h^n$ can be approximated using the right-hand side of (DG-SG).

**2. Space-time reconstruction $\hat{\mathbf{u}}^{st}$:**

With the temporal reconstruction $\hat{\mathbf{u}}^t$ at hand, we define the space-time reconstruction $\hat{\mathbf{u}}^{st}$ of the DG-solutions of (SG). To prove an optimal decay of the error estimator, the authors of [28] consider numerical fluxes $\hat{\mathbf{F}}$ which admit a special representation. In particular, there needs to exist a locally Lipschitz function $w : \mathcal{U} \times \mathcal{U} \to \mathbb{R}^{m\cdot(K+1)}$, with the additional property $w(\mathbf{u},\mathbf{u}) = \mathbf{u}$, such that $\hat{\mathbf{F}}$ can either be expressed as

$$\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v})), \quad \forall\, \mathbf{u},\mathbf{v} \in \mathcal{U}, \tag{5.2}$$

or as

$$\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v})) - \mu(\mathbf{u},\mathbf{v};h)h^{\nu}(\mathbf{v}-\mathbf{u}), \quad \forall\, \mathbf{u},\mathbf{v} \in \mathcal{U}, \tag{5.3}$$

where $\nu \in \mathbb{N}$ and for some matrix-valued function $\mu$, which has the property that for any compact $K \subset \mathcal{U}$ there exists a $\mu_K > 0$, such that $|\mu(\mathbf{u},\mathbf{v};h)| \leq \mu_K(1 + \frac{|\mathbf{v}-\mathbf{u}|}{h})$, for $h$ small enough.

**Remark 5.1.**
*For our numerical computations we consider the following numerical fluxes.*

- *The upwind numerical flux:* $\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v}))$ *with* $w(\mathbf{u},\mathbf{v}) = \mathbf{u}$ *satisfies* (5.2)
- *The Lax-Wendroff flux:* $\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v}))$ *with* $w(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}+\mathbf{v}}{2} - \frac{\Delta t}{2h}(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}))$*, satisfies* (5.2)*.*
- *The Lax-Friedrichs flux :* $\hat{\mathbf{F}}(u,\mathbf{v}) = \frac{1}{2}\Big(F(u) + F(w)\Big) + \lambda(w-u)$ *satisfies* (5.3)*, with* $\nu = 0$*,* $w(\mathbf{u},\mathbf{v}) := \frac{1}{2}(\mathbf{u}+\mathbf{v})$ *and* $\mu(\mathbf{u},\mathbf{v};h) := \lambda I - \frac{\mathbf{F}(\mathbf{u})-2\mathbf{F}(w(\mathbf{u},\mathbf{v}))+\mathbf{F}(\mathbf{v})}{2|\mathbf{v}-\mathbf{u}|^2} \otimes (\mathbf{u}-\mathbf{v})$.

We define the spatial reconstruction which is applied to the temporal reconstruction $\hat{\mathbf{u}}^t(t,\cdot)$ for each $t \in (0,T)$ using the function $w$ (cf. [28, 46]).

**Definition 5.2** (Space-time reconstruction)**.**
*Let* $\hat{\mathbf{u}}^t$ *be the temporal reconstruction of a sequence* $\{\mathbf{u}_h\}_{n=0}^{N_t}$ *of solutions of the fully discrete scheme of* (SG) *using a numerical flux satisfying* (5.2) *or* (5.3)*. The space-time reconstruction* $\hat{\mathbf{u}}^{st}(t,\cdot) \in \mathscr{V}_h^{q+1}$ *is defined as the solution of*

$$\sum_{l=0}^{N_s-1} \int_{x_l}^{x_{l+1}} (\hat{\mathbf{u}}^{st}(t,\cdot) - \hat{\mathbf{u}}^t(t,\cdot)) \cdot \mathbf{v_h}\, \mathrm{d}x = 0 \quad \forall\, \mathbf{v_h} \in \mathscr{V}_h^{q-1},$$

$$\hat{\mathbf{u}}^{st}(t,x_l^{\pm}) = w(\hat{\mathbf{u}}^t(t,x_l^-),\hat{\mathbf{u}}^t(t,x_l^+)) \quad \forall\, l = 0,\dots,N_s.$$

We have the following property of the space-time reconstruction.

**Lemma 5.3** ([28], Lemma 24)**.**
*Let $\hat{\mathbf{u}}^{st}$ be the space-time reconstruction from Definition 5.2. For each $t \in (0,T)$, the function $\hat{\mathbf{u}}^{st}(t,\cdot)$ is well defined. Moreover,*

$$\hat{\mathbf{u}}^{st} \in W_\infty^1((0,T); \mathscr{V}_h^{q+1} \cap C^0(D;\mathcal{U})).$$

Since $\hat{\mathbf{u}}^{st}$ is Lipschitz continuous in space and time, we can compute the following space-time residual.

**Definition 5.4** (Space-time residual)**.**
*We call the function $\mathbf{R}^{st} \in L^2((0,T)\times D; \mathbb{R}^{m\cdot(K+1)})$ satisfying*

$$\mathbf{R}^{st}(t,x) := \partial_t \hat{\mathbf{u}}^{st}(t,x) + \partial_x \mathbf{F}(\hat{\mathbf{u}}^{st}(t,x)) \tag{5.4}$$

*the space-time (or deterministic) residual .*

Expanding the space-time reconstruction $\hat{\mathbf{u}}^{st}$ in the finite orthonormal system $\{\Psi_k(y)\}_{k=0}^K$ enables us to consider the so-called space-time-stochastic residual. We define

$$\overline{\mathscr{V}}_h^{q+1} := \{u : D \to \mathbb{R}^m \mid u\big|_{(x_l,x_{l+1})} \in \mathbb{P}_{q+1}((x_l,x_{l+1});\mathbb{R}^m), \text{ for } 0 \le l \le N_s - 1\}.$$

**Definition 5.5** (Space-time-stochastic reconstruction and space-time stochastic residual for SG)**.**
*The function $\hat{u}^{sts} \in \mathscr{W}_K(\Xi) \otimes W_\infty^1((0,T); \overline{\mathscr{V}}_h^{q+1} \cap C^0(D;\mathbb{R}^m))$, which is defined as*

$$\hat{u}^{sts}(t,x,y) := \sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k(t,x)\Psi_k(y) \tag{5.5}$$

*is called space-time-stochastic reconstruction of the numerical approximation of* (5.1)*, which is defined as (cf.* (4.7)*): $u_h^n(x,y) := \sum_{k=0}^{K} (\mathbf{u}_h^n)_k(x)\Psi_k(y)$, for all $n = 0,\dots N_t$.*

*Moreover, we define the space-time-stochastic residual $\mathscr{R}^{sts} \in L_w^2(\Xi; L^2((0,T)\times D;\mathbb{R}^m))$ by*

$$\mathscr{R}^{sts}(t,x,y) := \partial_t\Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k(t,x)\Psi_k(y)\Big) + \partial_x F\Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k(t,x)\Psi_k(y)\Big). \tag{5.6}$$

We are now ready to state the main a posteriori error estimate for the difference between the random entropy admissible solution and the numerical approximation obtained with SG. Its proof follows immediately from Theorem 2.15.

**Theorem 5.6** (A posteriori error bound for the reconstruction of the numerical solution)**.**
*Let $u$ be a random entropy solution of* (5.1)*. Let the reconstruction $\hat{u}^{sts}$ from Definition 5.5*

*only take values in a compact, convex set $\mathscr{C} \subset \mathscr{U}$. Then, the difference between u and the reconstruction $\hat{u}^{sts}$ satisfies*

$$\begin{aligned}\|u(s,\cdot,\cdot)-\hat{u}^{sts}(s,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \leq & 2C_{\underline{\eta}}^{-1}\Big(\mathscr{E}^{sts}(s)+C_{\overline{\eta}}\mathscr{E}_0^{sts}\Big)\\ & \times\exp\Big(\int_0^s\Big(\frac{C_{\overline{F}}C_{\overline{\eta}}\|\partial_x\hat{u}^{sts}(t,\cdot,\cdot)\|_{L^\infty_w(\tilde{\Xi};L^\infty(D))}+C_{\overline{\eta}}^2}{C_{\underline{\eta}}}\Big)\ \mathrm{dt}\Big),\end{aligned}$$

*for $0\leq s\leq T$ and for any $\tilde{\mathbb{P}}$-measurable set $\tilde{\Xi}\subset\Xi$. Here,*

$$\mathscr{E}^{sts}(s) := \|\mathscr{R}^{sts}\|^2_{L^2_w(\tilde{\Xi};L^2((0,s)\times D))}, \tag{5.7}$$

$$\mathscr{E}_0^{sts} := \|u^0-\hat{u}^{sts}(0,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))}. \tag{5.8}$$

By means of the triangle inequality we reformulate Theorem 5.6 in terms of the numerical approximation.

**Corollary 5.7** (A posteriori error bound for the numerical solution)**.**
*Let u be a random entropy solution of* (5.1)*. Then, the difference between u and the numerical solution $u_h$ from Definition 5.5 satisfies*

$$\begin{aligned}\|u(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \leq & 2\|\hat{u}^{sts}(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))}+2C_{\underline{\eta}}^{-1}\Big(\mathscr{E}^{sts}(t_n)+C_{\overline{\eta}}\mathscr{E}_0^{sts}\Big)\\ & \times\exp\Big(\int_0^s\Big(\frac{C_{\overline{F}}C_{\overline{\eta}}\|\partial_x\hat{u}^{sts}(t,\cdot,\cdot)\|_{L^\infty_w(\tilde{\Xi};L^\infty(D))}+C_{\overline{\eta}}^2}{C_{\underline{\eta}}}\Big)\ \mathrm{dt}\Big),\end{aligned}$$

*for all $n=0,\dots N_t$ and for any $\tilde{\mathbb{P}}$-measurable set $\tilde{\Xi}\subset\Xi$.*

While Corollary 5.7 provides error control between the exact and numerical solution we are not able to identify if the numerical error is dominated by the spatio-temporal discretization error introduced by the deterministic RKDG scheme, or the stochastic discretization error due to truncation of the infinite gPC-series in (4.2). Hence, we would like to derive a splitting of the upper bound into a space-time (deterministic) and stochastic part. This is the statement of the following theorem, where we decompose the space-time-stochastic residual into a deterministic and stochastic residual. Before stating the orthogonal decomposition we give a definition of the stochastic residual.

**Definition 5.8** (Stochastic residual)**.**
*Let $K\in\mathbb{N}_0$ be the order of the space-time-stochastic reconstruction from* (5.5)*. The entries of the stochastic residual are defined as*

$$(\mathbf{R}^{stoch})_k := \Big\langle\mathscr{R}^{sts},\Psi_k\Big\rangle = \Big\langle\partial_x F(\hat{u}^{sts}),\Psi_k\Big\rangle, \tag{5.9}$$

*for all $k=K+1,K+2,\dots$.*

**Theorem 5.9** (Orthogonal decomposition of the space-time-stochastic residual)**.**
*The space-time-stochastic residual $\mathscr{R}^{sts}$, from Definition 5.5, admits the following orthogonal decomposition in $L^2_w(\Xi)$,*

$$\mathscr{R}^{sts} = \sum_{k=0}^{K} (\mathbf{R}^{st})_k \Psi_k + \sum_{k=K+1}^{\infty} (\mathbf{R}^{stoch})_k \Psi_k. \tag{5.10}$$

*Further, $\mathscr{E}^{sts}(s)$ from* (5.7) *admits the following decomposition,*

$$\mathscr{E}^{sts}(s) = \mathscr{E}^{st}(s) + \mathscr{E}^{stoch}(s), \tag{5.11}$$

*where $\mathscr{E}^{st}(s) := \sum_{k=0}^{K} \|(\mathbf{R}^{st})_k\|^2_{L^2((0,s)\times D)}$ and $\mathscr{E}^{stoch}(s) := \sum_{k=K+1}^{\infty} \|(\mathbf{R}^{stoch})_k\|^2_{L^2((0,s)\times D)}$, for all $s \in (0,T]$.*

*Proof.* We first consider the projection of the space-time-stochastic residual $\mathscr{R}^{sts}$ onto the subspace $\operatorname{span}\{\Psi_0,\dots,\Psi_K\}$. Let $l = 0,\dots,K$:

$$\begin{aligned}\left\langle \mathscr{R}^{sts}, \Psi_l \right\rangle = \left\langle \partial_t \hat{u}^{sts} + \partial_x F(\hat{u}^{sts}), \Psi_l \right\rangle &= \int_{\Xi} \Big( \partial_t \sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_l + \partial_x F\big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k\big)\Big) \Psi_l \; w(y)\mathrm{d}y \\ &= \partial_t (\hat{\mathbf{u}}^{st})_l + (\partial_x \mathbf{F}(\hat{\mathbf{u}}^{st}))_l = (\mathbf{R}^{st})_l,\end{aligned}$$

where the last equality follow from the orthogonality relation (4.1). Thus, $\left\langle \mathscr{R}^{sts}, \Psi_l \right\rangle$ coincides with the coefficients of the space-time residual $\mathbf{R}^{st}$ from Definition 5.4. Analogously we consider for $l > K$:

$$\left\langle \mathscr{R}^{sts}, \Psi_l \right\rangle = \left\langle \partial_t \hat{u}^{sts} + \partial_x F(\hat{u}^{sts}), \Psi_l \right\rangle = \left\langle \partial_x F(\hat{u}^{sts}), \Psi_l \right\rangle = (\mathbf{R}^{stoch})_l$$

Formula (5.10) follows from the previous computations. Finally, Formula (5.11) is an application of the Pythagorean theorem for the weighted Hilbert space $L^2_w(\Xi)$, applied to $\mathscr{E}^{sts}(s) = \|\mathscr{R}^{sts}\|^2_{L^2_w(\Xi;L^2((0,T)\times D))}$. □

**Remark 5.10.**
*In the same manner as for Theorem 5.9 we can find an orthogonal decomposition of the approximation error in the initial condition measured by $\mathscr{E}^{sts}_0$ given in* (5.8)*. Let us define the orthogonal projection $\Pi_K(u^0) := \sum_{k=0}^{K} \langle u^0, \Psi_k\rangle \Psi_k$. The Pythagorean theorem implies*

$$\begin{aligned}\mathscr{E}^{sts}_0 = \|u^0 - \hat{u}^{sts}(0,\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))} &= \|u^0 - \Pi_K(u^0)\|^2_{L^2_w(\Xi;L^2(D))} + \|\Pi_K(u^0) - \hat{u}^{sts}(0,\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))} \\ &= \|\sum_{k=K+1}^{\infty} \langle u^0, \Psi_k\rangle\|^2_{L^2(D)} + \sum_{k=0}^{K} \|\langle u^0, \Psi_k\rangle - \hat{\mathbf{u}}^{st}_k(0,\cdot)\|^2_{L^2(D)} \\ &=: \mathscr{E}^{stoch}_0 + \mathscr{E}^{st}_0. \end{aligned} \tag{5.12}$$

Summarizing Theorem 5.9, the coefficients of $\mathscr{R}^{sts}$ in the basis $\{\Psi_0,\dots,\Psi_K\}$ are the coefficients of $\mathbf{R}^{st}$ from Definition 5.4 and thus these coefficients are associated with the space-time error that arises when approximating the truncated (deterministic) (SG)-system. On the other hand for the remaining coefficients of $\mathscr{R}^{sts}$ there is no such interpretation. But since the (SG)-system is obtained by an orthogonal projection onto $\operatorname{span}\{\Psi_0,\dots,\Psi_K\}$ the remaining coefficients of $\mathscr{R}^{sts}$ contribute to the stochastic error that arises when truncating the infinite Fourier series (4.2). We can now combine Corollary 5.7, Theorem 5.9 and Remark 5.10 to obtain an a posteriori error bound for the error between the random entropy admissible solution and its numerical approximation, which provides separate bounds for the spatial discretization error and the stochastic discretization error

**Theorem 5.11** (A posteriori error bound for the numerical solution with error splitting)**.**
*Let $u$ be a random entropy solution of* (5.1)*. Then, the difference between $u$ and the numerical solution $u_h$ from Definition 5.5 satisfies*

$$\begin{aligned}\|u(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))} \le\ & 2\|\hat{u}^{sts}(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))}\\ &+2C_{\underline{\eta}}^{-1}\Big(\mathscr{E}^{st}(t_n)+\mathscr{E}^{stoch}(t_n)+C_{\overline{\eta}}(\mathscr{E}_0^{st}+\mathscr{E}_0^{stoch})\Big)\\ &\times\exp\Big(\int_0^s\Big(\frac{C_{\overline{F}}C_{\overline{\eta}}\|\partial_x\hat{u}^{sts}(t,\cdot,\cdot)\|_{L^\infty_w(\Xi;L^\infty(D))}+C_{\overline{\eta}}^2}{C_{\underline{\eta}}}\Big)\ \mathrm{d}t\Big),\end{aligned}$$

*for all $n=0,\dots N_t$.*

Theorem 5.11 allows us to decompose the error estimator for the space-time-stochastic error into parts quantifying the stochastic and the spatio-temporal discretization error, respectively. The stochastic error, introduced by truncating the Fourier series in (4.2), can be quantified by $\mathscr{E}^{stoch}(t_n)+\mathscr{E}_0^{stoch}$. The spatio-temporal discretization error, i.e., the error which arises by discretizing the system (SG) in space and time, can be quantified by $\mathscr{E}^{st}(t_n)+\mathscr{E}_0^{st}$.

In the case of the linear advection and Burgers equation, we can derive an analytical expression for $\mathbf{R}^{stoch}$.

**Example 5.12.** *(a) In the case of the linear advection equation $F(u)=au$, $a\in\mathbb{R}$, we have*

$$\begin{aligned}(\mathbf{R}^{stoch})_l =& \int_\Xi\Big(\partial_x F\big(\sum_{k=0}^K(\hat{\mathbf{u}}^{st})_k\Psi_k(y)\big)\Big)\Psi_l(y)\ w(y)\mathrm{d}y\\ =& a\sum_{k=0}^K\partial_x(\hat{\mathbf{u}}^{st})_k\int_\Xi\Psi_k(y)\Psi_l(y)\ w(y)\mathrm{d}y=0,\end{aligned}$$

*for $l>K$, due to orthogonality. Thus, $\mathbf{R}^{stoch}$ and consequently $\mathscr{E}^{stoch}$ vanishes. This means that the stochastic error is only inferred from projecting the initial condition onto*

*the orthonormal system, and can be measured by $\mathscr{E}_0^{stoch}$ . Due to linearity of the advection equation the initial stochastic error does not increase when advancing in time.*

*(b) If the advection velocity is random, i.e. $F(u,y) = a(y)u$, then the stochastic residual becomes*

$$(\mathbf{R}^{stoch})_l = \int_\Xi \Big(\partial_x F\big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k(y), y\big)\Big)\Psi_l(y)\; w(y)\mathrm{d}y$$
$$= \sum_{k=0}^{K} \partial_x (\hat{\mathbf{u}}^{st})_k \int_\Xi a(y)\Psi_k(y)\Psi_l(y)\; w(y)\mathrm{d}y,$$

*and in this case, due to the uncertain flux function, the stochastic error increases when advancing in time.*

*(c) In the case of Burgers' equation $F(u) = \frac{u^2}{2}$, we compute for $l > K$,*

$$(\mathbf{R}^{stoch})_l = \int_\Xi \Big(\partial_x F\big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k(y)\big)\Big)\Psi_l(y)\; w(y)\mathrm{d}y$$
$$= \frac{1}{2}\int_\Xi \partial_x \Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k(y)\Big)^2 \Psi_l(y)\; w(y)\mathrm{d}y.$$

*Because $\Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k\Big)^2$ is a polynomial of degree $2K$ and due to orthogonality, it follows that*

$$\int_\Xi \partial_x F\Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k(y)\Big)\Psi_l(y)\; w(y)\mathrm{d}y = 0$$

*for $l > 2K$. We therefore only need to compute $(\mathbf{R}^{stoch})_l$ for $l = K+1,\dots,2K$. Moreover,*

$$\frac{1}{2}\int_\Xi \partial_x \Big(\sum_{k=0}^{K} (\hat{\mathbf{u}}^{st})_k \Psi_k(y)\Big)^2 \Psi_l(y)\; w(y)\mathrm{d}y = (\partial_x \hat{\mathbf{u}}^{st})^\top C_l \hat{\mathbf{u}}^{st},$$

*where $[C_k]_{k,l=0}^{K} := \int_\Xi \Psi_k(y)\Psi_l(y)\Psi_k(y)\; w(y)\mathrm{d}y$, $k = K+1,\dots,2K$, is the symmetric and sparse polynomial chaos tensor [85].*

Before presenting numerical experiments concerning the scaling behavior of the space-time and stochastic residual we want to make some important remarks.

**Remark 5.13.** *1. In [28] it has been shown that for linear equations, i.e. $F(u) = au$, $a \in \mathbb{R}$, using a numerical flux satisfying* (5.2) *and provided the numerical error converges as $\mathscr{O}(h^{q+1})$, the deterministic residual $\mathscr{E}^{st}$ converges with the same order as the numerical error. Hence, in this case we have optimality of the residual and an optimal a posteriori error control.*

2. *In the nonlinear case, the optimality of the residual for smooth solution in the first time-step is still open, cf. the discussion in [49].*

3. *Whenever the solution of system (SG) exhibits discontinuities in the spatial variable, the quantity* $\|\partial_x \hat{u}^{sts}(t,\cdot,\cdot)\|_{L^\infty_w(\tilde{\Xi};L^\infty(D))}$ *is expected to scale like* $h^{-1}$ *and hence the estimator will blow up for* $h \to 0$ *in the vicinity of discontinuities. We therefore have only reliable a posteriori error control in the pre-shock case. However, as the residuals clearly capture important flow information we use the residuals in Section 5.2 as indicators for local mesh refinement, even in the post-shock case.*

4. *For linear equations we have that* $C_{\overline{F}} = 0$*. Therefore, for linear equations we expect no blow-up of the estimator for* $h \to 0$ *even in case the solution is discontinuous.*

5. *Our numerical experiments in Section 5.1.2 indicate that* $\mathscr{E}^{stoch}$ *exhibits spectral convergence. We are currently not able to give a proof for this assertion. From a heuristic point of view, the quantities* $(\mathbf{R}^{stoch})_k = \left\langle \partial_x F(\hat{u}^{sts}), \Psi_k \right\rangle$ *are the Fourier coefficients of* $\partial_x F(\hat{u}^{sts})$ *in the basis* $\{\Psi_k\}_{k=0}^{\infty}$ *and hence, if* $y \mapsto \partial_x F(\hat{u}^{sts}(t,x,y))$ *is sufficiently regular, the coefficients decay spectrally.*

**Remark 5.14.**
*For scalar random conservation laws it is also possible to derive an a posteriori error estimator using the deterministic a posteriori error estimate in Theorem 3.2. We decided to use the relative entropy framework of Dafermos and DiPerna based on the following two reasons. First, for smooth solutions, upper bounds for the numerical error that result from Kružkov's theory are only of half-order, i.e. have a rate of convergence of* $(q+1)/2$*, when* $q \in \mathbb{N}$ *is the polynomial degree, cf. [54, 71]. In contrast, the worst case for error bounds for smooth solutions obtained using the relative entropy framework is loosing one order, i.e. the rate of convergence is q. Second, the relative entropy framework requires only one strictly convex entropy/entropy flux pair. In contrast to Kružkov's theory it is also applicable for random hyperbolic systems of conservation laws.*

### 5.1.2 Numerical experiments

In this section we present numerical experiments for scalar random conservation laws, where we examine the scaling behavior of the space-time-stochastic residual. For the following test cases, we consider the classical polynomial chaos expansion using Legendre orthonormal polynomials for uniformly distributed random variables. As numerical solver for the SG system we use SG-FLEXI [12, 63]. For the time-stepping we use a low storage SSP RK-method of order three as in

[65], a time-reconstruction of order three and for the physical space we use DG polynomials of degree one or two. As numerical flux for the linear advection equation, we choose the upwind flux,

$$\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v})),\ w(\mathbf{u},\mathbf{v}) = \mathbf{u},$$

and as projection operator for the initial data, we choose the Radau-projection operator $\mathscr{L}_{\mathscr{V}_h^q} = \mathbb{R}_h^+$, as defined in [111]. For the Burgers equation, in contrast to the linear advection equation, the direction of transport of information depends on the state and we therefore select the Lax-Wendroff numerical flux

$$\hat{\mathbf{F}}(\mathbf{u},\mathbf{v}) = \mathbf{F}(w(\mathbf{u},\mathbf{v})),\ w(\mathbf{u},\mathbf{v}) = \frac{1}{2}\Big((\mathbf{u},\mathbf{v}) + \frac{\Delta t}{h}\big(\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{u})\big)\Big).$$

For Burgers' equation we use Gauss-Legendre interpolation of the initial data in our numerical experiments. We have also tried other interpolation and projection operators and all of them lead to the same scaling behavior of the space-time residual described in Remark 5.15. The space, time and stochastic integrals are numerically evaluated by a Gauß–Legendre quadrature. For the time integration we use 7 points, in physical space we use 25 points and in random space 80 points. In the following we call $\|u(T,\cdot,\cdot) - u_h^{N_t}\|_{L_w^2(\Xi;L^2(D))}$ numerical error and we plot the quantities $\mathscr{E}^{st}(T)$, $\mathscr{E}^{stoch}(T)$ and the numerical error as in Theorem 5.11 at final computational time $T$.

**Remark 5.15.** *1. As already mentioned in Remark 5.14, for nonlinear hyperbolic equations, the space-time residual is suboptimal (by one order) on the first time-step. Indeed, we loose half an order of convergence in the (global) space-time residual, i.e. when the error is of order $h^{p+1}$ the error estimator is of order $h^{p+1/2}$. This is due to a lack of compatibility between the projection/interpolation of the initial data into the DG space and the spatial-reconstruction. For the linear advection equation, where we compute the numerical solution using an upwind numerical flux, the Radau projection is the compatible choice, as it accounts for the upwind direction. Indeed, we have used it in our numerical experiments and observed optimal rates for the error estimator. A similar concept for nonlinear equations is up to now missing.*

*2. If we start to reconstruct the numerical solution of the Burgers equation from $t = 0$, we loose half an order of convergence in the space-time residual. Therefore, we start to reconstruct the numerical solution after the first time-step on the coarsest mesh used for our computations. This corresponds to $t = 0.008$. We also start to integrate the space-time residual from $t = 0.008$ and obtain the full order of convergence in the space-time residual.*

### The linear advection equation

We consider the linear advection equation

$$\partial_t u + 2\partial_x u = 0, \tag{5.13}$$

on the spatial domain $[0,2]_{per}$ and with $T = 0.2$. We start the computation with 16 elements and a time-step size of $\Delta t = 0.02$. We then subsequently reduce $h$ and $\Delta t$ by a factor of two. The initial condition is given by $u^0(x,y) = y(1-0.5\cos(\pi x))$, where we assume $\xi \sim \mathscr{U}[1,3]$ to be uniformly distributed.

In Figures 5.1 and 5.2 we show the numerical error between the exact solution $u(t,x,y) = y(1-0.5\cos(\pi(x-2t)))$ and the numerical solution computed by the RKDG method for one and three chaos polynomials ($K = 0,2$), evaluated at $t_n = T$. Thanks to $C_{\overline{F}} = 0$ in this special case, the exponential term of the error indicator from Theorem 5.11 vanishes. In Figure 5.1 we can see that the numerical error is not decreasing when $h$ tends to zero. This is due to the term $\mathscr{E}_0^{stoch}$, cf. Remark 5.10. The overall error is dominated by the error we make in projecting the initial condition onto $\text{span}\{\Psi_0\}$. If we increase the number of orthonormal polynomials to three, we obtain an exact representation of the initial condition in the orthonormal basis, i.e., $\mathscr{E}_0^{stoch} = 0$. Therefore, the numerical error only consists of the spatio-temporal discretization error, quantified by $\mathscr{E}^{st}$, this can be seen in Figure 5.2. After increasing the polynomial chaos degree, the numerical error decreases with the same order as the spatial residual. Furthermore, for $q = 1,2$ both residuals have the correct order of convergence, that is $q+1$.

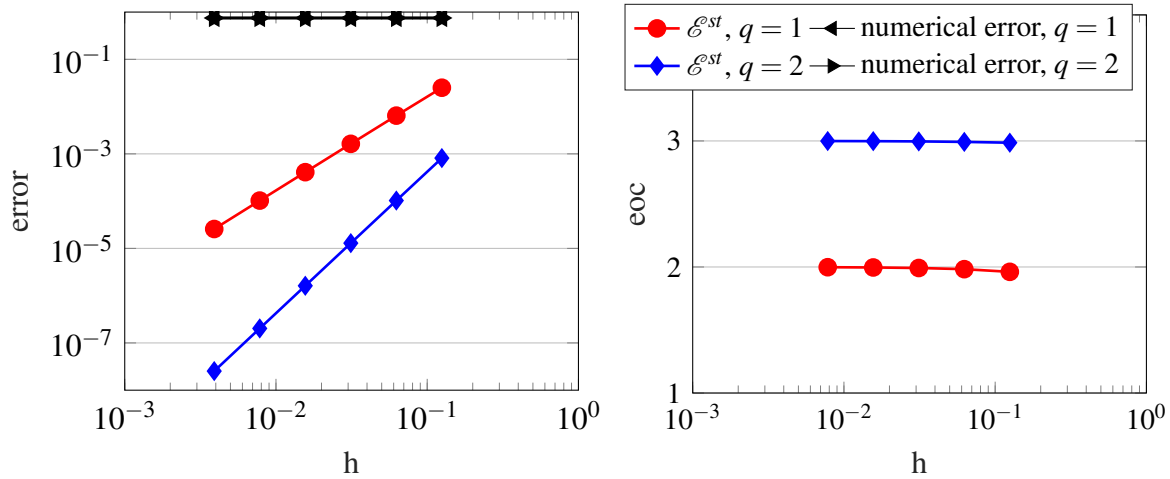



Figure 5.1: Error and eoc plot for the linear advection equation in the case of one orthonormal polynomial and DG polynomial degrees $q = 1,2$. Example (5.13).

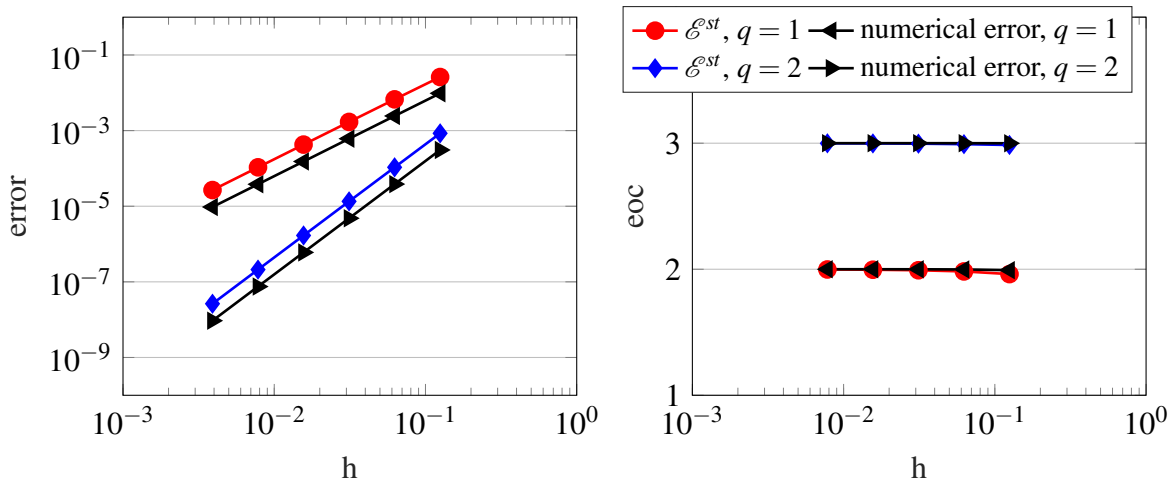



Figure 5.2: Error and eoc plot for the linear advection equation in the case of three orthonormal polynomials and DG polynomial degrees $q = 1,2$. Example (5.13).

**The Burgers equation with smooth solution**

In this numerical example we consider Burgers' equation

$$\partial_t u + \partial_x(\frac{u^2}{2}) = S \tag{5.14}$$

with source term

$$S(t,x,y) = \pi y^2 \sin(\pi(x-yt))\big(\cos(\pi(x-yt))-1\big),$$

$\xi \sim \mathscr{U}[1,3]$. For the initial condition $u^0(x,y) = y\cos(\pi x)$, the exact solution for Burgers' equation is given by

$$u(t,x,y) = y\cos(\pi(x-yt)).$$

The numerical solution is computed up to time $T = 0.2$ on the spatial domain $D = [0,2]_{per}$ and we start the computations initially on a mesh with 16 elements and time-step size $\Delta t = 0.008$. Again we reduce $h$ and $\Delta t$ by a factor of two. The DG polynomial degree is two and the reconstruction in time is of order three. In the following numerical computations we consider different cases, where on the one hand we refine the physical space and on the other hand we increase the polynomial degree of the orthonormal polynomials. The latter corresponds to a refinement in the random space.

We start our computations by considering mesh refinements in the physical space with a fixed polynomial chaos degree. In Figure 5.3 (a) we display the numerical solution using only one chaos polynomial, which corresponds to $K = 0$. We can see that the numerical error is clearly dominated by $\mathscr{E}^{stoch}$. To reduce the overall error significantly, we have to increase the polynomial chaos degree. We increase the number of polynomials to five, corresponding to $K = 4$. We

observe in Figure 5.3 (b), that for the coarse space discretization with 16 elements the numerical error is dominated by $\mathscr{E}^{st}$. However, after this point any significant reduction of the numerical error requires again an increase of the polynomial chaos degree as the numerical error is then again dominated by $\mathscr{E}^{stoch}$. When increasing the polynomial degree to thirteen ($K = 12$), we can see in Figure 5.3 (c) that the numerical error is now dominated by $\mathscr{E}^{st}$, as the stochastic discretization is fine enough. The numerical error now converges with the same rate as the space-time residual. Additionally, in Figure 5.3 (d) we plot the exponential factor from Theorem 5.11 for different polynomial degrees $K$ and different mesh sizes $h$. We can see that in the smooth case the exponential factor stays bounded for $h \to 0$. The exponential factor for $K = 0$ is smaller than for $K = 4, 12$ because we solve the (SG)-system only for the mean value.

In the previous computations we have considered spatial refinements for a fixed polynomial chaos degree. Now we want to examine the behavior of $\mathscr{E}^{stoch}$ for different mesh sizes and a DG polynomial degree two. In Figure 5.4 (a) we show results for a fixed spatial discretization with 16 elements. We can see that the numerical error is dominated by $\mathscr{E}^{st}$, because the spatial discretization is too coarse. We can also see that $\mathscr{E}^{st}$ remains unchanged by increasing the polynomial chaos degree. Additionally, we note that $\mathscr{E}^{stoch}$ exhibits spectral convergence. To reduce the numerical error we therefore need to increase the number of spatial elements or the DG polynomial degree. Finally, in Figure 5.4 (b) we consider a very fine mesh, consisting of $N_s = 1024$ elements. Due to the fine resolution of the physical space, the numerical error is dominated by $\mathscr{E}^{stoch}$ up to $K = 8$. After that point, the numerical error can only be significantly decreased by performing a spatial refinement. We can now also see how the numerical error converges spectrally until its convergence is again dominated by the spatial error.

Let us note that the spectral convergence of $\mathscr{E}^{stoch}$ is in accordance with what is known theoretically for stochastic discretization errors in SG schemes. Indeed, the authors of [55] prove spectral convergence of the SG method for the linear transport equation with random transport velocity. This indicates that the stochastic residual allows us to assess the magnitude of the stochastic discretization error in an optimal (spectral) way.

**The Burgers equation, the artificial shock case**

We study now the same example as in the previous section,

$$\partial_t u + \partial_x(\frac{u^2}{2}) = S, \tag{5.15}$$

but we compute numerical solutions up to $T = 0.56$ and use the slope limiter $\Lambda\Pi_h$ from [24]. It is a well-known drawback of the SG-methodology, cf. the numerical example in Section 6 of

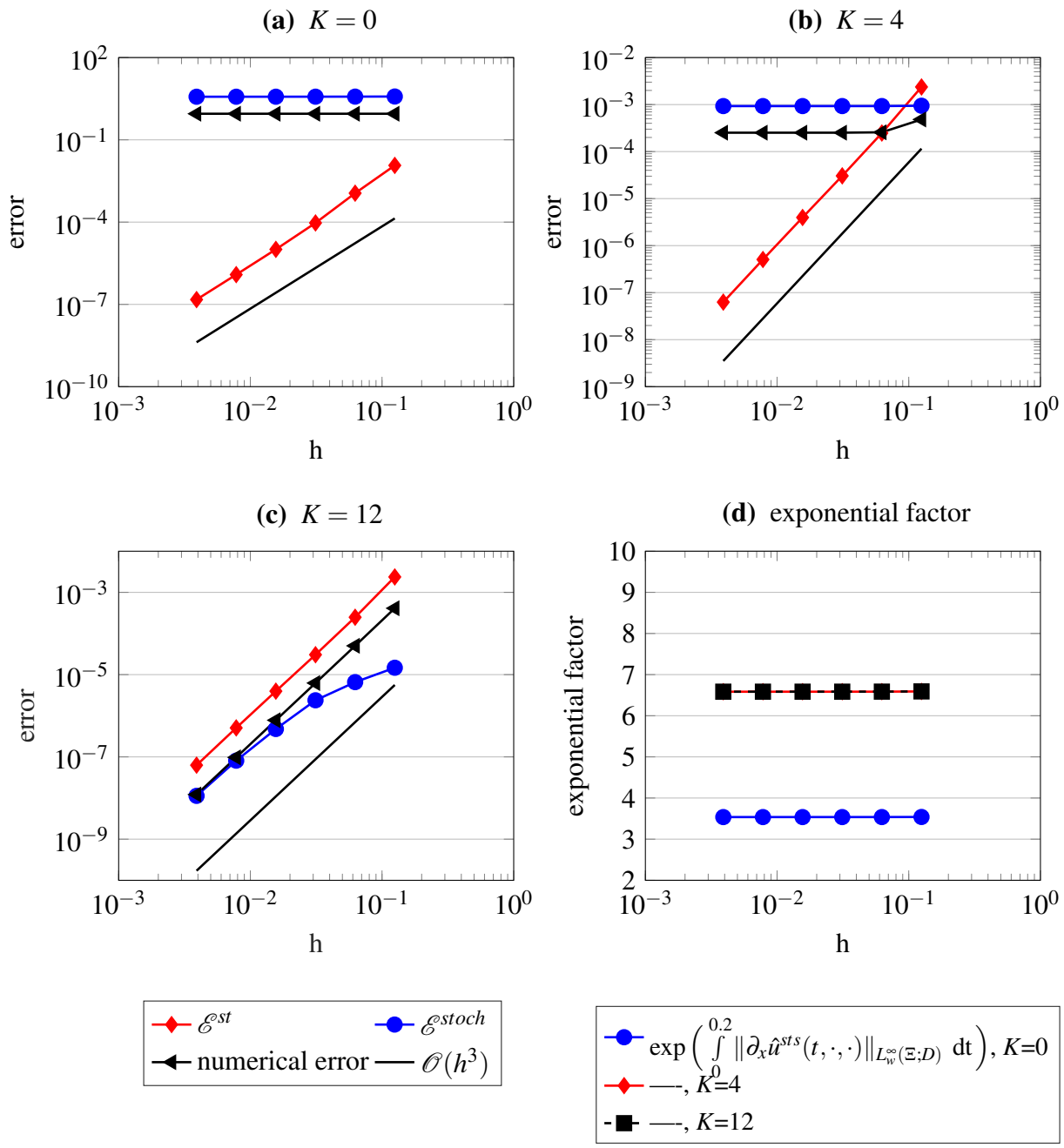



Figure 5.3: Error plots for the Burgers equation for $h$-refinement, $q = 2$. Example (5.14).

[85], that, even if the solution to (5.1) is smooth up to time $T$, the solutions of the (SG)-system may develop discontinuities before that time. Indeed, this is the case in the example at hand. In Figure 5.5(a), we display the numerical solutions and residuals for $K = 1$ at time $T$. The solution of the zeroth mode $u_0$ appears to contain a shock at approximately $x \approx 1.6$. We see that the spatial discontinuity is clearly picked up by the zeroth mode $\mathbf{R}_0^{st}$ of the spatial residual and, to some extent, also in the second mode $\mathbf{R}_2^{st}$ of the stochastic residual. We may therefore use

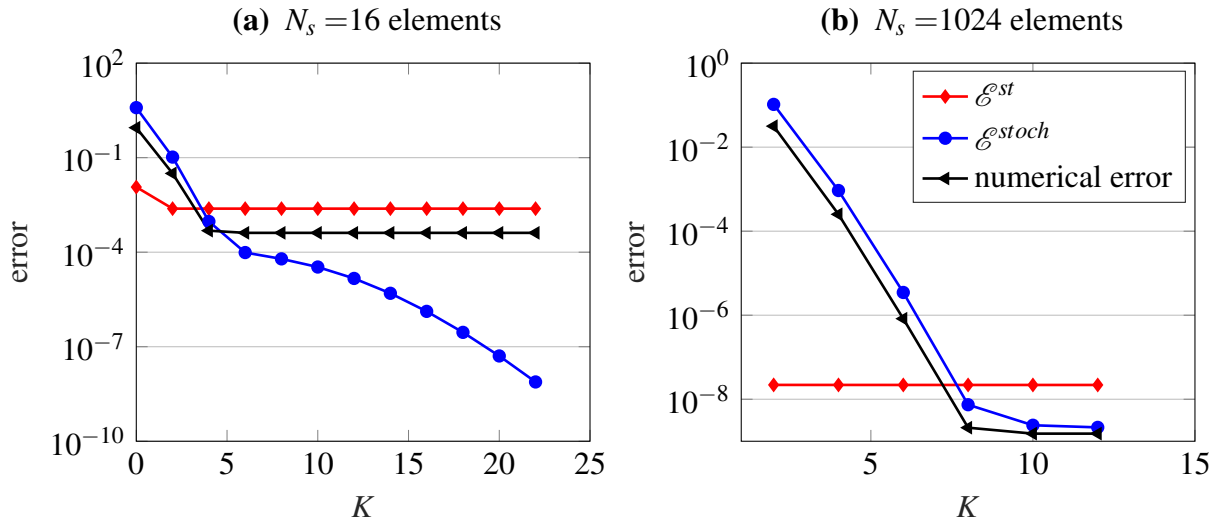
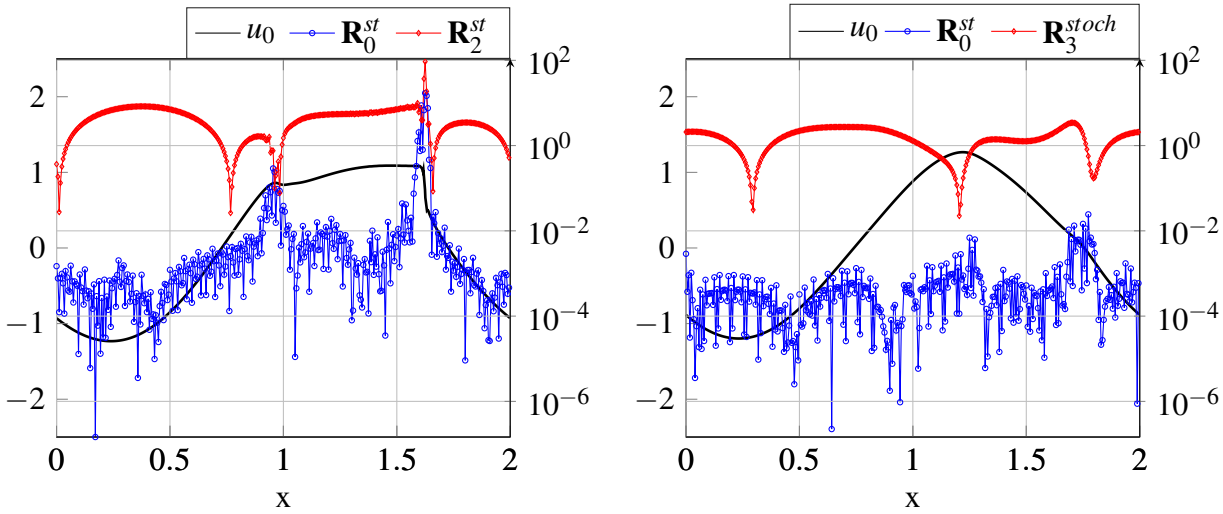



Figure 5.4: Error plots for the Burgers equation for $K$-refinement, $q = 2$. Example (5.14).

the spatial residual as an indicator for local spatial mesh refinements, which will be considered in Section 5.2.




(a) $K = 1$, residuals indicated on the right y-axis.

(b) $K = 2$, residuals indicated on the right y-axis.

Figure 5.5: Plot of the numerical solution, spatial residual and stochastic residual for the Burgers equation in the case of two and three orthonormal polynomial and DG polynomial degree $q = 2$. Example (5.15).

Moreover, we can see in Figure 5.6, that for $h \to 0$ the spatial residual blows up, although the numerical error stays constant. Also $\mathscr{E}^{stoch}$ grows with $h \to 0$, however very slow compared to $\mathscr{E}^{st}$. Increasing the polynomial chaos degree to $K = 2$, also increases the smoothness of the

numerical solution, which can be seen in Figure 5.5 (b). In Figure 5.6, we can see that for $K = 2$, $\mathscr{E}^{st}$ decreases when $h$ tends to zero.

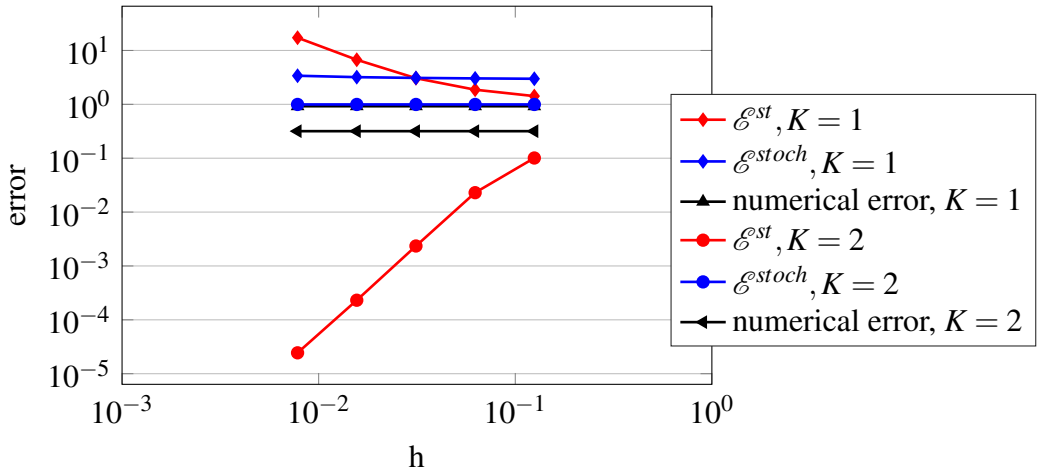



Figure 5.6: Comparison of residuals and numerical error for one and two orthonormal polynomials and DG polynomial degree $q = 2$. Example (5.15).

**The Burgers equation, the shock case**

As last numerical example we consider a Riemann problem for Burgers' equation, where the jump size of the discontinuity and thus the shock speed is random. The spatial domain is now $[-1,1]$ and we set $T = 0.1$. The initial condition is given by

$$u^0(x,y) = \begin{cases} 1+y, & \text{if } x \leq 0, \\ 0.5+y & \text{else,} \end{cases}$$

where $\xi \sim \mathscr{U}[-0.2,0.2]$. The shock speed can be computed as $s(y) = 1.5 + 2y$ and therefore the exact solution for Burgers' equation is given by

$$u(t,x,y) = \begin{cases} 1+y, & \text{if } x \leq s(y)t \\ 0.5+y & \text{else.} \end{cases}$$

As the shock speed is $\mathbb{P}-$a.s. positive, we use the upwind numerical flux in this numerical experiment. We use the slope limiter $\Lambda\Pi_h$ from [24] and consider a very fine physical resolution with 512 elements and DG polynomial degree of two. Although the exact solution is discontinuous we can see in Figure 5.7 that $\mathscr{E}^{stoch}$ exhibits exponential convergence an increasing number of orthonormal polynomials. This is due to the fact that the coefficients of $u$ in the polynomial chaos expansion are smooth, cf. [85, Section 6]. Hence, although the exact solution

is discontinuous $\mathscr{E}^{stoch}$displays the correct (spectral) type of convergence and moreover, it gives us information about the resolution in the stochastic space independent of the spatial resolution.

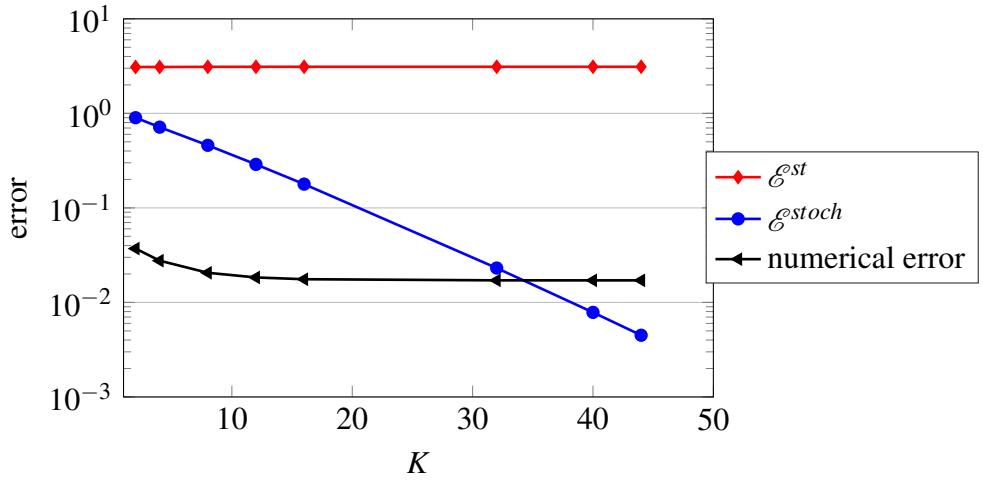



Figure 5.7: Error plot for the Burgers equation for a fixed mesh with 512 elements and DG polynomial degree $q = 2$.

## 5.2 A posteriori error analysis based on the SC method

In this section we derive an a posteriori error estimate for numerical approximations which rely on the SC method from Section 4.1.4. The content of this section led to the work [47].

For our problem of interest we now allow for random flux functions and consider an $N$-dimensional random space $\Xi \subset \mathbb{R}^N$, $N \in \mathbb{N}$. The spatial domain remains $D = [0,1]_{\text{per}}$ and the initial-value problem that we consider reads as follows.

$$\begin{cases} \partial_t u(t,x,y) + \partial_x F(u(t,x,y),y) = 0, & (t,x,y) \in (0,T) \times D \times \Xi, \\ u(0,x,y) = u^0(x,y), & (x,y) \in D \times \Xi. \end{cases} \tag{5.16}$$

### 5.2.1 SC on time-dependent physical meshes

Recalling the SC method from Section 4.1.4 we consider a set of nodal collocation points, denoted by $\{y_k\}_{k \in \mathscr{K}} \subset \Xi$. Using the collocation points as input parameters in (5.16) we obtain

the following deterministic initial-value problems

$$\begin{cases} \partial_t u(t,x,y_{\mathrm{k}}) + \partial_x F(u(t,x,y_{\mathrm{k}}),y_{\mathrm{k}}) = 0, & (t,x) \in (0,T)\times D, \\ u(0,x,y_{\mathrm{k}}) = u^0(x,y_{\mathrm{k}}), & x \in D, \end{cases} \tag{5.17}$$

for all $\mathrm{k} \in \mathscr{K}$. For the space and time discretization of (5.17) we use again the RKDG method described in Section 3.1 but now we want to allow for adaptive spatial mesh refinement. Therefore, we formulate the DG scheme on time-dependent spatial meshes. For the ease of presentation we neglect the dependence of the flux $F$, the spatial mesh and the DG spaces on the collocation points $\{y_{\mathrm{k}}\}_{\mathrm{k}\in\mathscr{K}}$.

We let $0 = x_0 < x_1 \ldots < x_{N_s} = 1$ be a quasi-uniform triangulation of $[0,1]$, which we denote by $\mathscr{T}$ and $0 = t_0 < t_1 < \ldots < t_{N_t} = T$ be a temporal decomposition of $[0,T]$. Furthermore, we identify $x_0 = x_{N_s}$ to account for the periodic boundary conditions. With each time-interval $(t_n, t_{n+1}]$ we associate a (possibly different) partition $\mathscr{T}_n$ and associated DG space

$$\mathscr{V}_{h,n}^q := \{u : D \to \mathbb{R}^m \mid u\big|_I \in \mathbb{P}_q(I,\mathbb{R}^m), \text{ for all } I \in \mathscr{T}_n\}.$$

With $\mathscr{L}_{\mathscr{V}_{h,n}^q}$ we denote the $L^2$-projection mapping into the DG space $\mathscr{V}_{h,n}^q$.

Following [30] we call the function $u_h$ a generalized semi-discrete DG approximation of (5.17) if it satisfies for $u_h^{-1} := \mathscr{L}_{\mathscr{V}_{h,0}^q} u^0$ the following equations. For every $n = 0,\ldots,N_t$, $u_h^n\big|_{[t_n,t_{n+1}]} \in C^1((t_n,t_{n+1});\mathscr{V}_{h,n}^q) \cap C^0([t_n,t_{n+1}];\mathscr{V}_{h,n}^q)))$,

$$\begin{cases} \sum\limits_{l=0}^{N_s-1} \int\limits_{x_l}^{x_{l+1}} \partial_t u_h^n \cdot \psi_h \,\mathrm{d}x &= \sum\limits_{l=0}^{N_s-1} \int\limits_{x_l}^{x_{l+1}} L_h^n(u_h^n)\cdot\psi_h \,\mathrm{d}x \quad \forall \psi_h \in \mathscr{V}_{h,n}^q \\ u_h^n(t_n) &= \mathscr{L}_{\mathscr{V}_{h,n}^q} u_h^{n-1}(t_n), \end{cases} \tag{GDG}$$

where $L_h^n : \mathscr{V}_{h,n}^q \to \mathscr{V}_{h,n}^q$ is defined by

$$\begin{aligned} \sum_{l=0}^{N_s-1} \int_{x_l}^{x_{l+1}} L_h^n(v)\cdot\psi_h \,\mathrm{d}x = &\sum_{l=0}^{N_s-1} \int_{x_l}^{x_{l+1}} F(v)\cdot\partial_x\psi_h \,\mathrm{d}x \\ &- \sum_{l=0}^{N_s-1} \hat{F}(v(x_l^-), v(x_l^+))\cdot[\![\,\psi_h\,]\!]_l, \quad \forall v,\psi_h \in \mathscr{V}_{h,n}^q. \end{aligned} \tag{5.18}$$

The numerical solution $u_h$ is defined through $u_h(0) := u_h^{-1}$ and $u_h\big|_{(t_n,t_{n+1}]} := u_h^n\big|_{(t_n,t_{n+1}]}$.

The $S$-stage time-marching algorithm for the initial-value problem (GDG) for given $n$-th time-iterate $u_h^n(t_n) \in \mathscr{V}_{h,n}^q$ reads as follows.

**Algorithm 3** `Runge-Kutta Time-Step on time-dependent meshes`

1: Set $u_h^{(0)} = u_h^n(t_n)$.
2: **for** $j = 1, \dots, S$ **do**
3: Compute: $u_h^{(j)} = \Lambda\Pi_h\Big(\sum_{l=0}^{j-1} \alpha_{jl} w_h^{jl}\Big), \quad w_h^{jl} = u_h^{(l)} + \frac{\beta_{jl}}{\alpha_{jl}} \Delta t_n L_h^n(u_h^{(l)})$.
4: **end for**
5: Set $u_h^n(t_n) = u_h^{(S)}$.

The spatial mesh is adapted after every iteration of Algorithm 3. We describe the local mesh adaption procedure in detail in Section 5.2.3.

### 5.2.2 Reconstruction and residuals

The reconstruction process for the SC approximation (4.19) is very similar to that of the SG method, which we already described in detail in Section 5.1.1. For the SG method we reconstructed the deterministic modes of the truncated gPC expansion, now we reconstruct the numerical approximation $\{u_h^n(\cdot, y_\mathrm{k})\}_{n=0}^{N_t}$ for every collocation point $\{y_\mathrm{k}\}_{\mathrm{k}\in\mathscr{K}}$ to a Lipschitz function in space and time. For simplicity we assume that all approximate solutions are interpolated onto a reference mesh $\mathscr{T}$, which is a common refinement of all meshes. With $\mathscr{V}_h^q$ we denote the DG space associated with $\mathscr{T}$, hence $u_h^n(\cdot, y_\mathrm{k}) \in \mathscr{V}_h^q$ for all $n = 0, \dots, N_t$ and for all $\mathrm{k} \in \mathscr{K}$.

The temporal and space-time reconstruction process from Section 5.1.1 provides us with a computable space-time reconstruction $\hat{u}^{st}(y_\mathrm{k}) := \hat{u}^{st}(\cdot, \cdot, y_\mathrm{k}) \in W_\infty^1((0,T); \mathscr{V}_h^{q+1} \cap C^0(D; \mathscr{U}))$ of the numerical approximation $\{u_h^n(\cdot, y_\mathrm{k})\}_{n=0}^{N_t}$, for every collocation point $y_\mathrm{k}$, $\mathrm{k} \in \mathscr{K}$.

This allows us to define the space-time residual as follows.

**Definition 5.16** (Space-time residual)**.**
*We call the function $\mathscr{R}^{st}(y_\mathrm{k}) := \mathscr{R}^{st}(\cdot, \cdot, y_\mathrm{k}) \in L^2((0,T) \times D; \mathbb{R}^m)$, defined by*

$$\mathscr{R}^{st}(t, x, y_\mathrm{k}) := \partial_t \hat{u}^{st}(t, x, y_\mathrm{k}) + \partial_x F(\hat{u}^{st}(t, x, y_\mathrm{k}), y_\mathrm{k}), \tag{5.19}$$

*the space-time residual associated with the collocation point $y_\mathrm{k}$, for all $\mathrm{k} \in \mathscr{K}$.*

In the next step we expand the space-time reconstruction into the corresponding random basis, i.e. in the Lagrange basis, to obtain the so-called space-time-stochastic reconstruction.

**Definition 5.17** (Space-time-stochastic reconstruction for SC)**.**
*We call the function $\hat{u}^{sts} \in \mathscr{P}_K(\Xi) \otimes W_\infty^1((0,T); \mathscr{V}_h^{q+1} \cap C^0(D; \mathscr{U}))$ defined by*

$$\hat{u}^{sts}(t, x, y) := \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(t, x, y_\mathrm{k}) l_\mathrm{k}(y),$$

*the space-time-stochastic reconstruction of the numerical approximation* (4.19).

Similar to the space-time reconstruction, we may plug $\hat{u}^{sts}$ into the random conservation law (5.16) to obtain the so called space-time-stochastic residual.

**Definition 5.18** (Space-time-stochastic residual for SC)**.**
*We define the space-time-stochastic residual* $\mathscr{R}^{sts} \in L^2_w(\Xi; L^2((0,T)\times D;\mathbb{R}^m))$ *by*

$$\mathscr{R}^{sts}(t,x,y) := \partial_t \hat{u}^{sts}(t,x,y) + \partial_x F(\hat{u}^{sts}(t,x,y),y). \tag{5.20}$$

We now have all ingredients together to state the following main a posteriori error estimate that can be directly derived from Theorem 2.15.

**Theorem 5.19** (A posteriori error bound for the numerical solution)**.**
*Let u be a random entropy admissible solution of* (5.16)*. Provided the reconstruction* $\hat{u}^{sts}$ *from Definition 5.17 only take values in a compact, convex set* $\mathscr{C} \subset \mathscr{U}$*, the difference between u and the numerical solution* $u_h^n$ *from* (4.19) *satisfies*

$$\begin{aligned}\|u(t_n,\cdot,\cdot) - u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} &\le 2\|\hat{u}^{sts}(t_n,\cdot,\cdot) - u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \\ &\quad + 2\int_{\tilde{\Xi}} \Big[\Big(C^{-1}_{\underline{\eta}}(y)\Big(\mathscr{E}^{sts}(t_n,y) + C_{\overline{\eta}}(y)\mathscr{E}_0^{sts}(y)\Big)\Big) \\ &\quad \times \exp\Big(\int_0^{t_n} \frac{C_{\overline{\eta}}(y)C_{\overline{F}}(y)\|\partial_x \hat{u}^{sts}(t,\cdot,y)\|_{L^\infty(D)} + C^2_{\overline{\eta}}(y)}{C_{\underline{\eta}}(y)}\ \mathrm{dt}\Big)\Big]\ w(y)\mathrm{dy},\end{aligned}$$

*for all* $n = 0,\dots,N_t$ *and for any* $\tilde{\mathbb{P}}$*-measurable set* $\tilde{\Xi} \subseteq \Xi$*. Here*

$$\mathscr{E}^{sts}(t_n,y) := \|\mathscr{R}^{sts}(\cdot,\cdot,y)\|^2_{L^2((0,t_n)\times D)}, \tag{5.21}$$

$$\mathscr{E}_0^{sts}(y) := \|u^0(\cdot,y) - \hat{u}^{sts}(0,\cdot,y)\|^2_{L^2(D)}. \tag{5.22}$$

Similar to the SG method we want to derive a splitting of the space-time-stochastic residual into a space-time (deterministic) and stochastic part. The decomposition of the residual is shown in the following lemma.

**Lemma 5.20** (Splitting of the space-time-stochastic residual for SC)**.**
*The space-time-stochastic residual* $\mathscr{R}^{sts}$ *admits the decomposition*

$$\mathscr{R}^{sts} = \mathscr{R}^{st} + \mathscr{R}^{stoch}, \tag{5.23}$$

*with*

$$\mathscr{R}^{st} := \sum_{\mathrm{k}\in\mathscr{K}} \mathscr{R}^{st}(y_\mathrm{k}) l_\mathrm{k} \text{ and} \tag{5.24}$$

$$\mathscr{R}^{stoch} := \partial_x F\Big(\sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_\mathrm{k}) l_\mathrm{k}, \cdot\Big) - \sum_{\mathrm{k}\in\mathscr{K}} \partial_x F(\hat{u}^{st}(y_\mathrm{k}), y_\mathrm{k}) l_\mathrm{k}. \tag{5.25}$$

*$\mathscr{R}^{st}$ and $\mathscr{R}^{stoch}$ are called the space-time (deterministic) and stochastic residual.*

*Proof.* For every collocation point $y_{\mathrm{k}}$, $\mathrm{k} \in \mathscr{K}$, we compute the space-time reconstruction $\hat{u}^{st}(\cdot,\cdot,y_{\mathrm{k}})$ which fulfills

$$\mathscr{R}^{st}(y_{\mathrm{k}}) = \partial_t \hat{u}^{st}(y_{\mathrm{k}}) + \partial_x F(\hat{u}^{st}(y_{\mathrm{k}}), y_{\mathrm{k}}). \tag{5.26}$$

Moreover, we know from (5.47) that the space-time-stochastic reconstruction $\hat{u}^{sts} = \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(t,x,y_{\mathrm{k}}) l_{\mathrm{k}}(y)$ satisfies the relation

$$\mathscr{R}^{sts} = \partial_t \hat{u}^{sts} + \partial_x F(\hat{u}^{sts},\cdot) = \partial_t \Big( \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} \Big) + \partial_x F\Big( \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}}, \cdot \Big). \tag{5.27}$$

Multiplying (5.26) by $l_{\mathrm{k}}$ and summing over $\mathrm{k} \in \mathscr{K}$ yields

$$\sum_{\mathrm{k}\in\mathscr{K}} \mathscr{R}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} = \sum_{\mathrm{k}\in\mathscr{K}} \partial_t \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} + \sum_{\mathrm{k}\in\mathscr{K}} \partial_x F(\hat{u}^{st}(y_{\mathrm{k}}), y_{\mathrm{k}}) l_{\mathrm{k}}. \tag{5.28}$$

Inserting (5.28) into (5.27) yields

$$\begin{aligned}
\mathscr{R}^{sts} &= \partial_t \Big( \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} \Big) + \partial_x F\Big( \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}}, \cdot \Big) \\
&\quad + \sum_{\mathrm{k}\in\mathscr{K}} \mathscr{R}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} - \Big( \sum_{\mathrm{k}\in\mathscr{K}} \partial_t \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} + \sum_{\mathrm{k}\in\mathscr{K}} \partial_x F(\hat{u}^{st}(y_{\mathrm{k}}), y_{\mathrm{k}}) l_{\mathrm{k}} \Big) \\
&= \sum_{\mathrm{k}\in\mathscr{K}} \mathscr{R}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}} + \Big( \partial_x F\Big( \sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(y_{\mathrm{k}}) l_{\mathrm{k}}, \cdot \Big) - \sum_{\mathrm{k}\in\mathscr{K}} \partial_x F(\hat{u}^{st}(y_{\mathrm{k}}), y_{\mathrm{k}}) l_{\mathrm{k}} \Big) \\
&= \mathscr{R}^{st} + \mathscr{R}^{stoch}.
\end{aligned}$$

□

**Remark 5.21.**
*The residual $\mathscr{R}^{st}$ in* (5.24) *interpolates spatio-temporal residuals and contains information about the discretization error in physical space, i.e. the space-time resolution of* (GDG) *using the RKDG method. In contrast to $\mathscr{R}^{st}$, the stochastic residual $\mathscr{R}^{stoch}$ in* (5.25) *indicates the quality of the interpolation in stochastic space.*

To simplify Theorem 5.19 let us assume that the eigenvalues of the Hessian $H_u\eta(u,y)$ are bounded from above and below by positive numbers, for any $u \in \mathscr{C}$ uniformly in $\Xi$. We let $C_{\overline{\eta}} := \operatorname{ess\,sup}_{y\in\Xi} C_{\overline{\eta}}(\Xi) < \infty$, $C_{\underline{\eta}} := \operatorname{ess\,inf}_{y\in\Xi} C_{\overline{\eta}}(y) > 0$ and $C_{\overline{F}} := \operatorname{ess\,sup}_{y\in\Xi} C_{\overline{F}}(\Xi) < \infty$. In our numerical examples, where we consider the compressible Euler equations, the dependence of the flux function $F$ and $\eta$ on $y$ is explicitly known and we can compute the constants $C_{\underline{\eta}}, C_{\overline{\eta}}, C_{\overline{F}}$ numerically.

**Corollary 5.22** (A posteriori error bound with error splitting and simplified bounds)**.**
*Let $u$ be a random entropy admissible solution of* (5.16)*. Then, the difference between $u$ and the numerical solution $u_h^n$ from* (4.19) *satisfies*

$$\begin{aligned}\|u(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))} \leq\ & 2\|\hat{u}^{sts}(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))}\\ &+2C_{\underline{\eta}}^{-1}\left(2\mathscr{E}^{st}(t_n)+2\mathscr{E}^{stoch}(t_n)+C_{\overline{\eta}}\mathscr{E}_0^{sts}\right)\\ &\times\exp\left(C_{\underline{\eta}}^{-1}\int_0^{t_n}\left(C_{\overline{\eta}}C_{\overline{F}}\|\partial_x\hat{u}^{sts}(t,\cdot,y)\|_{L^\infty_w(\tilde{\Xi};L^\infty(D))}+C_{\overline{\eta}}^2\right)\,\mathrm{dt}\right)\end{aligned}$$

*for $n=0,\dots,N_t$ and for all $\tilde{\mathbb{P}}$-measurable sets $\tilde{\Xi}\subseteq\Xi$. Here,*

$$\mathscr{E}^{st}(t_n) := \|\mathscr{R}^{st}\|^2_{L^2_w(\tilde{\Xi};L^2((0,t_n)\times D))}, \tag{5.29}$$

$$\mathscr{E}^{stoch}(t_n) := \|\mathscr{R}^{stoch}\|^2_{L^2_w(\tilde{\Xi};L^2((0,t_n)\times D))}, \tag{5.30}$$

$$\mathscr{E}_0^{sts} := \|u^0(\cdot,\cdot)-\hat{u}^{sts}(0,\cdot,\cdot)\|^2_{L^2_w(\tilde{\Xi};L^2(D))}. \tag{5.31}$$

**Remark 5.23.** *1. In order for the upcoming space-stochastic adaptive algorithm based on $\mathscr{E}^{st}$, $\mathscr{E}^{stoch}$ to be efficient, we need $\mathscr{E}^{st}$ to depend solely on the spatio-temporal discretization and to be independent of the stochastic discretization. Similarly, we need $\mathscr{E}^{stoch}$ to decay when the stochastic resolution is increased but to be independent of the spatio-temporal discretization. In Remark 5.24 we prove that $\mathscr{E}^{st}$ is indeed unaffected by stochastic refinement. Remark 5.25 discusses the independence of the stochastic residual from the spatial discretization.*

2. *The scaling properties of $\mathscr{E}^{st}$ , resp. $\mathscr{R}^{st}(y_k)$, were studied in [28]. Currently we are not able to prove any of the scaling properties of $\mathscr{E}^{stoch}$ w.r.t. to K and the number of Multi-Elements. However, our numerical experiments show that $\mathscr{E}^{stoch}$ scales as desired, i.e. $\mathscr{E}^{stoch}$ shows the same qualitative behavior as the stochastic interpolation error of the exact solution.*

3. *As described in Remark 5.13, $\mathscr{R}^{st}$ scales with $\frac{1}{h}$ in the vicinity of shocks and contact discontinuities, i.e., it blows up under spatial mesh refinement in these areas, although the numerical solution converges towards the exact solution. Hence, we only have reliable a posteriori error control for smooth solutions of* (5.16)*. However, as $\mathscr{R}^{st}$ precisely captures the positions of rarefaction waves, contact discontinuities and shocks we use $\mathscr{R}^{st}$ and $\mathscr{R}^{stoch}$, resp. $\mathscr{E}^{st}$ and $\mathscr{E}^{stoch}$, as local indicators for our adaptive mesh refinement algorithms described in Section 5.2.3.*

**Remark 5.24** (Uniformity of the deterministic residual in $\Xi$)**.**
*As noted above, the collocation points $y_k$ are chosen to be the zeros of the corresponding or-*

*thogonal polynomial depending on the distribution of y. The deterministic residual $\mathscr{R}^{st}$ from* (5.24) *consists of Lagrange polynomials associated with the corresponding collocation points, thus Gaussian quadrature in $\Xi$ yields*

$$\begin{aligned}\mathscr{E}^{st}(T) = \|\mathscr{R}^{st}\|^2_{L^2_w(\Xi;L^2((0,T)\times D))} &= \sum_{\mathrm{k}\in\mathscr{K}} \|\mathscr{R}^{st}(y_\mathrm{k})\|^2_{L^2((0,T)\times D))} w_\mathrm{k} \\ &\leq \max_{\mathrm{k}\in\mathscr{K}} \|\mathscr{R}^{st}(y_\mathrm{k})\|^2_{L^2((0,T)\times D)}.\end{aligned}$$

*Hence, $\mathscr{E}^{st}$ inherits the convergence order of $\mathscr{R}^{st}(y_\mathrm{k})$ and is thus independent of the stochastic discretization.*

**Remark 5.25** (Decay of stochastic residual)**.**
*Remark 5.24 shows that the deterministic residual $\mathscr{E}^{st}$ is indeed unaffected by stochastic discretization. To prove that the stochastic residual $\mathscr{R}^{stoch}$ is independent of the spatial discretization is not straightforward and can only be shown if we assume that the stochastic regularity does not depend on the spatial discretization. This is in general not true, see for example [82, Sec. 4.3], where the authors prove that in case of a wave equation with discontinuous wave speed, the stochastic regularity depends on the spatial mesh width h. If we define the SC operator via $\mathscr{I}_\mathscr{K}(u) := \sum_{\mathrm{k}\in\mathscr{K}} u(y_\mathrm{k}) l_\mathrm{k}$ we are able to write the stochastic residual $\mathscr{R}^{stoch}$ from* (5.25) *as follows*

$$\begin{aligned}\mathscr{R}^{stoch}(t,x,\cdot) &= \Big(\partial_x F\Big(\sum_{\mathrm{k}\in\mathscr{K}} \hat{u}^{st}(t,x,y_\mathrm{k}) l_\mathrm{k}(\cdot),\cdot\Big) - \sum_{\mathrm{k}\in\mathscr{K}} \partial_x F(\hat{u}^{st}(t,x,y_\mathrm{k}),y_\mathrm{k}) l_\mathrm{k}(\cdot)\Big) \\ &= \partial_x F(\hat{u}^{sts}(t,x,\cdot),\cdot) - \mathscr{I}_\mathscr{K}\Big(\partial_x F(\hat{u}^{sts}(t,x,\cdot),\cdot)\Big),\end{aligned}$$

*for a.e. $(t,x)\in(0,T)\times D$. Hence, $\mathscr{R}^{stoch}$ corresponds to the stochastic interpolation error, when interpolating the spatial derivative of the flux function $F$. As long as the regularity of the mapping $y\mapsto \partial_x F(\hat{u}^{sts}(\cdot,\cdot,y),y)$ does not depend on the spatial mesh width h, $\mathscr{R}^{stoch}$ would decay independently of h. For smooth solutions we expect that this is indeed the case, cf. [82, Remark 6]. Our numerical experiments confirm this assertion and show that the convergence of $\mathscr{E}^{sc}$ is unaffected by the spatial resolution.*

### 5.2.3 Adaptive Algorithms

The splitting of the space-time-stochastic residual into a deterministic and a stochastic residual helps us in developing adaptive numerical schemes where we use the residuals as local error indicators for spatial and stochastic mesh refinement. We describe the deterministic spatially adaptive algorithm, which we use to solve (GDG) for every collocation point $y_\mathrm{k}$, $\mathrm{k}\in\mathscr{K}$. We

slightly abuse the notation from (5.29) and write for every physical cell $I \in \mathscr{T}_n$, $\mathscr{E}_{\mathrm{k}}^{st}(t_n, t_{n+1}, I) := \|\mathscr{R}^{st}(y_{\mathrm{k}})\|_{L^2((t_n,t_{n+1})\times I)}$, which is the cell-wise indicator for the spatial refinement in $D$.

The local physical mesh refinement is achieved by uniformly dividing one cell into two new children cells or merging two cells into one parent cell. To mark elements for refinement we compute the deterministic residual $\mathscr{E}_{\mathrm{k}}^{st}(t_n, t_{n+1}, I)$ on every cell $I \in \mathscr{T}_n$ and based on the residual we mark a fixed fraction of the cells for refinement. To coarsen the mesh, we can only merge cells that have the same parent element and both siblings are marked for coarsening. For coarsening we also choose a fixed fraction of all elements according to the local residual $\mathscr{E}_{\mathrm{k}}^{st}(t_n, t_{n+1}, I)$, cf. [60]. Additionally, each cell is augmented with a variable denoting its current mesh-level which is initially zero. We fix a maximum mesh-level $L \in \mathbb{N}$, to restrict the fineness of the adaptive mesh. The algorithm reads as follows:

**Algorithm 4** `Deterministic` $h$`-adaptive algorithm`

**Input:** final time $T$, max mesh-level $L$, initial mesh $\mathscr{T}_0$

1: Compute $u_h^{n+1}$ on the current mesh $\mathscr{T}_n$ using Algorithm 3
2: Compute $\mathscr{E}_{\mathrm{k}}^{st}(t_n, t_{n+1}, I)$ for $I \in \mathscr{T}_n$ and mark a fixed fraction of the elements for refinement and coarsening
   a: Refinement: If the cell's mesh-level is $L$ do nothing. Else divide it uniformly into two new cells and increase the two new cells' mesh-level by one
   b: Coarsening: If the cell's mesh-level is zero do nothing. Else check if its sibling is marked for coarsening. If yes merge the two cells into one and decrease its mesh-level by one
3: Project $u_h^{n+1}$ onto the new mesh $\mathscr{T}_{n+1}$ using the $\mathscr{L}_{\mathscr{V}_{h,n+1}^q}$-projection
4: If $t_{n+1} < T$ go to step 1

**Remark 5.26.**

*After every projection step in line three of Algorithm 4 we apply the TVBM slope limiter* $\Lambda\Pi_h$.

In the numerical experiments in Section 5.2.4 we observed that setting the refinement and coarsening fractions to 1% and 20% provided the best error reduction. We restrict the maximal refinement level to $L = 3$, since the time step size is linked to the size of the smallest spatial cell (confer (3.5)), and allowing for smaller cells would make the time steps infeasible small. This limitation could be overcome by local time-stepping.

Next, we describe the stochastic adaptive algorithm, where we use the Multi-Element method from Section 4.1.2 in combination with the SC method. We call this method ME-SC. The idea is the compute the stochastic residual $\mathscr{E}^{stoch}$ on every ME and use this information as local indicator for stochastic refinement.

**Algorithm 5** `Stochastic` $N_\Xi$`-Adaptive Algorithm`

**Input:** initial number of Multi-Elements $M_\Xi$, max no. of Multi-Elements $N_\Xi$, $K+1$ number of collocation points in each stochastic dimension

1: For every Multi-Element $\mathrm{D_m}$ compute $(K+1)^N$ numerical samples using Algorithm 4
2: Compute $\mathscr{E}^{stoch}(T)$ on every Multi-Element $\mathrm{D_m}$ and uniformly subdivide the Multi-Element with the biggest residual, set $M_\Xi := M_\Xi + (2^N - 1)$
3: If $M_\Xi < N_\Xi$ compute $M$ samples on every new Multi-Element and go to 2

### 5.2.4 Numerical experiments

In this section we present various numerical examples concerning the scaling properties of the residuals and the performance of the adaptive algorithms. As deterministic numerical solver we use the RKDG Code FLEXI [63]. The DG polynomial degrees are always one or two and for the time-stepping we use the low storage SSP RK-method of order three as in [65]. The time-reconstruction is also of order three. As numerical fluxes we choose either the Lax-Wendroff numerical flux

$$\hat{F}(u,v) := F(w(u,v)), \quad w(u,v) := \frac{1}{2}\Big((u+v) + \frac{\Delta t}{h}(F(v)-F(u))\Big), \tag{5.32}$$

or the Lax-Friedrichs numerical flux

$$\hat{F}(u,v) := \frac{1}{2}\Big(F(u)+F(v)\Big) + \lambda(v-u). \tag{5.33}$$

In our example, the uncertainty is uniformly distributed. Therefore, we use the zeros of the Gauß–Legendre polynomials as collocation points. Computing $\mathscr{E}^{st}, \mathscr{E}^{stoch}$ requires computing integrals, we approximate them via Gauß–Legendre quadrature where we use seven points in time, ten points in physical space and ten points in random space, except for $K$-refinement, where for the global interpolation the number of quadrature points in random space is $2K$.

In the following experiments we consider as instance of (5.16) the one-dimensional compressible Euler equations for the flow of an ideal gas, which are given by

$$\begin{aligned} \partial_t \rho + \partial_x m &= 0, \\ \partial_t m + \partial_x \left(\frac{m^2}{\rho} + p\right) &= 0, \\ \partial_t E + \partial_x \left((E+p)\frac{m}{\rho}\right) &= 0, \end{aligned} \tag{5.34}$$

where $\rho$ describes the mass density, $m$ the momentum and $E$ the energy of the gas. The constitutive law for pressure $p$ reads

$$p = (\gamma - 1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right),$$

with the adiabatic constant $\gamma = 1.4$ if not specified otherwise. In the following figures we refer to the quantity $\|m(T,\cdot,\cdot) - m_h^{N_t}(\cdot,\cdot)\|_{L^2_w(\Xi;L^2(D))}$ at final computational time $T$ as numerical error, unless otherwise stated. We also plot the residuals $\mathscr{E}^{st}(T)$ and $\mathscr{E}^{stoch}(T)$ as in (5.29) and (5.30) from the momentum equation.

**Remark 5.27.**

*Due to the structure of the flux Jacobian for the Euler equations* (5.34)*,*

$$\mathrm{D}F(u) = \begin{pmatrix} 0 & 1 & 0 \\ -0.5(\gamma-3)\frac{m^2}{\rho^2} & (3-\gamma)\frac{m}{\rho} & \gamma-1 \\ -\gamma\frac{Em}{\rho} + (\gamma-1)\frac{m^3}{\rho^3} & \gamma\frac{E}{\rho} - \frac{3}{2}(\gamma-1)\frac{m^2}{\rho^2} & \gamma\frac{m}{\rho} \end{pmatrix},$$

*the first component of the stochastic residual $\mathscr{R}^{stoch}$ from* (5.25) *vanishes when considering the Euler equations without source term. We therefore use the residuals for the momentum and the energy balance as indicators for our space-stochastic mesh refinements.*

**Deterministic adaptivity: Sod Shock Tube Problem**

In this numerical experiment we apply the adaptive spatial mesh refinement from Algorithm 4 to the Sod shock tube problem. The Riemann data for this problem is given by

$$\begin{aligned} \rho(t=0,x,y) &= \begin{cases} 1, & x < 0.5 \\ 0.125, & x \geq 0.5, \end{cases} \\ m(t=0,x,y) &= 0, \\ E(t=0,x,y) &= \begin{cases} 2.5, & x < 0.5, \\ 0.25, & x \geq 0.5. \end{cases} \end{aligned} \tag{5.35}$$

The numerical solution is computed on the domain $D = [0,1]$ up to $T = 0.2$ using the Lax-Friedrichs flux (5.33) and a DG polynomial degree of two. In this example we use exact boundary conditions. In Figure 5.8(a) we compare the $L^1(D)$- and $L^2(D)$-error at time $T$ between the numerical solution and the exact solution obtained with an exact Riemann solver [4]. We can see that for the same number of spatial cells $N_s$, the numerical error obtained with the adaptive numerical algorithm is smaller than for the uniform mesh refinement. The adaptive algorithm is also computationally more efficient than the uniform algorithm, which can be seen in the error vs. cpu time plot in Figure 5.8(b).

**Remark 5.28.**

*As discussed in Remark 5.23, $\mathscr{R}^{st}$ scales with $\frac{1}{h}$ in the vicinity of shocks and contact discontinuities, i.e., it blows up under spatial mesh refinement in these areas. Thus, if we view the residual*

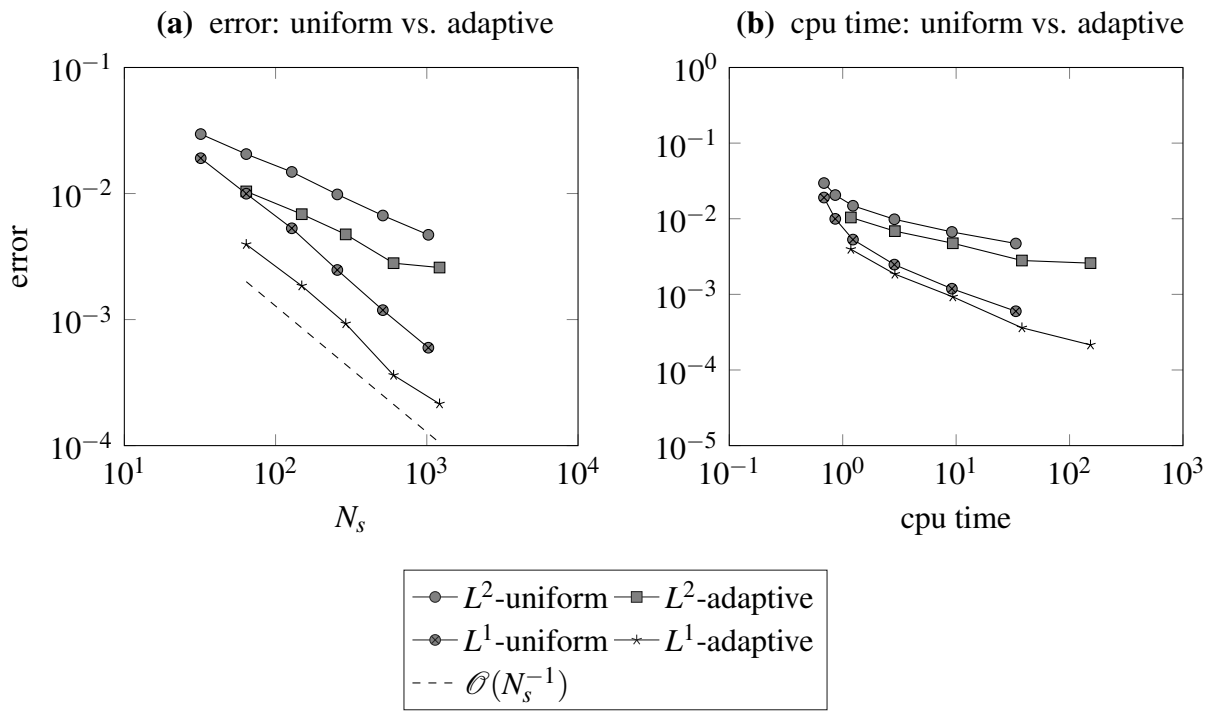



Figure 5.8: Error plot for the deterministic Sod shock tube problem. Example 5.35.

*as an error indicator, it severely over-estimates the error so that it is to be called "inefficient" in these areas, according to the nomenclature of e.g. [100]. From the point of view of mesh adaptation however, refinement based on $\mathscr{R}^{st}$ leads to a reasonable refinement strategy that yields a considerable improvement in error decay compared to uniform mesh refinement (cf. Figure 5.8). In particular, $\mathscr{R}^{st}$ precisely captures the positions of rarefaction waves, contact discontinuities and shocks.*

*Over-estimating the error at discontinuities leads to maximal refinement at discontinuities and some refinement strategies for hyperbolic conservation laws suggest a maximal refinement close to shocks [88].*

### A one-dimensional random space, $K$-refinement

We now consider the following exact solution.

$$\begin{pmatrix} \rho(t,x,y) \\ m(t,x,y) \\ E(t,x,y) \end{pmatrix} = \begin{pmatrix} 2+0.1\cos(4\pi(x-yt)) \\ \Big(2+0.1\cos(4\pi(x-yt))\Big)\Big(1+0.1\sin(4\pi(x-yt))\Big) \\ \Big(2+0.1\cos(4\pi(x-yt))\Big)^2 \end{pmatrix}. \tag{5.36}$$

The numerical solution is computed on $D = [0,1]_{\text{per}}$ up to $T = 0.2$, the uncertainty stems from a uniform distribution, i.e. $\xi \sim \mathscr{U}(0,8)$. We consider three different spatial meshes consisting of $N_s = 8, 32, 512$ elements, DG polynomial degrees of $q = 1, 2$ and we use the Lax-Wendroff numerical flux (5.32). In this numerical example we globally approximate the function (5.36) in $\Xi$, i.e. we increase the polynomial degree $K$ and consider one ME.

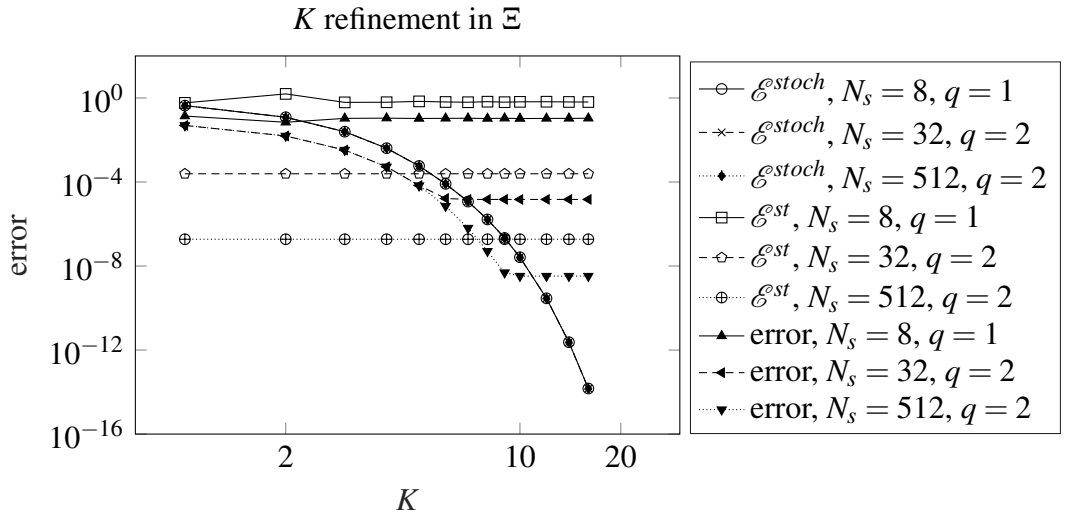



Figure 5.9: Error plot for stochastic smooth problem (5.36).

Figure 5.9 shows the behavior of the error and the spatial, resp. stochastic residual, when we globally interpolate the smooth function (5.36). We see that the stochastic residual $\mathscr{E}^{stoch}$ exhibits spectral convergence. Also the numerical error exhibits spectral convergence until it starts to stagnate because of the spatial resolution error. This is the correct behavior of the stochastic residual as we are globally increasing the polynomial degree in the random space and, therefore, expect spectral convergence with increasing polynomial degree. We also observe that the exponential convergence of $\mathscr{E}^{stoch}$ is not altered by a finer or coarser space discretization, even for the very coarse discretization consisting of eight spatial cells and DG polynomial degree of one. Moreover, the deterministic residual $\mathscr{E}^{st}$ is unaffected by the increasing resolution in the random space, which we expect from the residual's splitting into a space-time and a stochastic part.

**Mesh refinement in $\Xi$ and random flux function**

In this example we examine the scaling properties of $\mathscr{E}^{stoch}$ under mesh refinements for a two-dimensional random space $\Xi \subset \mathbb{R}^2$. We consider the same smooth function as before,

$$\begin{pmatrix} \rho(t,x,y_1) \\ m(t,x,y_1) \\ E(t,x,y_1) \end{pmatrix} = \begin{pmatrix} 2+0.1\cos(4\pi(x-y_1t)) \\ \Big(2+0.1\cos(4\pi(x-y_1t))\Big)\Big(1+0.1\sin(4\pi(x-y_1t))\Big) \\ \Big(2+0.1\cos(4\pi(x-y_1t))\Big)^2 \end{pmatrix}. \tag{5.37}$$

with $\xi_1 \sim \mathscr{U}(0,8)$. Moreover, we consider a random adiabatic constant. We assume that $\gamma = \xi_2 \sim \mathscr{U}(1.4,1.6)$ and thus the flux function is also random. The randomness of the adiabatic-constant corresponds to considering a gas mixture of uncertain composition. The numerical solution is computed on $D = [0,1]_{\text{per}}$ up to $T = 0.2$. We consider a fixed spatial mesh consisting of $N_s = 32$ elements. For the ME-SC method we perform a linear and a quadratic interpolation, i.e. $K \in \{1,2\}$.

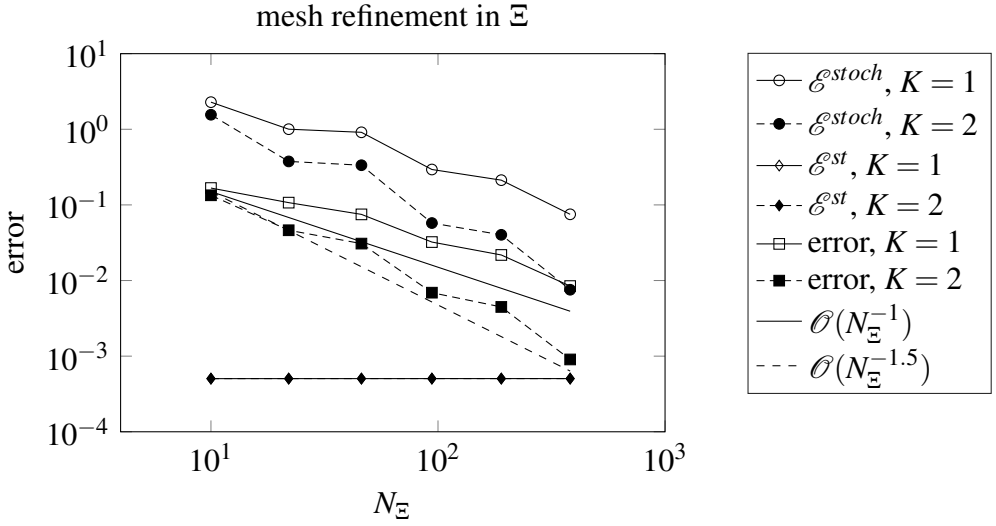



Figure 5.10: Error plot for stochastic smooth problem (5.37).

Figure 5.10 illustrates the behavior of the stochastic residual $\mathscr{E}^{stoch}$, when we consider a local interpolation, i.e., when we consider the ME method from Section 4.1.2. We observe that for a local linear and quadratic interpolation, i.e. $K \in \{1,2\}$, the stochastic residual converges approximately with the expected rate of convergence, which is $(K+1)/2$, cf. [102]. Like for the $K$-refinement in the previous section, the deterministic residual $\mathscr{E}^{st}$ stays constant, when we increase the number of MEs, i.e. it is independent of the stochastic discretization.

**Stochastic adaptivity: stochastic problem with discontinuous solution**

We apply the stochastic adaptive Algorithm 5 without spatial adaptivity to a solution which has a discontinuity in the random variable and compare the results with uniform space-stochastic mesh refinements. We therefore consider the following discontinuous function,

$$\begin{pmatrix} \rho(t,x,y_1,y_2) \\ m(t,x,y_1,y_2) \\ E(t,x,y_1,y_2) \end{pmatrix} = \begin{pmatrix} 1+A(y_1,y_2)\cos(4\pi(x-y_1t)) \\ \Big(1+A(y_1,y_2)\cos(4\pi(x-y_1t))\Big)\Big(1+0.1\sin(4\pi(x-y_1t))\Big) \\ \Big(1+A(y_1,y_2)\cos(4\pi(x-y_1t))\Big)^2 \end{pmatrix}, \quad (5.38)$$

where

$$A(y_1,y_2) = \begin{cases} 0.1, \text{ if } y_1^2+y_2^2 \leq 0.5^2 \\ 0.2, \text{ else .} \end{cases}$$

is a discontinuous amplitude. For the spatial domain $D=[0,1]_{\text{per}}$ we use $N_s=32$ elements and a DG polynomial degree of two. The solution is computed up to $T=0.2$ using the Lax-Wendroff numerical flux (5.32) and for the uncertainty we assume that $\xi_1,\xi_2 \sim \mathscr{U}(0,1)$. For the ME-SC method we consider a linear interpolant, i.e. $K=1$.

In Figure 5.11(a) we plot the error and the spatial resp. stochastic residual versus the number of MEs and in Figure 5.11(b) we show the error of the uniform and adaptive method versus cpu time. In Figure 5.11(a) we can observe that for the uniform stochastic refinement, both the error and the stochastic residual $\mathscr{E}^{stoch}$ converge with a rate of approximately $1/4$. This is in accordance with what we expect when interpolating a two-dimensional discontinuous function. For the adaptive refinement the error and the residual exhibit a rate of convergence of approximately $1/2$. The advantage of the stochastic adaptive algorithm is also reflected in Figure 5.11(b), where we reach an error reduction in significantly less time compared to uniform refinement.

**Space-stochastic adaptivity: an uncertain Riemann problem**

Finally, we assess the efficiency of the space-stochastic adaptive algorithm by considering a random Riemann Problem. The initial data for this problem reads as follows

$$\begin{aligned} \rho(t=0,x,y) &= 1 \\ m(t=0,x,y) &= \begin{cases} y_1, & x \leq 0.5 \\ y_2, & x > 0.5 \end{cases} \\ p(t=0,x,y) &= 1, \end{aligned} \quad (5.39)$$

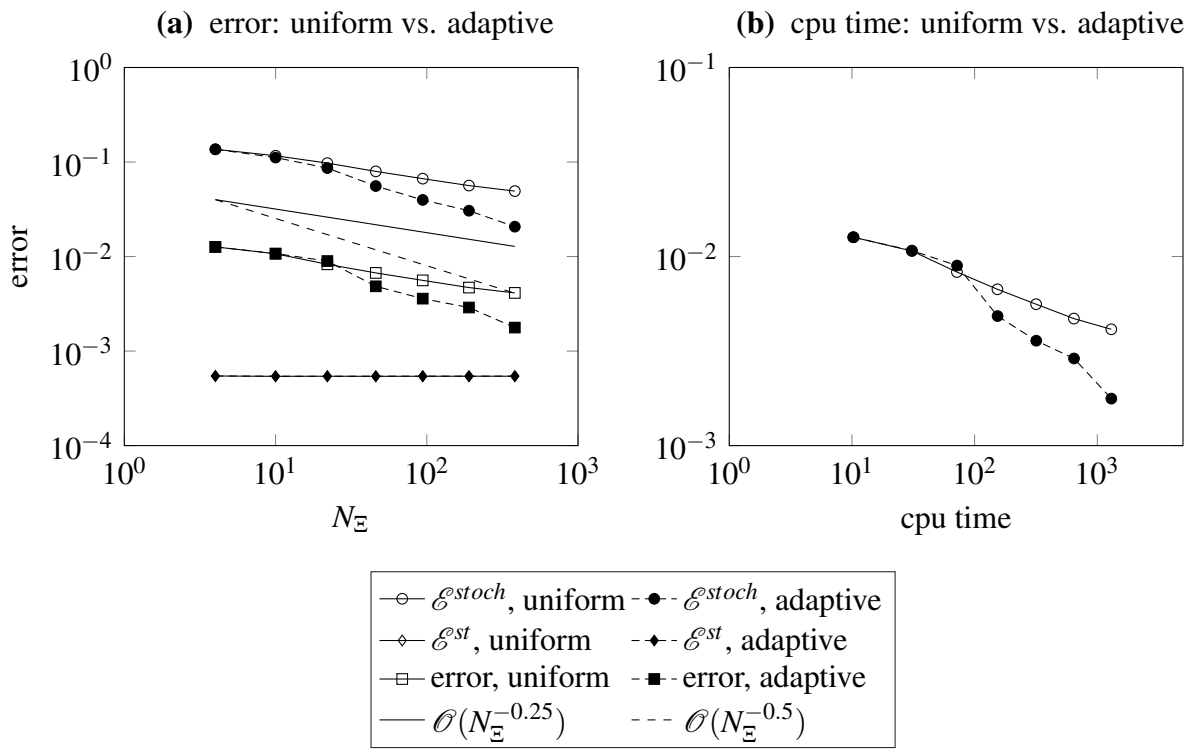



Figure 5.11: Error plot for discontinuous stochastic problem (5.38).

where $\xi_1, \xi_2 \sim \mathscr{U}(-1,1)$ and $D = [0,1]$. We compare the space-stochastic adaptive Algorithm 5 with uniform refinement, both in physical and random space. For this problem we use the Lax-Friedrichs numerical flux (5.33) and for the uniform spatial mesh we consider $N_s = 512$ spatial elements. We prescribe exact boundary conditions. For the adaptive algorithm we always start on a spatial mesh consisting of $N_s = 128$ elements. The DG polynomial degree is two and we consider a linear interpolation in the random space, i.e. $K = 1$. The solution is computed up to $T = 0.2$. The error is measured in the expected value rather than the $L^2_w(\Xi; L^2(D))$-norm. Note that we do not have an exact solution at hand for this problem, but due to Jensen's inequality,

$$\begin{aligned} \|\mathbb{E}\big(u(T,\cdot,\cdot)\big) - \mathbb{E}\big(u_h^{N_t}(\cdot,\cdot)\big)\|^2_{L^2(D)} &\leq \mathbb{E}\big(\|u(T,\cdot,\cdot) - u_h^{N_t}(\cdot,\cdot)\|^2_{L^2(D)}\big) \\ &= \|u(T,\cdot,\cdot) - u_h^{N_t}(\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))}. \end{aligned} \tag{5.40}$$

The reference expectation $\mathbb{E}\big(u(T,\cdot,\cdot)\big)$ is computed using a Monte Carlo method with an exact Riemann solver with 500000 samples. In Figure 5.12(a) we show the numerical error as in (5.40) and we also consider the error $\|\mathbb{E}\big(u(T,\cdot,\cdot)\big) - \mathbb{E}\big(u_h^{N_t}(\cdot,\cdot)\big)\|_{L^1(D)}$ for an increasing number of MEs, i.e. for increasing $N_\Xi$. We can see that the adaptive algorithm decreases the error considerably faster than the uniform refinement. This is also depicted in the cpu time vs. error plot (Figure 5.12(b)), where we can see that the adaptive algorithm reaches an absolute error in significantly less computational time than the uniform algorithm. This demonstrates, in particular,

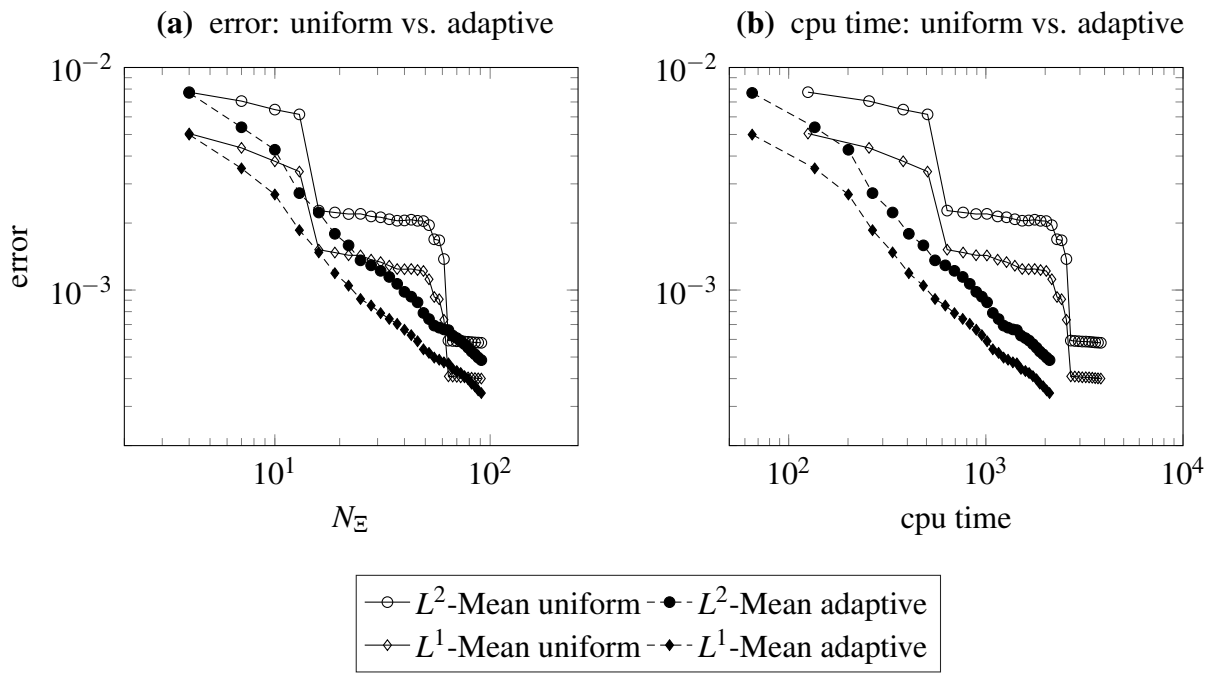



Figure 5.12: Error plot uncertain Riemann problem (5.39).

the efficiency of our proposed method.

## 5.3 A posteriori error analysis based on the NISP method

As last application of our a posteriori error analysis framework we derive an a posteriori error estimator for numerical approximations which rely on the NISP method. The content of this section is published in [48].

Since the reconstruction procedure and the main a posteriori error estimate are very similar to that of the SC method from Section 5.2, we only present the notion of reconstruction in the context of NISP and the corresponding splitting of the residual. We consider (2.1) with uncertain initial data, deterministic flux function $F : \mathscr{U} \to \mathbb{R}^m$ and one-dimensional random space $\Xi \subset \mathbb{R}$. Our problem of interest is the following initial-value problem.

$$\begin{cases} \partial_t u(t,x,y) + \partial_x F(u(t,x,y)) = 0, & (t,x,y) \in (0,T) \times D \times \Xi, \\ u(0,x,y) = u^0(x,y), & (x,y) \in D \times \Xi. \end{cases} \tag{5.41}$$

### 5.3.1 Discretization, reconstruction and residuals

Recalling Section 4.1.3 we consider a set of quadrature points denoted by $\{y_\mathrm{q}\}_{\mathrm{q}=0}^{Q_\Xi} \subset \Xi$ and use them as collocation points in (5.41). Consequently, we obtain the following deterministic initial-value problems

$$\begin{cases} \partial_t u(t,x,y_\mathrm{q}) + \partial_x F(u(t,x,y_\mathrm{q})) = 0, & (t,x) \in (0,T)\times D, \\ u(0,x,y_\mathrm{q}) = u^0(x,y_\mathrm{q}), & x \in D. \end{cases} \tag{5.42}$$

for all $\mathrm{q} = 0,\dots,Q_\Xi$. For a simple notation let us assume that the time partition $\{t_n\}_{n=0}^{N_t}$ and the triangulation $\mathscr{T}$ used for (5.42) are the same for every quadrature point $\{y_\mathrm{q}\}_{\mathrm{q}=0}^{Q_\Xi}$. The numerical approximation of the random entropy admissible weak solution of (5.41) at time $t = t_n$ can be written as (cf. (4.16) and (4.17))

$$u_h^n(x,y) := \sum_{\mathrm{k}=0}^{K} \hat{u}_\mathrm{k}^n(x)\Psi_\mathrm{k}(y) = \sum_{\mathrm{k}=0}^{K} \Big( \sum_{\mathrm{q}=0}^{Q_\Xi} u_h^n(x,y_\mathrm{q})\Psi_\mathrm{k}(y_\mathrm{q})w_\mathrm{q} \Big) \Psi_\mathrm{k}(y). \tag{5.43}$$

For every collocation point $\{y_\mathrm{q}\}_{\mathrm{q}=0}^{Q_\Xi}$, the reconstruction process from the previous sections provides us with a computable space-time reconstruction $\hat{u}^{st}(y_\mathrm{q}) := \hat{u}^{st}(\cdot,\cdot,y_\mathrm{q}) \in W_\infty^1((0,T);\mathscr{V}_h^{q+1} \cap C^0(D;\mathscr{U}))$ of the numerical solution $\{u_h^n(y_\mathrm{q})\}_{n=0}^{N_t} \subset \mathscr{V}_h^q$. This allows us to define a space-time residual as follows.

**Definition 5.29** (Space-time residual)**.**
*For all* $\mathrm{q} = 0,\dots,Q_\Xi$*, we define* $\mathscr{R}^{st}(y_\mathrm{q}) := \mathscr{R}^{st}(\cdot,\cdot,y_\mathrm{q}) \in L^2((0,T)\times D;\mathbb{R}^m)$ *by*

$$\mathscr{R}^{st}(y_\mathrm{q}) := \partial_t \hat{u}^{st}(y_\mathrm{q}) + \partial_x F(\hat{u}^{st}(y_\mathrm{q})) \tag{5.44}$$

*to be the space-time residual associated with the quadrature point* $y_\mathrm{q}$.

Next we define the reconstructed mode, the space-time-stochastic reconstruction and the space-time-stochastic residual. The latter is obtained by plugging the space-time-stochastic reconstruction into the random conservation law (5.41).

**Definition 5.30** (Space-time-stochastic reconstruction and residual)**.**
*Let* $\{\hat{u}^{st}(y_\mathrm{q})\}_{\mathrm{q}=0}^{Q_\Xi} : (0,T)\times D \to \mathbb{R}^m$ *be the sequence of space-time reconstructions at quadrature points* $\{y_\mathrm{q}\}_{\mathrm{q}=0}^{Q_\Xi}$*. The reconstructed modes of* (5.43) *are defined as*

$$\hat{u}_\mathrm{k}^{st} := \sum_{\mathrm{q}=0}^{Q_\Xi} \hat{u}^{st}(y_\mathrm{q})\Psi_\mathrm{k}(y_\mathrm{q})w_\mathrm{q}, \tag{5.45}$$

*for* $\mathrm{k}=0,\dots,K$. *The space-time-stochastic reconstruction* $\hat{u}^{sts}\in\mathscr{W}_K(\Xi)\otimes W_\infty^1((0,T);\mathscr{V}_h^{q+1}\cap C^0(D;\mathscr{U}))$ *is defined as*

$$\hat{u}^{sts}(t,x,y):=\sum_{\mathrm{k}=0}^{K}\hat{u}^{st}_{\mathrm{k}}(t,x)\Psi_{\mathrm{k}}(y). \tag{5.46}$$

*Finally, we define the space-time-stochastic residual* $\mathscr{R}^{sts}\in L^2_w(\Xi;L^2((0,T)\times D;\mathbb{R}^m))$ *by*

$$\mathscr{R}^{sts}:=\partial_t\hat{u}^{sts}+\partial_x F(\hat{u}^{sts}). \tag{5.47}$$

In Lemma 5.31 we show a splitting of the space-time-stochastic residual $\mathscr{R}^{sts}$ into three parts. A deterministic residual, which corresponds to the spatial error when approximating (5.42) using the RKDG method, a quadrature residual that reflects the quadrature error from the discrete orthogonal projection in (4.16) and a stochastic cut-off error, which occurs when truncating the infinite Fourier series in (4.2).

**Lemma 5.31** (Orthogonal decomposition of the space-time-stochastic residual)**.**
*The space-time-stochastic residual* $\mathscr{R}^{sts}$ *from* (5.47) *admits the following orthogonal decomposition,*

$$\mathscr{R}^{sts}=\sum_{j=0}^{K}\Big(\mathscr{R}^{st}_j+\mathscr{R}^{sq}_j\Big)\Psi_j+\sum_{j>K}^{\infty}\mathscr{R}^{sc}_j\Psi_j, \tag{5.48}$$

*where*

$$\begin{aligned}
\mathscr{R}^{st}_j&:=\sum_{\mathrm{q}=0}^{Q_\Xi}\mathscr{R}^{st}(y_{\mathrm{q}})\Psi_j(y_{\mathrm{q}})w_{\mathrm{q}}\quad \textit{for } j=0,\dots,K\\
\mathscr{R}^{sq}_j&:=\Big\langle\partial_x F\Big(\sum_{\mathrm{k}=0}^{K}\hat{u}^{st}(y_{\mathrm{k}})\Psi_{\mathrm{k}}\Big),\Psi_j\Big\rangle-\sum_{\mathrm{q}=0}^{Q_\Xi}\partial_x F(\hat{u}^{st}(y_{\mathrm{q}}))\Psi_j(y_{\mathrm{q}})w_{\mathrm{q}}\quad \textit{for } j=0,\dots,K\\
\mathscr{R}^{sc}_j&:=\Big\langle\partial_x F\Big(\sum_{\mathrm{k}=0}^{K}\hat{u}^{st}(y_{\mathrm{k}})\Psi_{\mathrm{k}}\Big),\Psi_j\Big\rangle\quad \textit{for } j>K
\end{aligned}$$

*are called the* $j$*-th mode of the space-time, stochastic quadrature and stochastic cut-off residual. Moreover, we have*

$$\begin{aligned}
\mathscr{E}^{sts}(t):=\|\mathscr{R}^{sts}\|^2_{L^2_w(\Xi;L^2((0,t)\times D))}&=\sum_{j=0}^{K}\|\mathscr{R}^{st}_j+\mathscr{R}^{sq}_j\|^2_{L^2((0,t)\times D)}+\sum_{j>K}^{\infty}\|\mathscr{R}^{sc}_j\|^2_{L^2((0,t)\times D)}\\
&\le 2\mathscr{E}^{st}(t)+2\mathscr{E}^{sq}(t)+\mathscr{E}^{sc}(t),
\end{aligned} \tag{5.49}$$

*where, for any* $t\in(0,T)$,

$$\mathscr{E}^{st}(t):=\sum_{j=0}^{K}\|\mathscr{R}^{st}_j\|^2_{L^2((0,t)\times D)},\ \mathscr{E}^{sq}(t):=\sum_{j=0}^{K}\|\mathscr{R}^{sq}_j\|^2_{L^2((0,t)\times D)},\ \mathscr{E}^{sc}(t):=\sum_{j>K}^{\infty}\|\mathscr{R}^{sc}_j\|^2_{L^2((0,t)\times D)}.$$

*Proof.* We recall that the space-time reconstruction $\hat{u}^{st}(y_{\mathrm{q}})$ satisfies

$$\mathscr{R}^{st}(y_{\mathrm{q}}) = \partial_t \hat{u}^{st}(y_{\mathrm{q}}) + \partial_x F(\hat{u}^{st}(y_{\mathrm{q}})) \tag{5.50}$$

for all $\mathrm{q} = 0, \dots, Q_\Xi$. Moreover, the reconstructed mode $\hat{u}^{st}_j$ was defined as (cf. (5.45))

$$\hat{u}^{st}_j = \sum_{\mathrm{q}=0}^{Q_\Xi} \hat{u}^{st}(y_{\mathrm{q}}) \Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}} \tag{5.51}$$

for all $j = 0, \dots, K$. Multiplying (5.50) by $\Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}}$ and suming over $\mathrm{q} = 0, \dots, Q_\Xi$ yields, using (5.51), the following relationship

$$\sum_{\mathrm{q}=0}^{Q_\Xi} \mathscr{R}^{st}(y_{\mathrm{q}}) \Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}} = \partial_t \hat{u}^{st}_j + \sum_{\mathrm{q}=0}^{Q_\Xi} \partial_x F(\hat{u}^{st}(y_{\mathrm{q}})) \Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}}. \tag{5.52}$$

By definition of the space-time-stochastic residual we have

$$\mathscr{R}^{sts} = \partial_t \hat{u}^{sts} + \partial_x F(\hat{u}^{sts}) = \partial_t \Big( \sum_{\mathrm{k}=0}^{K} \hat{u}^{st}_{\mathrm{k}} \Psi_{\mathrm{k}} \Big) + \partial_x F \Big( \sum_{\mathrm{k}=0}^{K} \hat{u}^{st}_{\mathrm{k}} \Psi_{\mathrm{k}} \Big).$$

Let us begin by studying the $j$-th mode of $\mathscr{R}^{sts}$ for $j = 0, \dots, K$. In this case the orthogonality relation (4.1) yields

$$\Big\langle \mathscr{R}^{sts}, \Psi_j \Big\rangle = \Big\langle \partial_t \hat{u}^{sts} + \partial_x F(\hat{u}^{sts}), \Psi_j \Big\rangle = \partial_t \hat{u}^{st}_j + \Big\langle \partial_x F \Big( \sum_{\mathrm{k}=0}^{K} \hat{u}^{st}_{\mathrm{k}} \Psi_{\mathrm{k}} \Big), \Psi_j \Big\rangle. \tag{5.53}$$

Using (5.52) we obtain

$$\begin{aligned} \Big\langle \mathscr{R}^{sts}, \Psi_j \Big\rangle &= \sum_{\mathrm{q}=0}^{Q_\Xi} \mathscr{R}^{st}(y_{\mathrm{q}}) \Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}} \\ &\quad + \Big\langle \partial_x F \Big( \sum_{\mathrm{k}=0}^{K} \hat{u}^{st}_{\mathrm{k}} \Psi_i \Big), \Psi_j \Big\rangle - \sum_{\mathrm{q}=0}^{Q_\Xi} \partial_x F(\hat{u}^{st}(y_{\mathrm{q}})) \Psi_j(y_{\mathrm{q}}) w_{\mathrm{q}} = \mathscr{R}^{st}_j + \mathscr{R}^{sq}_j. \end{aligned} \tag{5.54}$$

For $j > K$ the $j$-th moment of $\mathscr{R}^{sts}$ is

$$\Big\langle \mathscr{R}^{sts}, \Psi_j \Big\rangle = \Big\langle \partial_x F \Big( \sum_{\mathrm{k}=0}^{K} \hat{u}^{st}_{\mathrm{k}} \Psi_{\mathrm{k}} \Big), \Psi_j \Big\rangle = \mathscr{R}^{sc}_j. \tag{5.55}$$

Formula (5.48) then follows from (5.54) and (5.55). Formula (5.49) is an application of the Pythagorean theorem for $L^2_w(\Xi)$. □

We directly present the following a posteriori error estimate with separable error bounds, which follows from Theorem 2.15 and Lemma 5.31.

**Theorem 5.32** (A posteriori error bound for the numerical solution with error splitting)**.**
*Let $u$ be the random entropy admissible weak solution of* (5.41)*. Provided the reconstruction $\hat{u}^{sts}$ from Definition 5.46 only takes values in a compact, convex set $\mathscr{C} \subset \mathscr{U}$, the difference between $u$ and the numerical solution $u_h^n$ from* (5.43) *satisfies*

$$\begin{aligned}\|u(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))} \leq\ & 2\|\hat{u}^{sts}(t_n,\cdot,\cdot)-u_h^n(\cdot,\cdot)\|^2_{L^2_w(\Xi;L^2(D))}\\ &+2C_{\underline{\eta}}^{-1}\left(2\mathscr{E}^{st}(t_n)+2\mathscr{E}^{sq}(t_n)+\mathscr{E}^{sc}(t_n)+C_{\overline{\eta}}\mathscr{E}_0^{sts}\right)\\ &\times\exp\left(C_{\underline{\eta}}^{-1}\int_0^{t_n}\left(C_{\overline{\eta}}C_{\overline{F}}\|\partial_x\hat{u}^{sts}(t,\cdot,\cdot)\|_{L^\infty_w(\Xi;L^\infty(D))}+C_{\overline{\eta}}^2\right)\,\mathrm{dt}\right),\end{aligned}$$

*for all $n=1,\ldots,N_t$.*

Theorem 5.32 shows that the entire space-stochastic error can be decomposed into three parts, where $\mathscr{E}^{st}$ quantifies the spatio-temporal discretization error of the RKDG scheme, $\mathscr{E}^{sq}$ assesses the quality of the discrete orthogonal projection and $\mathscr{E}^{sc}$ quantifies the stochastic error by truncation of the generalized polynomial chaos series.

# 6 Statistical Uncertainty Quantification methods

The previous chapters were concerned with UQ methods for random conservation laws, which relied on polynomial approximations of the underlying random field. In this chapter we discuss a simpler and in most cases more robust approach for UQ of compressible flows, based on Monte Carlo (MC) and Multilevel Monte Carlo (MLMC) methods. As our novel contribution we extend the classical MLMC method, which considers hierarchies of meshes of different resolution ($h$-refinement), to hierarchies which consist of arbitrarily $h$-, $p$- or $hp$-refined meshes. Due to the multilevel structure, the bulk of computational time is shifted towards coarse levels and only a few computations are conducted on fine levels, yielding significant savings in computational work. We extend the well-known complexity results for the $h$-MLMC method to what we call $hp$-MLMC method and prove that in the best case the asymptotic work scales quadratically with the prescribed tolerance.

An important aspect of an iterative version of $hp$-MLMC is the optimal number of samples which have to be computed on each level during one iteration. Due to random fluctuations in computational time (for example due to a random time-step) and the fact that statistical quantities like variances are only estimated, the optimal number of samples is in fact a statistical quantity. A reliable estimate of the optimal number of samples requires a sufficiently large sample size, which might be not available for deep levels or in the warm-up phase of the algorithm. Consequently, a wrong estimate may lead to a severe overestimation of the actual optimal number of samples, leading to a decrease of efficiency of the $hp$-MLMC method. Therefore, to account for this additional uncertainty we are going to derive a lower confidence bound for the optimal number of samples, which ensures that the optimal number of samples is not overestimated.

Finally, we present numerical results for our numerical scheme and apply the $h$-,$p$- and $hp$-MLMC method to different uncertain, viscous, compressible flow problems, governed by the Navier–Stokes equations. In particular, we consider an important flow problem from computa-

tional acoustics, a so-called cavity flow. It exhibits physical phenomena with high sensitivity with respect to the problem parameters, thus it poses a challenging problem for UQ. Results from this chapter have been published in [11] and the cavity example is from [74].

## 6.1 Monte Carlo method

The most straightforward approach for UQ of random conservation laws is the MC method. It generates random inputs of the uncertain data, performs a deterministic simulation for every random input and approximates mean and variance of the underlying random field by computing sample mean and sample variance. More specifically, we are interested to determine

$$\mathbb{E}\big(Q(u(t,x,\cdot))\big) = \int_{\Xi} Q(u(t,x,y))\, w(y)\mathrm{d}y \text{ and} \tag{6.1}$$

$$\mathrm{Var}\big(Q(u(t,x,\cdot))\big) = \int_{\Xi} \Big(\mathbb{E}\big(Q(u(t,x,y))\big) - Q(u(t,x,y))\Big)^2 w(y)\mathrm{d}y \tag{6.2}$$

for a.e. $(t,x) \in (0,T)\times D$. Here $Q$ can be any arbitrary non-linear function or functional of $u$.

We let $\{\hat{y}_j\}_{j=1}^M \subset \Xi$ denote a set of randomly drawn input samples. Using $\hat{y}_j$ as an input parameter in (2.1) we compute a numerical approximation $\{u_h^j(t_n,\cdot)\}_{n=0}^{N_t} := \{u_h(t_n,\cdot,\hat{y}_j)\}_{n=0}^{N_t} \subset \mathscr{V}_h^q$ using the RKDG method from Section 3.1. We assume that the samples $\{u_h^j\}_{j=1}^M$ are independent and identically distributed (iid). The MC estimator of (6.1) is defined as follows:

$$\mathrm{E}_{\mathrm{MC}}^M\big[Q(u_h)\big] := \frac{1}{M}\sum_{j=1}^{M} Q(u_h^j) \approx \mathbb{E}\big(Q(u)\big). \tag{6.3}$$

The following lemma illustrates the rather slow convergence of the classical MC method with respect to the sample size $M \in \mathbb{N}$.

**Lemma 6.1.**
*Let $M \in \mathbb{N}$ and assume that $Q(u) \in L_w^2(\Xi;E)$ for some suitable Banach space $E$. Then the mean square error (MSE) satisfies*

$$\mathbb{E}\big(\|\mathbb{E}\big(Q(u)\big) - E_{MC}^M\big[Q(u)\big]\|_E\big)^2 \le M^{-1}\sigma^2,$$

*where $\sigma^2 := \mathbb{E}\big(\|\mathbb{E}\big(u\big) - u\|_E\big)^2$.*

*Proof.* See the proof of Lemma 4.1 in [8]. □

Hence, the classical MC method converges with a rate of $M^{-1/2}$ and is thus not suitable for problems where the computation of a single sample is computationally expensive. In our numerical

experiments where we consider two-dimensional flow problems governed by the compressible Navier–Stokes equations, MC is not feasible because the computation of a single sample is too time-consuming. The situation becomes even worse when considering three-dimensional flow problems.

## 6.2 $hp$-Multilevel Monte Carlo method

To overcome this problem Heinrich [61] and later Giles [51] extended the MC method to the Multilevel Monte Carlo (MLMC) method, where they considered different mesh hierarchies instead of one fixed mesh, to discretize the deterministic equation of interest. The main idea of the MLMC method is that the global behavior of the exact expectation can be approximated by the behavior of the expectation of numerical solutions with a low spatial resolution, which can be computed at low cost. The coarse expectation is then subsequently corrected by only a few computations on finer levels, which are computationally more expensive per sample. For a successful application of MC and MLMC for UQ of conservation laws, we refer to [6, 7, 40, 79, 80, 81, 86] and references therein.

The classical MLMC method of Giles considers a hierarchy of different spatial meshes, which corresponds to a $h$-refinement in spatial domain. We therefore dub this method $h$-MLMC. On the other hand, mesh hierarchies can also be obtained by varying other discretization parameters, for example the degree of the ansatz polynomials. The resulting method is called $p$-MLMC, or Multiorder MC (MOMC) as in [81]. In this thesis we generalize the $h$- and $p$-MLMC method to arbitrarily, i.e. $hp$-refined mesh hierarchies. Consequently, we call this method $hp$-MLMC method. To give a precise definition of the $hp$-MLMC method we consider for $l = 0,\dots,L \in \mathbb{N}$, spatial meshes with $N_l \in \mathbb{N}$ elements and ansatz spaces of polynomial degree $q_l \in \mathbb{N}_0$. We choose the number of cells $N_l$ and DG polynomial degrees $q_l$, such that $N_0 < \ldots < N_L$ and $q_0 < \ldots < q_l$ holds, i.e., we simultaneously increase the mesh size and the DG polynomial degree. With $\mathscr{V}_{h_l}^{q_l}$ we denote the DG polynomial space corresponding to level $l = 0,\dots,L$ . Moreover, $u_l(t,\cdot,y) \in \mathscr{V}_{h_l}^{q_l}$ (a.e. $(t,y) \in (0,T)\times\Xi$) is the DG numerical approximation associated with level $l = 0,\dots,L$. Additionally, the deterministic numerical solution on level $l = 0,\dots,L$ for input parameter $\hat{y}_j \in \Xi$, is denoted by $u_l^j := u_l(\cdot,\cdot,\hat{y}_j)$ and will subsequently be called sample.

Next we advance the MC estimator $\mathrm{E}_{\mathrm{MC}}^{M}\big[\,\cdot\,\big]$ to the $hp$-MLMC estimator $\mathrm{E}_{hp}^{L}\big[\,\cdot\,\big]$ by using the

linearity of the expectation in combination with a telescoping sum. We then write (see [51])

$$\mathbb{E}\big(Q(u_L)\big) = \sum_{l=0}^{L} \mathbb{E}\big(Q(u_l) - Q(u_{l-1})\big), \tag{6.4}$$

where we define $Q(u_{-1}) = 0$. Now each term in (6.4) can be estimated by the MC estimator (6.3). If we let $M_l \in \mathbb{N}$ denote a level-dependent number of samples for each level $l = 0, \dots, L$ and assume that the samples $\{Q(u_l^j)\}_{j=1}^{M_l}$, $l = 0, \dots, L$, on different levels are independent from each other, we obtain the $hp$-MLMC estimator via

$$\begin{aligned} \mathrm{E}_{hp}^{L}\big[Q(u_L)\big] &:= \sum_{l=0}^{L} \frac{1}{M_l} \sum_{j=1}^{M_l} (Q(u_l^j) - Q(u_{l-1}^j)) = \sum_{l=0}^{L} \mathrm{E}_{\mathrm{MC}}^{M_l}\big[Q(u_l) - Q(u_{l-1})\big] \\ &\approx \sum_{l=0}^{L} \mathbb{E}\big(Q(u_l) - Q(u_{l-1})\big) = \mathbb{E}\big(Q(u_L)\big). \end{aligned}$$

Here, $Q(u_l^j)$ and $Q(u_{l-1}^j)$, are computed using the same sample $y_l^j \in \Xi$. The main idea of the MLMC estimator is that the global behavior of the exact expectation can be approximated by the behavior of the expectation of numerical solutions with a low resolution, where each sample can be computed with low cost. Thus, $M_l$ should be big for coarse levels. The coarse-level expectation is then subsequently corrected by a few computations on finer levels, for which each sample is computationally expensive and therefore $M_l$ should be small on fine levels. Hence, the most important aspect for the efficiency of the $hp$-MLMC estimator is the correct choice of $M_l$. In the following section we show how to obtain $M_l$ as solution of a constrained optimization problem.

### 6.2.1 Optimal number of samples

For the following analysis we set the QoI to

$$Q(u) = u(t, \cdot, \cdot) \tag{6.5}$$

for a fixed $t \in [0, T]$. We note that our analysis can be analogously performed using any other QoI, including functional ones and suitable norms.

**Remark 6.2.**

*We consider the QoI* (6.5) *because we are mainly interested in statistical quantities of the solution u itself. However, many other QoIs have a higher regularity than the solution u, especially if Q is a functional. Therefore, using a QoI with high regularity yields a faster decrease of the variance across different levels, which increases the performance of the MLMC method*

*compared to MC. In [1, Fig. 10] it has been numerically shown that for an uncertain Kelvin-Helmholtz problem the level variance of the solution u does not decrease, because upon each mesh refinement additional smaller scale structures are detected. Thus, MLMC method provides no computational gains compared to the MC when the QoI is the solution u itself. We observe a similar behavior for the open cavity problem in Example 6.4.2 and for this example we consider a QoI based on the pressure fluctuations of the solution. This QoI admits a fast decrease of the variance across different levels.*

With the help of the following representation of the MSE we derive an optimal number of samples $M_l$ for all $l = 0, \dots, L$.

$$\begin{aligned} \text{MSE} &:= \|\mathbb{E}\big(u(t,\cdot,\cdot)\big) - \mathrm{E}_{hp}^{L}\big[u_L(t,\cdot,\cdot)\big]\|_{L_w^2(\Xi;L^2(D))} \leq \varepsilon_{\text{det}} + \varepsilon_{\text{stat}}, \\ \varepsilon_{\text{det}} &:= \|\mathbb{E}\big(u(t,\cdot,\cdot)\big) - \mathbb{E}\big(u_L(t,\cdot,\cdot)\big)\|_{L^2(D)}, \\ \varepsilon_{\text{stat}} &:= \|\mathbb{E}\big(u_L(t,\cdot,\cdot)\big) - \mathrm{E}_{hp}^{L}\big[u_L(t,\cdot,\cdot)\big]\|_{L_w^2(\Xi;L^2(D))}. \end{aligned} \tag{6.6}$$

The first term $\varepsilon_{\text{det}}$ in (6.6) is the deterministic approximation error (bias). It accounts for the insufficient resolution of the deterministic system. The second term $\varepsilon_{\text{stat}}$ corresponds to the statistical (sampling) error. It occurs due to the finite number of samples in (6.3). The optimal number of samples $M_l$ is then chosen to minimize this term. For notational convenience we will in the following suppress the explicit dependence on $t \in [0,T]$. Using the independence of the samples, we rewrite the statistical error in (6.6) (cf. [83]):

$$\begin{aligned} \varepsilon_{\text{stat}}^2 &= \mathbb{E}\Big[\|\mathbb{E}\big(u_L\big) - \mathrm{E}_{hp}^{L}\big[u_L\big]\|_{L^2(D)}^2\Big] \\ &= \mathbb{E}\Big[\|\sum_{l=0}^{L}\mathbb{E}\big(u_l - u_{l-1}\big) - \mathrm{E}_{\text{MC}}^{M_l}\big[u_l - u_{l-1}\big]\|_{L^2(D)}^2\Big] \\ &= \sum_{l=0}^{L}\frac{1}{M_l^2}\sum_{i=1}^{M_l}\mathbb{E}\Big[\|\mathbb{E}\big(u_l - u_{l-1}\big) - (u_l^i - u_{l-1}^i)\|_{L^2(D)}^2\Big] \\ &= \sum_{l=0}^{L}\frac{1}{M_l}\mathbb{E}\Big[\|\mathbb{E}\big(u_l - u_{l-1}\big) - (u_l - u_{l-1})\|_{L^2(D)}^2\Big] \\ &=: \sum_{l=0}^{L}\frac{\sigma_l^2}{M_l}. \end{aligned} \tag{6.7}$$

From the representation (6.7), the optimal number of samples can be obtained by an error-complexity analysis as in [51, 81, 83]. We introduce the total work

$$\mathrm{W}_{\text{tot}} := \mathrm{W}_{\text{tot}}(M_0, \dots, M_L) := \sum_{l=0}^{L} M_l w_l, \tag{6.8}$$

where $w_l$ is the work needed to create one sample $u_l^j - u_{l-1}^j$. Following [51] we obtain the optimal number of samples on different levels by considering the following minimization problem.

$$\text{For a tolerance } \varepsilon > 0, \underset{M_0,\dots,M_L \in \mathbb{N}}{\text{minimize}} \, \mathrm{W}_{\text{tot}} \text{ under the constraint } \sum_{l=0}^{L} \frac{\sigma_l^2}{M_l} \leq \frac{1}{4}\varepsilon^2. \tag{6.9}$$

The minimization problem can be explicitly solved by (cf. [22, 51, 81])

$$M_l = \left\lceil \left(\frac{\varepsilon}{2}\right)^{-2} \sqrt{\frac{\sigma_l^2}{w_l}} \sum_{k=0}^{L} \sqrt{\sigma_k^2 w_k} \right\rceil. \tag{6.10}$$

Since we consider an iterative version of the $hp$-MLMC method, we denote by $M_{\text{tot}_l}$ the number of samples on level $l \in \{0,\dots,L\}$ that have already been calculated during the iterations of the algorithm. We want to emphasize that $M_{\text{tot}_l}$ is not equal to $M_l$, which is the estimated optimal number of samples from (6.10).

The level variances $\sigma_0^2, \dots \sigma_L^2$ in (6.7) are not known in general and we therefore use the unbiased estimator as in [83]:

$$\hat{\sigma}_l^2 := \frac{1}{M_{\text{tot}_l} - 1} \sum_{j=1}^{M_{\text{tot}_l}} \int_D \left( \left( \frac{1}{M_{\text{tot}_l}} \sum_{i=1}^{M_{\text{tot}_l}} (u_i^j - u_{l-1}^i) \right) - (u_{l-1}^j - u_{l-1}^j) \right)^2 dx. \tag{6.11}$$

The work required for the simulation of one sample can vary with an uncertain parameter (e.g. when uncertain viscosity influences the time-step restriction). Moreover, on high performance computing systems, random variations in work can occur between two executions of the same simulation. In order to account for this uncertainty, we estimate the work $w_l$ on level $l = 0,\dots,L$ by the average work per sample $U_l^i$, denoted by $w_l^i$, and define the average work by

$$\hat{w}_l := \frac{1}{M_{\text{tot}_l}} \sum_{i=1}^{M_{\text{tot}_l}} w_l^j \tag{6.12}$$

As a matter of fact, $M_l$ from (6.10) is also only estimated and we call the estimator in the following $\hat{M}_l$.

We now have all ingredients to state the classical MLMC algorithm proposed by Giles in [51]. It essentially consists of two parts. The first part aims at finding a maximum number of levels $L$ to bound the bias term $\varepsilon_{\text{det}}$ in (6.6). In the second part of the algorithm the optimal number of samples $M_l$ is computed such that (6.9) is satisfied. With this choice we obtain MSE $\leq \varepsilon$ with the smallest possible computational cost. For the sake of illustration we shortly recapitulate the most basic form of the $hp$-MLMC algorithm, where we fix the number of levels $L$ beforehand. The algorithm is in most parts similar to the ones presented in [51, 83]. The basic idea is to

compute a number of warm-up samples $K_l$ on each level $l = 0, \dots, L$, then estimate $w_l$, $\sigma_l^2$ and compute $M_l$ by formula (6.12). Finally, we add $M_l - M_{\text{tot}_l}$ new samples and repeat the process as long as no new samples have to be added. The algorithm reads as follows.

**Algorithm** $hp$-MLMC

1: Fix a tolerance $\varepsilon > 0$, the maximum level $L \in \mathbb{N}$ and set $\mathscr{L} := \{0, \dots, L\}$
2: Compute $K_l$ (warm-up) samples on level $l = 0, \dots, L$ and set $M_{\text{tot}_l} := K_l$
3: **while** $\mathscr{L} \neq \emptyset$ **do**
4:  **for** $l \in \mathscr{L}$ **do**
5:  Estimate $w_l$ by (6.12), $\sigma_l^2$ by (6.11) and then $M_l$ by (6.10)
6:  **if** $M_l > M_{\text{tot}_l}$ **then**
7:  Add $(M_l - M_{\text{tot}_l})$ new samples of $u_l^j - u_{l-1}^j$ and update $M_{\text{tot}_l}$
8:  **else**
9:  Set $\mathscr{L} := \mathscr{L} \setminus \{l\}$
10:  **end if**
11:  **end for**
12: **end while**
13: Compute $\mathrm{E}_{hp}^L\big[u_L\big]$

Based on Algorithm $hp$-MLMC, we want to discuss several important aspects of the $hp$-MLMC method. First, the complexity of the algorithm will be analyzed in Theorem 6.4. The choice of the maximum level $L$ will be considered in Remark 6.8. The discussion of the number of warm-up samples $K_0, \dots, K_L$ (line two in the algorithm), resp. the additional samples (line six and seven in the algorithm) will be postponed to Section 6.2.3, where we derive lower confidence bounds for the optimal number of samples $M_l$.

### 6.2.2 Computational complexity of $hp$-MLMC

Let us first consider the computational complexity of Algorithm $hp$-MLMC. To analyze the $hp$-MLMC method we impose the following assumptions. To simplify the estimates in Theorem 6.4 below, we set $\tilde{q}_l = q_l + 1$ for all $l = 0, \dots, L$,

(A1) Asymptotic work: $\exists\, \gamma_1, c_1 > 0$ (independent of $h_l, \tilde{q}_l$): $w_l \leq c_1 (h_l^{-1} \tilde{q}_l)^{\gamma_1}$ for all $l \in \mathbb{N}$

(A2) Bias reduction: $\exists\, \kappa_1, c_2 > 0$ (independent of $h_l, \tilde{q}_l$): $\|\mathbb{E}\big(U\big) - \mathbb{E}\big(U_l\big)\|_{L^2(D)} \leq c_2 h_l^{\kappa_1 \tilde{q}_l}$ for all $l \in \mathbb{N}$.

(A3) Variance reduction between two levels: $\exists\, c_3 > 0$, (independent of $h_l, \tilde{q}_l$): $\sigma_l^2 \leq c_3 h_l^{\kappa_2 \tilde{q}_l}$ for some $\kappa_2 > 0$ with $\kappa_1 \tilde{q}_0 \geq \frac{\kappa_2 \tilde{q}_0}{2}$ and for all $l \in \mathbb{N}$.

**Remark 6.3.** *1. In Assumption (A1) it is required that the computational work is bounded by the number of degrees of freedom on levels $l = 0, \dots, L$. For the RKDG method in $d$ spatial dimensions we have $h_l^{-d}(q_l+1)^d$ spatial degrees of freedom times the number of time-steps, which is proportional to $h_l^{-1}(q_l+1)^1$, or $h_l^{-2}(q_l+1)^2$, depending on evaluation of the maximum in the CFL condition* (6.52)*. Thus, the total work asymptotically equals $\mathcal{O}(h_l^{-(d+1)}(q_l+1)^{d+1})$, resp. $\mathcal{O}(h_l^{-(d+2)}(q_l+1)^{d+2})$ and we therefore expect the parameter $\gamma_1$ to satisfy $\gamma_1 = d+1$, or $\gamma_1 = d+2$.*

2. *Assumption (A2) assumes that the bias, i.e., the deterministic approximation error, converges with the order of the DG method. For smooth solutions the order of convergence is $\mathcal{O}(h_l^{q_l+1}) = \mathcal{O}(h_l^{\tilde{q}_l})$, cf. [110] and the numerical experiments in [63]. Therefore, we expect $\kappa_1 = 1$.*

3. *In Assumption (A3) we assume that the variance on level $l = 0 \dots, L$, decays similar to the bias term. If we consider regular solutions of a random pde, we expect $\kappa_2 = 2$, cf. the discussion in [86, p. 25].*

In Theorem 6.4 below we consider two different cases $\kappa_2\tilde{q}_0 \gtrless \gamma_1$. For the second case $\kappa_2\tilde{q}_0 < \gamma_1$ we introduce a critical level $L^* \in \mathbb{N}$ such that $\kappa_2\tilde{q}_{L^*} < \gamma_1$ and $\kappa_2\tilde{q}_{L^*+1} \geq \gamma_1$. With this notation and the assumptions from above we can prove an optimality result for the complexity of the $hp$-MLMC method. This result generalizes [22, Theorem 1] and [81, Theorem 3] to $hp$-refined mesh hierarchies.

**Theorem 6.4** (Complexity of the $hp$-MLMC method)**.**
*For $\beta \in \mathbb{N}$, and $q_0 \geq 1$ let $\{q_l := q_0 + \beta l\}_{l\in\mathbb{N}}$ be a sequence of DG polynomial degrees. Additionally, we consider a family of meshes with associated mesh size $h_l = \lambda^{-l}h_0$ for some $h_0 \in (0,1)$ and $\lambda \geq 2$. Let $\mathcal{V}_{h_l}^{q_l}$ be the corresponding DG spaces.*
*Under the assumptions* (A1) - (A3)*, there exists a constant $c > 0$, such that for any tolerance $0 < \varepsilon < \frac{1}{e}$, there exists a maximum level $L = L(\varepsilon) \in \mathbb{N}$, $L(\varepsilon) \geq 2$, a number of samples $M_l$ on each level $l \in \{0, \dots, L(\varepsilon)\}$, such that the mean square error from* (6.6) *satisfies* $\mathrm{MSE} \leq \varepsilon$ *with the computational complexity*

$$W_{\mathrm{tot}} \leq \begin{cases} c\varepsilon^{-2}, & \kappa_2\tilde{q}_0 > \gamma_1, \\ c\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\min\{\kappa_1\tilde{q}_L, \kappa_1\tilde{q}_{L^*}\}}}, & \kappa_2\tilde{q}_0 < \gamma_1. \end{cases}$$

**Remark 6.5.**
*Theorem 6.4 states that in the worst case $\kappa_2\tilde{q}_0 < \gamma_1$, there exists a threshold for the asymptotic complexity, which is $\mathcal{O}(\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L^*}}})$ which cannot be improved for $L > L^*$.*

Before we present the proof of Theorem 6.4 we quickly state the following two lemmas which will be needed for the proof of Theorem 6.4.

**Lemma 6.6** ([81])**.**
*For every real number $r \in (0,1)$ and any $p \in \mathbb{N}$, $p \geq 1$, we have*

$$\sum_{l=0}^{\infty} r^{(\tilde{q}_0+\beta l)}(\tilde{q}_0+\beta l)^p = \sum_{k=1}^{p} c_k r^k \Big(\frac{d}{dr}\Big)^k \frac{r^{\tilde{q}_0}}{1-r^\beta}. \tag{6.13}$$

*Proof.* Due to the geometric series formula we have

$$r^{\tilde{q}_0} \sum_{l=0}^{\infty} (r^\beta)^l = r^{\tilde{q}_0} \frac{1}{1-r^\beta} =: f(r).$$

Differentiation with respect to $r$ and then multiplication by $r$ yields

$$r \sum_{l=0}^{\infty} (\tilde{q}_0+\beta l) r^{(\tilde{q}_0+\beta l-1)} = r f'(r).$$

Repeating this process $p$ times yields the assertion of Lemma 6.6. □

**Lemma 6.7.**
*The following equality holds,*

$$h_{L+1}^{\kappa_1 \tilde{q}_{L+1}} = h_L^{\kappa_1 \tilde{q}_L} \lambda^{-\kappa_1(\tilde{q}_0+\beta)} \lambda^{-2\kappa_1 \beta L} h_0^{\kappa_1 \beta}. \tag{6.14}$$

*Proof.* A straightforward calculation yields

$$\begin{aligned} h_{L+1}^{\kappa_1 \tilde{q}_{L+1}} &= (\lambda^{-1} h_L)^{\kappa_1(\tilde{q}_L+\beta)} \\ &= \lambda^{-\kappa_1 \tilde{q}_L} \lambda^{-\kappa_1 \beta} h_L^{\kappa_1 \tilde{q}_L} h_L^{\kappa_1 \beta} \\ &= h_L^{\kappa_1 \tilde{q}_L} \lambda^{-\kappa_1 \tilde{q}_0} \lambda^{-\kappa_1 \beta L} \lambda^{-\kappa_1 \beta} (\lambda^{-L} h_0)^{\kappa_1 \beta} \\ &= h_L^{\kappa_1 \tilde{q}_L} \lambda^{-\kappa_1(\tilde{q}_0+\beta)} \lambda^{-2\kappa_1 \beta L} h_0^{\kappa_1 \beta}. \end{aligned}$$

□

*Proof of Theorem 6.4.* We structure the proof as follows. First, we choose the numbers of levels $L(\varepsilon) \in \mathbb{N}$ to bound the bias term $\varepsilon_{\text{det}}$ in (6.6). After fixing $L = L(\varepsilon)$ we choose the numbers of samples $M_l$ for every level $l = 0, \ldots, L$.

By Assumption (A2) the bias term satisfies

$$\varepsilon_{\text{det}} = \|\mathbb{E}(u) - \mathbb{E}(u_L)\|_{L^2(D)} \leq c_2 h_L^{\kappa_1 \tilde{q}_L} \leq \frac{\varepsilon}{2}. \tag{6.15}$$

The statement in (6.15) can be equivalently written as

$$2c_2\varepsilon^{-1}h_L^{\kappa_1\tilde{q}_L} = 2c_2\varepsilon^{-1}(\lambda^{-L}h_0)^{\kappa_1(\tilde{q}_0+\beta L)} \leq 1.$$

Taking the logarithm yields

$$P(L) := \log(2c_2\varepsilon^{-1}) + \kappa_1(\tilde{q}_0+\beta L)\Big(\log(h_0) - L\log(\lambda)\Big) \leq 0. \tag{6.16}$$

The roots of the quadratic polynomial $P$ are given by the real numbers

$$L_1 = \frac{\kappa_1(\beta\log(h_0) - \tilde{q}_0\log(\lambda)) + \Big(\beta^2\kappa_1^2\log(h_0)^2 + \kappa_1^2\tilde{q}_0^2\log(\lambda)^2}{2\beta\kappa_1\log(\lambda)}$$
$$\frac{+4\beta\kappa_1\log(2)\log(\lambda) + 4\beta\kappa_1\log(\lambda)\log(c_2/\varepsilon) + 2\beta\kappa_1^2\tilde{q}_0\log(\lambda)\log(h_0)\Big)^{\frac{1}{2}}}{2\beta\kappa_1\log(\lambda)},$$

$$L_2 = \frac{\kappa_1(\beta\log(h_0) - \tilde{q}_0\log(\lambda)) - \Big(\beta^2\kappa_1^2\log(h_0)^2 + \kappa_1^2\tilde{q}_0^2\log(\lambda)^2}{2\beta\kappa_1\log(\lambda)}$$
$$\frac{+4\beta\kappa_1\log(2)\log(\lambda) + 4\beta\kappa_1\log(\lambda)\log(c_2/\varepsilon) + 2\beta\kappa_1^2\tilde{q}_0\log(\lambda)\log(h_0)\Big)^{\frac{1}{2}}}{2\beta\kappa_1\log(\lambda)}.$$

From now on we fix the level

$$L = L(\varepsilon) = \max\{\lceil L_1\rceil, \lceil L_2\rceil, 2\}. \tag{6.17}$$

Without loss of generality we assume that the maximum in (6.17) is attained at $\lceil L_1\rceil$. Due to $L_1 \leq L \leq L_1 + 1$ we obtain

$$c_2 h_{L_1+1}^{\kappa_1\tilde{q}_{L_1+1}} < c_2 h_L^{\kappa_1\tilde{q}_L} \leq c_2 h_{L_1}^{\kappa_1\tilde{q}_{L_1}} \leq \frac{1}{2}\varepsilon. \tag{6.18}$$

Using Lemma 6.7 yields the lower bound

$$\frac{1}{2}\varepsilon\delta\lambda^{-2\kappa_1\beta L_1} < c_2 h_L^{\kappa_1\tilde{q}_L} \leq \frac{1}{2}\varepsilon, \tag{6.19}$$

with $\delta := \lambda^{-\kappa_1(\tilde{q}_0+\beta)}h_0^{\kappa_1\beta}$. Next, we consider two different cases.

**First case:** $\kappa_2\tilde{q}_0 > \gamma_1$

We choose the number of samples on level $l = 0,\ldots,L$ to be

$$M_l := \left\lceil 4\varepsilon^{-2}c_3 S h_l^{(\gamma_1+\kappa_2\tilde{q}_l)/2}\tilde{q}_l^{-\gamma_1/2}\right\rceil. \tag{6.20}$$

Here,

$$S = h_0^{-\gamma_1/2} \sum_{k=1}^{\lceil \gamma_1/2 \rceil} c_k r^k \Big(\frac{d}{dr}\Big)^k f(r), \quad f(r) = \frac{r^{\tilde{q}_0}}{1-r^\beta} \tag{6.21}$$

is the right-hand side in (6.13) with $r = h_0^{\kappa_2/2}$. Using (6.20) and Assumption (A3) we obtain

$$\begin{aligned}
\sum_{l=0}^{L} \frac{\sigma_l^2}{M_l} &\leq \frac{1}{4}\varepsilon^2 S^{-1} \sum_{l=0}^{L} h_l^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2} \\
&\leq \frac{1}{4}\varepsilon^2 S^{-1} \sum_{l=0}^{L} h_0^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2} \\
&\leq \frac{1}{4}\varepsilon^2 S^{-1} h_0^{-\gamma_1/2} \sum_{l=0}^{\infty} h_0^{\kappa_2(\tilde{q}_0+l\beta)/2} (\tilde{q}_0+l\beta)^{\gamma_1/2} \\
&\leq \frac{\varepsilon^2}{4}(S^{-1}S) = \frac{\varepsilon^2}{4}.
\end{aligned}$$

Here we used $\kappa_2\tilde{q}_l - \gamma_1 > 0$ for all $l = 0,\dots,L$, because $\kappa_2\tilde{q}_0 > \gamma_1$.

Next we want to derive a bound for the total work. Using Assumption (A1) we obtain

$$\sum_{l=0}^{L} M_l w_l \leq c_1 \sum_{l=0}^{L} M_l h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1}.$$

Using the definition of $M_l$ from (6.20) leads to

$$\begin{aligned}
\sum_{l=0}^{L} M_l w_l &\leq c_1 \sum_{l=0}^{L} M_l h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1} \\
&\leq c_1 \sum_{l=0}^{L} \Big(4\varepsilon^{-2} c_3 S h_l^{(\gamma_1+\kappa_2\tilde{q}_l)/2} \tilde{q}_l^{-\gamma_1/2} + 1\Big) h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1} \\
&\leq c_1 \Big(4\varepsilon^{-2} S c_3 \sum_{l=0}^{L} h_l^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2} + \sum_{l=0}^{L} h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1}\Big).
\end{aligned} \tag{6.22}$$

The first sum in (6.22) is bounded by the same arguments as before. For the second sum we first consider (6.19). After rearranging the lower bound in (6.19) we obtain

$$\begin{aligned}
\frac{1}{2}\varepsilon\delta < c_2 \lambda^{2\kappa_1\beta L_1} h_L^{\kappa_1\tilde{q}_L} &= c_2 \lambda^{2\kappa_1\beta L_1} \lambda^{-L\kappa_1\tilde{q}_L} h_0^{\kappa_1\tilde{q}_L} \\
&= c_2 \lambda^{2\kappa_1\beta L_1} \lambda^{-L\kappa_1\tilde{q}_0} \lambda^{-\kappa_1\beta L^2} h_0^{\kappa_1\tilde{q}_L} \\
&= c_2 \lambda^{\kappa_1\beta(2L_1-L^2)} \lambda^{-L\kappa_1\tilde{q}_0} h_0^{\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta L}.
\end{aligned} \tag{6.23}$$

Because $L \geq L_1$ and $L \geq 2$ it holds that $(2L_1 - L^2) \leq 0$ and thus $\lambda^{\kappa_1\beta(2L_1-L^2)} \leq 1$. Analogously we estimate $h_0^{\kappa_1\beta L} \leq 1$. Altogether we have

$$\frac{1}{2}\varepsilon\delta < c_2 h_L^{\kappa_1\tilde{q}_0},$$

and this yields

$$h_L^{-\gamma_1} < (2c_2\delta^{-1}\varepsilon^{-1})^{\gamma_1/\kappa_1\tilde{q}_0}. \tag{6.24}$$

On the other hand we have from (6.19)

$$\frac{1}{2}\varepsilon\delta\lambda^{-2\kappa_1\beta L_1} < c_2 h_L^{\kappa_1\tilde{q}_L} = c_2\lambda^{-L\kappa_1\tilde{q}_0}\lambda^{-\kappa_1\beta L^2}h_0^{\kappa_1\tilde{q}_L}.$$

Rearranging gives

$$\frac{1}{2}\varepsilon\delta\lambda^{\kappa_1\beta(L^2-2L_1)}\lambda^{L\kappa_1\tilde{q}_0} < c_2 h_0^{\kappa_1\tilde{q}_L}. \tag{6.25}$$

By the same arguments as above we derive the following lower bound

$$\frac{1}{2}\varepsilon\delta < c_2 h_0^{\kappa_1\tilde{q}_L}. \tag{6.26}$$

Thus, by rearranging (6.26) and taking the logarithm we obtain

$$\tilde{q}_L < \frac{\log(2c_2\delta^{-1}\varepsilon^{-1})}{\log(h_0^{-\kappa_1})} = \frac{\log(2c_2\delta^{-1})}{\log(h_0^{-\kappa_1})} + \frac{\log(\varepsilon^{-1})}{\log(h_0^{-\kappa_1})} =: \hat{c}_1 + \hat{c}_2\log\Big(\frac{1}{\varepsilon}\Big). \tag{6.27}$$

The leading term in (6.27) has logarithmic growth in $\varepsilon$, hence we can bound it by a term which grows algebraically, i.e. we find a constant $\hat{c}_3 > 0$, such that

$$\tilde{q}_L < \hat{c}_3\varepsilon^{\frac{-2+\frac{\gamma_1}{\kappa_1\tilde{q}_0}}{\gamma_1+1}}. \tag{6.28}$$

To have a negative exponent we need to have $\kappa_1\tilde{q}_0 > \frac{1}{2}\gamma_1$. This follows from Assumption (A2) because $\kappa_1\tilde{q}_0 \geq \frac{\kappa_2\tilde{q}_0}{2} > \frac{\gamma_1}{2}$. We then proceed to estimate

$$\sum_{l=0}^{L} h_l^{-\gamma_1}\tilde{q}_l^{\gamma_1} \leq h_L^{-\gamma_1}\tilde{q}_L^{\gamma_1}(L+1) \leq h_L^{-\gamma_1}\tilde{q}_L^{\gamma_1+1}, \tag{6.29}$$

where the last inequality follows from the fact that $(1+L) \leq \tilde{q}_0 + \beta L = \tilde{q}_L$. Now, using (6.24) and (6.28) yields with a constant $c = c(c_2, \hat{c}_3, \delta^{-1})$

$$\sum_{l=0}^{L} h_l^{-\gamma_1}\tilde{q}_l^{\gamma_1} \leq c\varepsilon^{-\frac{\gamma_1}{\kappa_1\tilde{q}_0}}\varepsilon^{-2+\frac{\gamma_1}{\kappa_1\tilde{q}_0}} = c\varepsilon^{-2}. \tag{6.30}$$

Thus, the first case follows.

**Second case:** $\kappa_2\tilde{q}_0 < \gamma_1$

Let $L^* \in \mathbb{N}$ be such that $\kappa_2\tilde{q}_{L^*} < \gamma_1$ and $\kappa_2\tilde{q}_{L^*+1} \geq \gamma_1$. For the second case we choose

$$M_l := \left\lceil 8\varepsilon^{-2}c_3 h_l^{(\gamma_1+\kappa_2\tilde{q}_l)/2}h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2}\big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\big)^{-1}\right\rceil, \tag{6.31}$$

for $l = 0, \dots, L^*$ and

$$M_l := \left\lceil 8\varepsilon^{-2} c_3 S h_l^{(\gamma_1 + \kappa_2 \tilde{q}_l)/2} \tilde{q}_l^{-\gamma_1/2} \right\rceil, \tag{6.32}$$

for $l = L^* + 1, \dots, L$. Here $S$ is the right hand side of (6.13) with $r = h_0^{\kappa_2/2}$ and $\tilde{q}_0$ replaced with $\tilde{q}_{L^*+1}$. We derive the following two bounds which we shall consider later. Let us assume that $L(\varepsilon)$ from (6.17) satisfies $L = L(\varepsilon) \leq L^*$. After rearranging (6.19) we have

$$h_L^{-\kappa_1 \tilde{q}_L} < 2c_2(\varepsilon\delta)^{-1} \lambda^{2\kappa_1 \beta L_1} \leq 2c_2 \delta^{-1} \lambda^{2\kappa_1 \beta L^*} \varepsilon^{-1}$$

Hence, we deduce

$$h_L^{-\gamma_1} < c(L^*)^{-\gamma_1/\kappa_1 \tilde{q}_L} \varepsilon^{-\gamma_1/\kappa_1 \tilde{q}_L}, \tag{6.33}$$

where we set $c(L^*) = 2c_2 \delta^{-1} \lambda^{2\kappa_1 \beta L^*}$. If $0 < c(L^*) < 1$ we estimate

$$c(L^*)^{-\gamma_1/\kappa_1 \tilde{q}_L} \leq c(L^*)^{-\gamma_1/\kappa_1 \tilde{q}_0} =: \tilde{c}(L^*),$$

and if $c(L^*) \geq 1$ we estimate

$$c(L^*)^{-\gamma_1/\kappa_1 \tilde{q}_L} \leq c(L^*)^{-\gamma_1/\kappa_1 \tilde{q}_{L^*}} =: \tilde{c}(L^*).$$

In both cases $\tilde{c}(L^*)$ is a constant independent of $L$.
If conversely $L^* \leq L - 1$ holds, we have

$$c_2 h_{L^*}^{\kappa_1 \tilde{q}_{L^*}} \geq c_2 h_{L-1}^{\kappa_1 \tilde{q}_{L-1}} \geq c_2 h_{L_1}^{\kappa_1 \tilde{q}_{L_1}} = c_2 \frac{\varepsilon}{2},$$

because $L - 1 \leq L_1$, cf. (6.17). It follows that

$$h_{L^*}^{-\gamma_1} < \varepsilon^{-\gamma_1/\kappa_1 \tilde{q}_{L^*}}. \tag{6.34}$$

Next, let us consider a splitting for the statistical error given by

$$\sum_{l=0}^{L} \frac{\sigma_l^2}{M_l} = \sum_{l=0}^{L^*} \frac{\sigma_l^2}{M_l} + \sum_{l=L^*+1}^{L} \frac{\sigma_l^2}{M_l}. \tag{6.35}$$

Note that in the case $L \leq L^*$, the second sum in (6.35) vanishes. We treat each term separately and start with the first sum in (6.35). Using the definition (6.31) of $M_l$ we have

$$\begin{aligned} \sum_{l=0}^{L^*} \frac{\sigma_l^2}{M_l} &\leq \frac{1}{8} \varepsilon^2 h_{L^*}^{(\gamma_1 - \kappa_2 \tilde{q}_0)/2} (1 - \lambda^{-(\gamma_1 - \kappa_2 \tilde{q}_0)/2}) \sum_{l=0}^{L^*} h_l^{(\kappa_2 \tilde{q}_l - \gamma_1)/2} \\ &\leq \frac{1}{8} \varepsilon^2 h_{L^*}^{(\gamma_1 - \kappa_2 \tilde{q}_0)/2} (1 - \lambda^{-(\gamma_1 - \kappa_2 \tilde{q}_0)/2}) \sum_{l=0}^{L^*} h_l^{(\kappa_2 \tilde{q}_0 - \gamma_1)/2}. \end{aligned}$$

For the last inequality we used the fact that $h_l^{(\kappa_2\tilde{q}_l-\gamma_1)/2} < h_l^{(\kappa_2\tilde{q}_0-\gamma_1)/2}$ holds as long as $0 > \kappa_2\tilde{q}_l - \gamma_1 > \kappa_2\tilde{q}_0 - \gamma_1$. We proceed to estimate

$$\begin{aligned}
\sum_{l=0}^{L^*} \frac{\sigma_l^2}{M_l} &\leq \frac{1}{8}\varepsilon^2 h_{L^*}^{(\gamma_1-\kappa_2\tilde{q}_0)/2}(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}) \sum_{l=0}^{L^*} (\lambda^{L^*-l} h_{L^*})^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \\
&= \frac{1}{8}\varepsilon^2 h_{L^*}^{(\gamma_1-\kappa_2\tilde{q}_0)/2}(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}) \sum_{l=0}^{L^*} (\lambda^{l} h_{L^*})^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \\
&\leq \frac{1}{8}\varepsilon^2 h_{L^*}^{(\gamma_1-\kappa_2\tilde{q}_0)/2} h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2}(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}) \sum_{l=0}^{\infty} (\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2})^l \\
&\leq \frac{1}{8}\varepsilon^2.
\end{aligned}$$

In the next step we estimate the second sum in (6.35) and use the definition (6.32) of $M_l$ to derive

$$\begin{aligned}
\sum_{l=L^*+1}^{L} \frac{\sigma_l^2}{M_l} &\leq \frac{1}{8}\varepsilon^2 S^{-1} \sum_{l=L^*+1}^{L} h_l^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2} \leq \frac{1}{8}\varepsilon^2 S^{-1} \sum_{l=L^*+1}^{L} h_0^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2} \\
&\leq \frac{1}{8}\varepsilon^2 S^{-1} \sum_{l=0}^{L-(L^*+1)} h_0^{(\kappa_2\tilde{q}_{(L^*+1)+l}-\gamma_1)/2} \tilde{q}_{(L^*+1)+l}^{\gamma_1/2} \leq \frac{1}{8}\varepsilon^2, \qquad (6.36)
\end{aligned}$$

which yields altogether

$$\sum_{l=0}^{L} \frac{\sigma_l^2}{M_l} \leq \frac{1}{4}\varepsilon^2.$$

Similarly to the previous sum (6.35) we rewrite the total work as

$$\sum_{l=0}^{L} M_l w_l = \sum_{l=0}^{L^*} M_l w_l + \sum_{l=L^*+1}^{L} M_l w_l \qquad (6.37)$$

and proceed with the first sum.

$$\begin{aligned}
\sum_{l=0}^{L^*} M_l w_l \leq c_1 \sum_{l=0}^{L^*} \Big( & 8\varepsilon^{-2} c_3 h_l^{(\gamma_1+\kappa_2\tilde{q}_l)/2} h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)^{-1} + 1 \Big) \\
& \times h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1}. \qquad (6.38)
\end{aligned}$$

Starting with the first sum in (6.38) we estimate

$$\begin{aligned}
&\sum_{l=0}^{L^*} c_3 c_1 8\varepsilon^{-2} h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)^{-1} h_l^{-(\gamma_1-\kappa_2\tilde{q}_l)/2} \tilde{q}_l^{\gamma_1} \\
&\leq c_3 c_1 8\varepsilon^{-2} h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)^{-1} \sum_{l=0}^{L^*} (\lambda^l h_{L^*})^{-(\gamma_1-\kappa_2\tilde{q}_0)/2} \tilde{q}_l^{\gamma_1} \\
&= c_3 c_1 8\varepsilon^{-2} h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)} \Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)^{-1} \sum_{l=0}^{L^*} (\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2})^l (\tilde{q}_0+\beta l)^{\gamma_1}.
\end{aligned}$$

By the ratio test the series $\sum_{l=0}^{L}(\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2})^l(\tilde{q}_0+\beta l)^{\gamma_1}$ converges as $L\to\infty$. Hence,

$$\sum_{l=0}^{L^*} c_3c_1 8\varepsilon^{-2}h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2}\Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)h_l^{-(\gamma_1-\kappa_2\tilde{q}_l)/2}\tilde{q}_l^{\gamma_1} \le c\varepsilon^{-2}h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)},$$

where $c := c_3c_1 8\Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)^{-1}\sum_{l=0}^{\infty}(\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2})^l(\tilde{q}_0+\beta l)^{\gamma_1}$. If $L^*\le L-1$ we use (6.34) to end up with

$$\begin{aligned}\sum_{l=0}^{L^*} c_3c_1 8\varepsilon^{-2}h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2}\Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)h_l^{-(\gamma_1-\kappa_2\tilde{q}_l)/2}\tilde{q}_l^{\gamma_1}&\\ \le c\varepsilon^{-2}h_{L^*}^{-(\gamma_1-\kappa_2\tilde{q}_0)}&\\ \le c\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L^*}}}.&\end{aligned}$$

If $L^*\ge L$ holds we use (6.33) to obtain

$$\begin{aligned}\sum_{l=0}^{L} c_3c_1 8\varepsilon^{-2}h_{L}^{-(\gamma_1-\kappa_2\tilde{q}_0)/2}\Big(1-\lambda^{-(\gamma_1-\kappa_2\tilde{q}_0)/2)}\Big)h_l^{-(\gamma_1-\kappa_2\tilde{q}_l)/2}\tilde{q}_l^{\gamma_1}&\\ \le c\tilde{c}(L^*)\varepsilon^{-2}h_{L}^{-(\gamma_1-\kappa_2\tilde{q}_0)}&\\ \le c\tilde{c}(L^*)\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L}}}.&\end{aligned}$$

Returning to the total work estimate, for the second sum in (6.38) we use (6.29) to obtain

$$\sum_{l=0}^{L^*} h_l^{-\gamma_1}\tilde{q}_l^{\gamma_1} \le h_{L^*}^{-\gamma_1}\tilde{q}_{L^*}^{\gamma_1+1}.$$

By the same argument as in (6.26)-(6.28) we find a constant $\hat{c}_4$ such that,

$$\tilde{q}_{L^*} < \hat{c}_4\varepsilon^{\frac{-2+\frac{\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L^*}}}{\gamma_1+1}}. \tag{6.39}$$

In this case we need $\kappa_1\tilde{q}_{L^*} > \frac{\kappa_2\tilde{q}_0}{2}$, which follows from Assumption (A2) because $\kappa_1\tilde{q}_{L^*} > \kappa_1\tilde{q}_0 \ge \frac{\kappa_2\tilde{q}_0}{2}$. Upon using (6.34) and (6.39) we end up with

$$\sum_{l=0}^{L^*} M_l w_l \le h_{L^*}^{-\gamma_1}\tilde{q}_{L^*}^{\gamma_1+1} \le \hat{c}_4\tilde{c}\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L^*}}},$$

which holds for $L^*\le L-1$. For $L\le L^*$ we obtain

$$\sum_{l=0}^{L} M_l w_l \le h_{L}^{-\gamma_1}\tilde{q}_{L}^{\gamma_1+1} \le \hat{c}_4\tilde{c}(L^*)\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\kappa_1\tilde{q}_{L}}}.$$

We combine the previous cases by writing

$$\sum_{l=0}^{L^*} M_l w_l \leq h_{L^*}^{-\gamma_1} \tilde{q}_{L^*}^{\gamma_1+1} \leq \hat{c}_4 \tilde{c}(L^*) \varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\min\{\kappa_1\tilde{q}_L,\kappa_1\tilde{q}_{L^*}\}}}.$$

For the second sum in (6.37) we have

$$\begin{aligned}\sum_{l=L^*+1}^{L} M_l w_l &\leq c_1 \sum_{l=L^*+1}^{L} \Big(8\varepsilon^{-2} c_3 S h_l^{(\gamma_1+\kappa_2\tilde{q}_l)/2} \tilde{q}_l^{-\gamma_1/2} + 1\Big) h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1} \\ &= 8c_1c_3S\varepsilon^{-2} \sum_{l=L^*+1}^{L} \Big(h_l^{(\kappa_2\tilde{q}_l-\gamma_1)/2} \tilde{q}_l^{\gamma_1/2}\Big) + c_1 \sum_{l=L^*+1}^{L} h_l^{-\gamma_1} \tilde{q}_l^{\gamma_1}.\end{aligned}$$

Upon using an index shift, the estimates (6.36) and (6.30) yield

$$\sum_{l=L^*+1}^{L} M_l w_l \leq c\varepsilon^{-2} \leq c\varepsilon^{-2-\frac{\gamma_1-\kappa_2\tilde{q}_0}{\min\{\kappa_1\tilde{q}_L,\kappa_1\tilde{q}_{L^*}\}}},$$

which concludes the proof. □

**Remark 6.8** (Choice of the maximum level)**.**
*The number of levels L in Algorithm hp-MLMC can be computed a priori using* (6.17)*. It is also possible to compute L on the fly. To this end we consider, similarly as in [51, 81],*

$$\begin{aligned}\|\mathbb{E}(u) - \mathbb{E}(u_L)\|_{L^2(D)} &= \|\sum_{l=L+1}^{\infty} (\mathbb{E}(u_l) - \mathbb{E}(u_{l-1}))\|_{L^2(D)} \\ &\leq \|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)} \sum_{l=L+1}^{\infty} \frac{\|\mathbb{E}(u_l) - \mathbb{E}(u_{l-1})\|_{L^2(D)}}{\|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)}}.\end{aligned}$$

*Assuming that* $\|\mathbb{E}(u_l) - \mathbb{E}(u_{l-1})\|_{L^2(D)} \leq c_2 h_l^{\kappa_1\tilde{q}_l}$ *we obtain*

$$\begin{aligned}\frac{\|\mathbb{E}(u_l) - \mathbb{E}(u_{l-1})\|_{L^2(D)}}{\|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)}} \approx \frac{h_l^{\kappa_1\tilde{q}_l}}{h_L^{\kappa_1\tilde{q}_L}} &= (\lambda^{-l} h_0)^{\kappa_1(\tilde{q}_0+\beta l)} (\lambda^{-L} h_0)^{-\kappa_1(\tilde{q}_0+\beta L)} \\ &= \lambda^{(L-l)\kappa_1\tilde{q}_0} \lambda^{\kappa_1\beta(L^2-l^2)} h_0^{\kappa_1\beta(l-L)} \\ &\leq \lambda^{(L-l)\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta(l-L)},\end{aligned}$$

*where we used that* $\lambda^{\kappa_1\beta(L^2-l^2)} \leq 1$ *as* $l > L$ *and* $\lambda \geq 2$*. Thus,*

$$\begin{aligned}\|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)} &\sum_{l=L+1}^{\infty} \frac{\|\mathbb{E}(u_l) - \mathbb{E}(u_{l-1})\|_{L^2(D)}}{\|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)}} \\ &\leq \|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)} \sum_{l=1}^{\infty} (\lambda^{-\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta})^l \\ &= \|\mathbb{E}(u_L) - \mathbb{E}(u_{L-1})\|_{L^2(D)} \frac{\lambda^{-\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta}}{1 - (\lambda^{-\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta})}.\end{aligned}$$

*Therefore, the condition to add new levels is*

$$\max_{j\in\{0,1,2\}} \frac{(\lambda^{-\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta})^{(j+1)}}{1-(\lambda^{-\kappa_1\tilde{q}_0} h_0^{\kappa_1\beta})} \|\mathbb{E}\big(u_{L-j}\big) - \mathbb{E}\big(u_{L-j-1}\big)\|_{L^2(D)} \leq \frac{1}{2}\varepsilon. \tag{6.40}$$

*This criterion ensures that the deterministic error approximated by an extrapolation from the three finest meshes is within the desired range, cf. [81]. For the modified hp-MLMC algorithm which we introduce in Section 6.2.3 we add new levels adaptively based on* (6.40)*.*

### 6.2.3 Confidence intervals for the number of additional samples

In this section we discuss the computation of the optimal number of samples $M_l$ based on confidence intervals, having in mind the use of queue-based HPC systems. In most modern large-scale computing systems, access to compute nodes is based on job schedulers. For execution of a job, a certain number of nodes can be requested for a specified time slot. The job is executed after some queuing time, which can be much longer than the actual job execution time. In the context of *hp*-MLMC, it is advisable to submit a new job to the queue for each iteration of Algorithm *hp*-`MLMC`, since otherwise idle times of the compute nodes are very difficult to avoid. As each iteration requires its own queuing time, it is our aim to compute as many new samples as possible during one iteration of Algorithm *hp*-`MLMC`. On the other hand, we want to avoid computing more samples than optimal, as this would decrease the efficiency of the *hp*-MLMC method. Moreover, from an economical point of view wasted computing time is expensive. In that sense, one is facing two competing issues. The first issue is to reduce queuing time by computing as many samples as possible per iteration. The second issue is to reduce the number of unnecessarily computed samples, i.e. saving computing time.

A straightforward approach to satisfy the first issue is to rely on the standard estimator $\hat{M}_l$. However, this approach contradicts the second aim of saving computing time. Let us recall that the quantities $w_l$ and $\sigma_l$ in (6.7) and (6.8) are not known exactly but are estimated by $\hat{w}_l, \hat{\sigma}_l$. In most cases the number of samples in the warm-up phase and after the first iteration of the *hp*-MLMC algorithm is too small to obtain a reliable estimate $\hat{M}_l$ of the optimal number of samples $M_l$. This wrong estimate may then lead to an severe overestimation of $M_l$ (see for example Figure 6.3(a)) and thus spoils the goal of avoiding the computation of unnecessary samples and saving computing time.

**Remark 6.9.**

*We note that our arguments and our proposed method of course also apply to sequential computing, where it is also desirable to avoid overestimating the optimal number of samples, while*

*keeping the iteration number (and associated post-processing costs) to a minimum. It becomes however especially important in the context of HPC systems, where large queuing times are to be averted.*

In order to satisfy the second goal of saving computing time, we want to properly account for the fact that $M_l$ is only estimated by constructing a confidence interval for $M_l$. More specifically, we construct a one-sided confidence interval $\mathrm{I}_{M_l} = [\underline{M_l}, \infty)$, such that $\mathbb{P}(M_l \in \mathrm{I}_{M_l}) \geq 1-\alpha$. To derive the desired confidence interval we construct corresponding one-sided confidence intervals for $\sigma_l$, $w_l$ denoted by

$$\mathrm{I}_{\underline{\sigma_l}} = [\underline{\sigma_l}, \infty), \quad \mathrm{I}_{\underline{w_l}} = [\underline{w_l}, \infty), \quad \mathrm{I}_{\overline{w_l}} = (-\infty, \overline{w_l}],$$

respectively. As we do not have any information about the underlying distributions of $\sigma_l$ and $w_l$ we construct the confidence interval based on asymptotic confidence intervals. Hence our approach is heuristic because the Central Limit Theorem implies that the number of samples needs to be sufficiently large to ensure that the estimators are asymptotically normally distributed. This seems to be a contradiction to the fact that we choose a small number of samples for the warm-up phase. However, we show in estimate (6.44) that our construction of the confidence interval is very conservative and yields a robust lower estimate on the optimal number of samples, although the number of warm-up samples is of order $\mathscr{O}(1)$. Indeed, we never overestimated the optimal number of samples in our computations, justifying our approach.

For the construction of the confidence interval for $\sigma_l$ we use the method described in [5, Formula (6)], which employs an adjustment to the degrees of freedom of the $\chi^2$-distribution. More precisely, we let

$$\begin{aligned}
\hat{r}_l &:= \frac{2M_{\mathrm{tot}_l}}{\hat{\gamma}_{e_l} + \left(\frac{2M_{\mathrm{tot}_l}}{M_{\mathrm{tot}_l}-1}\right)}, \\
\hat{\gamma}_{e_l} &= \frac{M_{\mathrm{tot}_l}(M_{\mathrm{tot}_l}+1)}{(M_{\mathrm{tot}_l}-1)(M_{\mathrm{tot}_l}-2)(M_{\mathrm{tot}_l}-3)} \frac{\hat{\mu}_l^4}{\hat{\sigma}_l^4} - \frac{3(M_{\mathrm{tot}_l}-1)^2}{(M_{\mathrm{tot}_l}-2)(M_{\mathrm{tot}_l}-3)},
\end{aligned}$$

where $\hat{\mu}_l^4 := \| \sum_{i=1}^{M_{\mathrm{tot}_l}} \Big( (u_l^i - u_{l-1}^i) - \frac{1}{M_{\mathrm{tot}_l}} \sum_{j=1}^{M_{\mathrm{tot}_l}} (u_l^j - u_{l-1}^j) \Big)^4 \|_{L^2(D)}$. For the lower confidence interval $\mathrm{I}_{\underline{\sigma_l}} = [\underline{\sigma_l}, \infty)$ we therefore define

$$\underline{\sigma_l} := \sqrt{\frac{\hat{r}_l \hat{\sigma}_l^2}{\chi^2_{1-\frac{\alpha}{2},\hat{r}_l}}}, \tag{6.41}$$

where for some $\alpha \in (0,1)$, $\chi^2_{1-\frac{\alpha}{2},\hat{r}_l}$ is the $(1-\frac{\alpha}{2})$-quantile of the $\chi^2$-distribution for $\hat{r}_l$ degrees of freedom. If the random samples are normally distributed, it follows that $\hat{\gamma}_{e_l} = 0$ and thus $\hat{r}_l = M_{\mathrm{tot}_l} - 1$, i.e. we obtain the standard confidence interval for the variance of a normal

distribution (cf. [5]). For $w_l$ we compute the confidence interval using the standard asymptotic confidence interval for the mean, i.e.

$$\underline{w_l} := \hat{w}_l - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{w_l}}{\sqrt{M_{\mathrm{tot}_l}}}, \quad \overline{w_l} := \hat{w}_l + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{w_l}}{\sqrt{M_{\mathrm{tot}_l}}}, \tag{6.42}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the normal distribution and $\hat{\sigma}^2_{w_l}$ is the unbiased estimator for the variance of $w_l$. We then define

$$\underline{M_l} = \frac{1}{\varepsilon^2} \frac{\sigma_l}{\sqrt{\overline{w_l}}} \left( \sum_{k=0}^{L} \underline{\sigma_k} \sqrt{\underline{w_k}} \right) \tag{6.43}$$

and the confidence interval $\mathrm{I}_{\underline{M_l}} := [\underline{M_l}, \infty)$. Moreover, for $l = 0, \dots, L$ we define the events

$$\begin{aligned}
X_l &:= \{M_l \in \mathrm{I}_{\underline{M_l}}\}, \\
\Sigma_{\varepsilon,l,\mathrm{lower}} &:= \left\{ \frac{1}{\varepsilon^2} \sigma_l \in \mathrm{I}_{\frac{1}{\varepsilon^2}\underline{\sigma_l}} \right\}, \quad \mathrm{I}_{\frac{1}{\varepsilon^2}\underline{\sigma_l}} := [\frac{1}{\varepsilon^2}\underline{\sigma_l}, \infty), \\
\Sigma_{l,\mathrm{lower}} &:= \{\sigma_l \in \mathrm{I}_{\underline{\sigma_l}}\}, \\
W_{l,\mathrm{lower}} &:= \{\sqrt{w_l} \in \mathrm{I}_{\sqrt{\underline{w_l}}}\}, \quad \mathrm{I}_{\sqrt{\underline{w_l}}} := [\sqrt{\underline{w_l}}, \infty) \\
W_{l,\mathrm{upper}} &:= \{\sqrt{w_l} \in \mathrm{I}_{\sqrt{\overline{w_l}}}\}, \quad \mathrm{I}_{\sqrt{\overline{w_l}}} := (-\infty, \sqrt{\overline{w_l}}].
\end{aligned}$$

It then follows that $Y_l \subseteq X_l$, with

$$Y_l := \bigcap_{k=0}^{L} \left( \Sigma_{k,\mathrm{lower}} \cap W_{k,\mathrm{lower}} \cap W_{l,\mathrm{upper}} \cap \Sigma_{\varepsilon,l,\mathrm{lower}} \right),$$

for all $l = 0, \dots, L$. Using elementary probability estimates and De Morgan's rule we estimate

$$\begin{aligned}
\mathbb{P}(X_l) \geq \mathbb{P}(Y_l) &= 1 - \mathbb{P}(Y_l^c) \\
&\geq 1 - \sum_{k=0}^{L} \left( (\mathbb{P}(\Sigma^c_{k,\mathrm{lower}}) + \mathbb{P}(W^c_{k,\mathrm{lower}}) + \mathbb{P}(W^c_{l,\mathrm{upper}}) + \mathbb{P}(\Sigma^c_{\varepsilon,l,\mathrm{lower}}) \right).
\end{aligned} \tag{6.44}$$

We construct the confidence intervals $\mathrm{I}_{\underline{\sigma_k}} = [\underline{\sigma_k}, \infty)$, $\mathrm{I}_{\sqrt{\underline{w_k}}} = [\sqrt{\underline{w_k}}, \infty)$ and $\mathrm{I}_{\sqrt{\overline{w_l}}} = (-\infty, \sqrt{\overline{w_l}}]$ such that

$$\mathbb{P}(\Sigma_{\varepsilon,l,\mathrm{lower}}) = \mathbb{P}(\Sigma_{k,\mathrm{lower}}) = \mathbb{P}(W_{k,\mathrm{lower}}) = \mathbb{P}(W_{l,\mathrm{upper}}) = 1 - \frac{\alpha}{4L} \tag{6.45}$$

for some $\alpha \in (0,1)$. This choice yields $\mathbb{P}(X_l) \geq 1 - \alpha$, for all $l = 0, \dots, L$.

Whereas the lower confidence bound $\underline{M_l}$ helps in saving computing time, it increases the number of queuing operations. Therefore, in order to balance the two competing issues of either reducing queuing time or saving computing time we introduce an additional parameter $\zeta \in [0,1]$ and let

$$\tilde{M}_l := \zeta \underline{M_l} + (1-\zeta)\hat{M}_l. \tag{6.46}$$

By setting $\zeta = 0$ we pursue the aim of reducing queuing time and by setting $\zeta = 1$ we try to save computing time. Choosing $\zeta \in (0,1)$ corresponds to finding a strategy in between. Moreover, it is possible to choose $\zeta = \zeta_l^{\text{iter}_l}$ iteration-dependent, such that each level $l = 0,\dots,L$ has its own iteration counter $\text{iter}_l$. We choose the following strategy which tries to realize a trade-off between reducing queuing time and saving computing time:

$$\zeta_l^{\text{iter}_l} = \begin{cases} 1, & \text{for iter}_l = 1, \\ 0.5, & \text{for iter}_l = 2, \\ 0, & \text{for iter}_l > 2. \end{cases} \tag{6.47}$$

In the first iteration we rely on the lower confidence bound $\underline{M_l}$ from (6.43) to avoid overestimating $M_l$. In the second iteration we start to approach $\hat{M}_l$ but still consider $\underline{M_l}$ as a safe-guard for overestimating $M_l$. After the second iteration, where $M_{\text{tot}_l}$ is sufficiently large, we trust the estimate $\hat{M}_l$. It might happen that during the first two iterations we have $\underline{M_l} \leq M_{\text{tot}_l} < \hat{M}_l$. In that case we set $\zeta_l^{\text{iter}_l} = 0$.

The modified $hp$-MLMC method which adds new levels adaptively based on the bias estimate (6.40), is summarized in Algorithm $hp$-`MLMC with confidence intervals`. Let us comment on Algorithm $hp$-`MLMC with confidence intervals`. Initially the algorithm considers three levels. As long as $M_{\text{tot}_l} < \hat{M}_l$ the algorihm adds samples according to the strategy defined in (6.47). After looping over all levels, the algorithm checks if the statistical tolerance is met (line 17). If the statistical tolerance is met it checks if the bias criterion is also satisfied (line 18). If the bias criterion is not satisfied it adds an additional level (line 21) and computes additional warm-up samples $K_{L+1}$. After this step the iteration is complete and the algorithm restarts from line 4. The algorithm terminates when the sum of the estimated statistical error and the estimated bias term is smaller than $\varepsilon$.

Thanks to this safety mechanism we avoid computing unnecessary samples and improve the efficiency of the plain $hp$-`MLMC` algorithm, while trying to reduce the overall queuing time. For the number of warm-up samples $K_l$ we typically choose ten to one hundred samples on the coarse levels and three samples on the fine levels. When we add a new level we set the number of warm-up samples also to three.

## 6.3 $hp$-MLMC for compressible Navier–Stokes equations

In this section we apply the $hp$-MLMC method for UQ of compressible flows described by the two-dimensional Navier–Stokes equations. Our application in mind is a problem arising from

**Algorithm** $hp$-MLMC with confidence intervals

Fix a tolerance $\varepsilon > 0$, set $L = 2$ set $\mathscr{L} := \{0, \dots, L\}$
Compute $K_l$ (warm-up) samples on $l = 0, \dots, L$, set $M_{\text{tot}_l} := K_l$ and $\text{iter}_l = 1$
**while** $\mathscr{L} \neq \emptyset$ **do**
  **for** $l \in \mathscr{L}$ **do**
    Estimate $\hat{w}_l$ by (6.12), $\hat{\sigma}_l^2$ by (6.11) and then $\hat{M}_l$ by (6.10)
    Estimate $\underline{w_l}$, $\overline{w_l}$ by (6.42), $\underline{\sigma_l^2}$ by (6.41) and then $\underline{M_l}$ by (6.43)
    **if** $M_{\text{tot}_l} < \hat{M}_l$ **then**
      Set $\zeta_l^{\text{iter}_l}$ according to (6.47) and compute $\tilde{M}_l$ by (6.46)
      Add $\lceil \tilde{M}_l - M_{\text{tot}_l} \rceil$ new samples of $u_l^j - u_{l-1}^j$
      Set $M_{\text{tot}_l} := M_{\text{tot}_l} + \lceil \tilde{M}_l - M_{\text{tot}_l} \rceil$
      Set $\text{iter}_l := \text{iter}_l + 1$
    **else**
      Set $\text{iter}_l := \text{iter}_l + 1$
      Skip level $l$
    **end if**
  **end for**
  **if** the statistical tolerance (6.9) is satisfied **then**
    **if** the bias term satisfies (6.40) **then**
      Set $\mathscr{L} := \emptyset$
    **else**
      Set $\mathscr{L} := \mathscr{L} \cup \{L+1\}$
      Compute $K_{L+1}$ (warm-up) samples and set $M_{\text{tot}_{L+1}} := K_{L+1}$
      Set $\text{iter}_{L+1} = 1$
      Set $L := L + 1$
    **end if**
  **end if**
**end while**
Compute $\mathrm{E}_{hp}^{L}\left[u_L\right]$

computational acoustics, where we consider the flow over an open cavity. Due to different noise generation mechanisms, like Rossiter feedback, or Helmholtz resonance, the cavity emits tonal noise which is very sensitive with respect to different parameters, cf. [73, 74] for a detailed description of the different noise generation mechanisms. We apply the $hp$-MLMC method to quantify the influence of uncertain parameters, for example an uncertain Mach number, on the development of tonal noise.

For the physical space we consider $D \subset \mathbb{R}^2$, to be a bounded spatial domain. We further define the space-time-stochastic domain $D_{T,\Xi} := (0,T) \times D \times \Xi$. Our equations of interest are the following random Navier–Stokes equations:

$$u_t + \nabla_x \cdot (G(u) - H(u, \nabla_x u)) = 0, \qquad \forall\, (t,x,y) \in D_{T,\Xi}, \tag{6.48}$$

where $u(t,x,y)$ denotes the solution vector of the conserved quantities, i.e. $u = (\rho, \rho v_1, \rho v_2, \rho e)^\top$, $G(u)$ and $H(u, \nabla_x u)$ are the advective and viscous fluxes, i.e.

$$G_i(u) = \begin{pmatrix} \rho\, v_i \\ \rho\, v_1 v_i + \delta_{1i}\, p \\ \rho\, v_2 v_i + \delta_{2i}\, p \\ \rho\, e v_i + p\, v_i \end{pmatrix}, \quad H_i(u, \nabla_x u) = \begin{pmatrix} 0 \\ \tau_{1i} \\ \tau_{2i} \\ \tau_{ij} v_j - q_i \end{pmatrix}, \quad i = 1,2. \tag{6.49}$$

Here, $\delta_{ij}$ is the Kronecker delta function and the physical quantities $\rho$, $v = (v_1, v_2)^\top$, $p$, and $e$ represent density, the velocity vector, the pressure and the specific total energy, respectively. With Stokes' and Fourier's hypothesis, the viscous stress tensor $\tau$ and the heat flux $q$ are given by:

$$\tau := \mu(\nabla v + (\nabla v)^\top - \frac{2}{3}(\nabla \cdot v) I), \quad q = -k \nabla \mathscr{T}, \tag{6.50}$$

with $\mu$ being the dynamic viscosity, $k$ the thermal conductivity and $\mathscr{T}$ the local temperature. In order to solve for the unknowns, the system has to be closed by an equation of state. We choose for the gas constant $R$, the adiabatic exponent $\gamma$ and specific heat at constant volume $c_v$ the perfect gas law assumption

$$p = \rho R \mathscr{T} = (\gamma - 1)\rho(e - \frac{1}{2} v \cdot v), \qquad e = \frac{1}{2} v \cdot v + c_v \mathscr{T}. \tag{6.51}$$

We augment (6.48) with suitable boundary and initial conditions, denoted by

$$B(u) = g, \quad \forall (x,y) \in \partial D \times \Xi$$

and

$$u(0,x,y) = u^0(x,y), \quad \forall (x,y) \in D \times \Xi.$$

The boundary operator $B$, the boundary data $g$ and the initial condition $u^0$ will be specified when we describe the numerical experiments in Section 6.4. We call $u \in L^2_w(\Xi; C^1([0,T]; L^2(D)))$ a weak random solution of (6.48), if it is a weak solution $\tilde{\mathbb{P}}$-a.s. $y \in \Xi$ and a measurable mapping

$$\big(\Xi, \mathscr{B}(\Xi)\big) \ni y \to u(\cdot,\cdot,y) \in \big(C^1([0,T]; L^2(D)), \mathscr{B}(C^1([0,T]; L^2(D)))\big).$$

**Remark 6.10.**

*We have to assume the existence and uniqueness of weak random solutions of* (6.48)*, since there are no well-posedness results for* (6.48) *available.*

## 6.4 Numerical experiments

In this section we present numerical results for the *hp*-MLMC method as introduced in Algorithm *hp*-`MLMC with confidence intervals`. In Section 6.4.1, we apply *h*-, *p*-, *hp*-MLMC and the plain MC method to a smooth benchmark problem to verify the assumptions of Theorem 6.4. In Section 6.4.2, we apply the *h*-, *p*- and *hp*-MLMC method to an open cavity flow problem, an important flow problem from computational acoustics. We give a detailed comparison of all methods and verify that for both problems, *h*-, *p*- and *hp*-MLMC yield an optimal asymptotic work. This shows that the *hp*-MLMC method is applicable for UQ of complex engineering problems in computational fluid dynamics.

The numerical solver relies on the Discontinuous Galerkin Spectral Element solver FLEXI [63]. The time-stepping uses a Runge–Kutta method of order four with a time-step restriction of the form

$$\Delta t \leq \min\Big\{ \frac{h_{min}}{\lambda^c_{max}(2q+1)}, \Big(\frac{h_{min}}{\lambda^v_{max}(2q+1)}\Big)^2 \Big\}, \tag{6.52}$$

where (6.52) $\lambda^c_{max} := \big((|v_1|+c)+(|v_2|+c)\big)$ is an estimate for the absolute value of the largest eigenvalue of the convective flux Jacobian with $c := \sqrt{\gamma \frac{p}{\rho}}$ being the speed of sound. Moreover, $\lambda^v_{max} := \Big(\max\Big(\frac{4}{3\rho}, \frac{\gamma}{p}\Big)\frac{\mu}{Pr}\Big)$, is an estimate of the largest eigenvalue of the diffusion matrix of the viscous flux, where $Pr = \frac{c_p \mu}{k}$ is the Prandtl number. The viscous fluxes normal to the cell interfaces are approximated by the procedure described by Bassi and Rebay [10] and the numerical flux is chosen to be the Roe Riemann solver with entropy fix [59]. For the following numerical experiments we let $\mathrm{DOF}_l := (h_l^{-1} q_l)^2$. All computations for the open cavity were performed on Cray XC40 at the High-Performance Computing Center Stuttgart.

### 6.4.1 Smooth benchmark solution

In this numerical example we verify Theorem 6.4 by means of a smooth manufactured solution, given by

$$\begin{pmatrix} \rho(t,x_1,x_2,y) \\ \rho v_1(t,x_1,x_2,y) \\ \rho v_2(t,x_1,x_2,y) \\ E(t,x_1,x_2,y)) \end{pmatrix} = \begin{pmatrix} 2+A\sin(4\pi((x_1+x_2)-ft)) \\ 2+A\sin(4\pi((x_1+x_2)-ft)) \\ 2+A\sin(4\pi((x_1+x_2)-ft)) \\ \big(2+A\sin(4\pi((x_1+x_2)-ft))\big)^2 \end{pmatrix}. \tag{6.53}$$

The benchmark solution (6.53) is obtained by introducing an additional source term in (6.48). We choose the amplitude $A$ and frequency $f$ of (6.53) to be uncertain, i.e. we let $A \sim \mathscr{U}(0.1, 0.9)$

and $f \sim \mathcal{N}(1, 0.05^2)$. The spatial domain is $D = (-1,1)^2$ and we consider periodic boundary conditions. The setup of the mesh hierarchies for MC, $h$-MLMC, $p$-MLMC and $hp$-MLMC can be found in Table 6.1. The final computational time for this example is $T = 1$ and the QoI in this example is the momentum in $x_1$-direction at final time $T$, i.e.

$$Q(u) = (\rho v_1)(T, \cdot, \cdot, \cdot).$$

For the confidence intervals in (6.45) we set $\alpha$ to be 0.025.

| | MC | | $h$-MLMC | | $p$-MLMC | | $hp$-MLMC | |
|---|---|---|---|---|---|---|---|---|
| level | $N_l$ | $q_l$ | $N_l$ | $q_l$ | $N_l$ | $q_l$ | $N_l$ | $q_l$ |
| 0 | 256 | 6 | 16 | 3 | 64 | 2 | 16 | 4 |
| 1 | – | – | 64 | 3 | 64 | 3 | 64 | 5 |
| 2 | – | – | 256 | 3 | 64 | 4 | 256 | 6 |
| 3 | – | – | 1024 | 3 | 64 | 5 | 1024 | 7 |

Table 6.1: Level setup for MC, $h$-, $p$- and $hp$-MLMC. Example 6.4.1.

In Figure 6.1(a) we plot the estimated bias, i.e. the quantity $\|\mathbb{E}(u_l) - \mathbb{E}(u_{l-1})\|_{L^2(D)}$ from Remark 6.8. The quantities $\mathbb{E}(u_l)$ and $\mathbb{E}(u_{l-1})$ have been estimated by the standard MC estimator (6.3). $hp$-MLMC yields the smallest bias compared to $h$- and $p$-MLMC but this is not surprising since $hp$-MLMC admits the finest resolution for all levels. In Figure 6.1(b) we plot the bias term vs. $h_l^{\tilde{q}_l}$ in a log-log plot. This allows us to estimate $\kappa_1$ from Assumption (A2) using a linear fit. For $h$-MLMC, where we consider a DG polynomial degree of three, we estimate $\kappa_1 \approx 0.9$. Since we expect the bias term to converge with a rate of four (cf. Remark 6.3 2.) we compute from $4 = \kappa \tilde{q}_0$ a rate of convergence (in terms of mesh width $h$) of $h$-MLMC of approximately 3.6.

Figure 6.1(c) shows the estimated level variances $\hat{\sigma}_l^2$ and its 95% confidence interval. We observe that for $hp$-MLMC the level variance decays faster than for $h$- and $p$-MLMC. To verify the assumptions of Theorem 6.4 we estimate $\kappa_2$ from Assumption (A3) in Figure 6.1(d). For $h$-MLMC we estimate $\kappa_2 \approx 1.35$. Since we expect the level variance to decay with a rate of eight (cf. Remark 6.3 3.) we compute from $8 = \kappa_2 \tilde{q}_0$ a rate of convergence of approximately six, which is slightly smaller than expected.

In Figure 6.2(b), Figure 6.2(d) and Figure 6.2(f) we check Assumption (A1) for all three methods under consideration. We estimate the average work using the sample mean from (6.12). Since we expect $\gamma_1 = 3$ and because $\mathrm{DOF}_l := (h_l^{-1} q_l)^2$, the average work should scale

as $\hat{w}_l = \mathscr{O}(\mathrm{DOF}_l^{3/2})$ (cf. Remark 6.3 1). Combining the estimate of $\kappa_2$ from Figure 6.1(d) and the fact that $\gamma_1 \approx 3$, we compute that $h$-, $p$- and $hp$-MLMC satisfy $\kappa_2 \tilde{q}_0 > \gamma_1$ (see the parameters of the coarsest level given in Table 6.1). Although $h$-MLMC does not satisfy $2\kappa_1 \geq \kappa_2$ from Assumption (A2), we observe in Figure 6.1(e) that all three methods yield an optimal asymptotic work of $\mathscr{O}(\varepsilon^{-2})$. Moreover, compared to MC, all three methods yield a speedup of approximately two orders. It appears that for small tolerances $hp$-MLMC proves to be the more efficient than $h$- and $p$-MLMC. This is probably due to the very small variance on level two, because for the same tolerance $hp$-MLMC requires significantly less samples on level two than $h$- and $p$-MLMC. Finally, we check whether the prescribed tolerance $\varepsilon$ is reached by all three methods. To this end we evaluate the statistical error $\varepsilon_{\mathrm{stat}}$ by (6.11) and compute the deterministic approximation error $\varepsilon_{\mathrm{det}}$ using the exact solution (6.53). Both terms $\varepsilon_{\mathrm{det}}$ and $\varepsilon_{\mathrm{stat}}$ are then summed up as in (6.6) and plotted in Figure 6.1(f).

The computed number of samples on each level for different tolerances is shown on the left-hand side of Figure 6.2. As expected, we only need a few computations on the fine levels, the majority of the computations is performed on coarse levels because the computation of coarse-level samples is very cheap compared to fine-level samples. The average work is plotted on the right-hand side of Figure 6.2. All three methods exhibit an average work of approximately $\gamma_1 = 3$, cf. the discussion from above.

Figure 6.3 illustrates the advantage of computing the lower confidence bound $\underline{M_l}$ from (6.43). In this figure we plot the values of $\underline{M_l}$ from (6.43), of $\tilde{M}_l$ from (6.46) and of $\hat{M}_l$ from (6.10), for each level $l = 0, \ldots, 3$, for the first three iterations of Algorithm $hp$-`MLMC with confidence intervals`. For this example we use $h$-MLMC. In Figure 6.3(a) we observe that relying on the estimate of $\hat{M}_0$ would have led to an overestimate of the optimal number of samples by more than 3000 samples. The same holds for level one (Figure 6.3(b)) where we would have overestimated the optimal number of samples by more than 300. On the other hand, our proposed strategy ensures that we reach the standard estimator $\hat{M}_l$ in (at most) three iterations. This shows in particular the advantage of our approach because we avoid computing unnecessary samples, hence we save computing time, while trying to keep the number of queuing operations low.

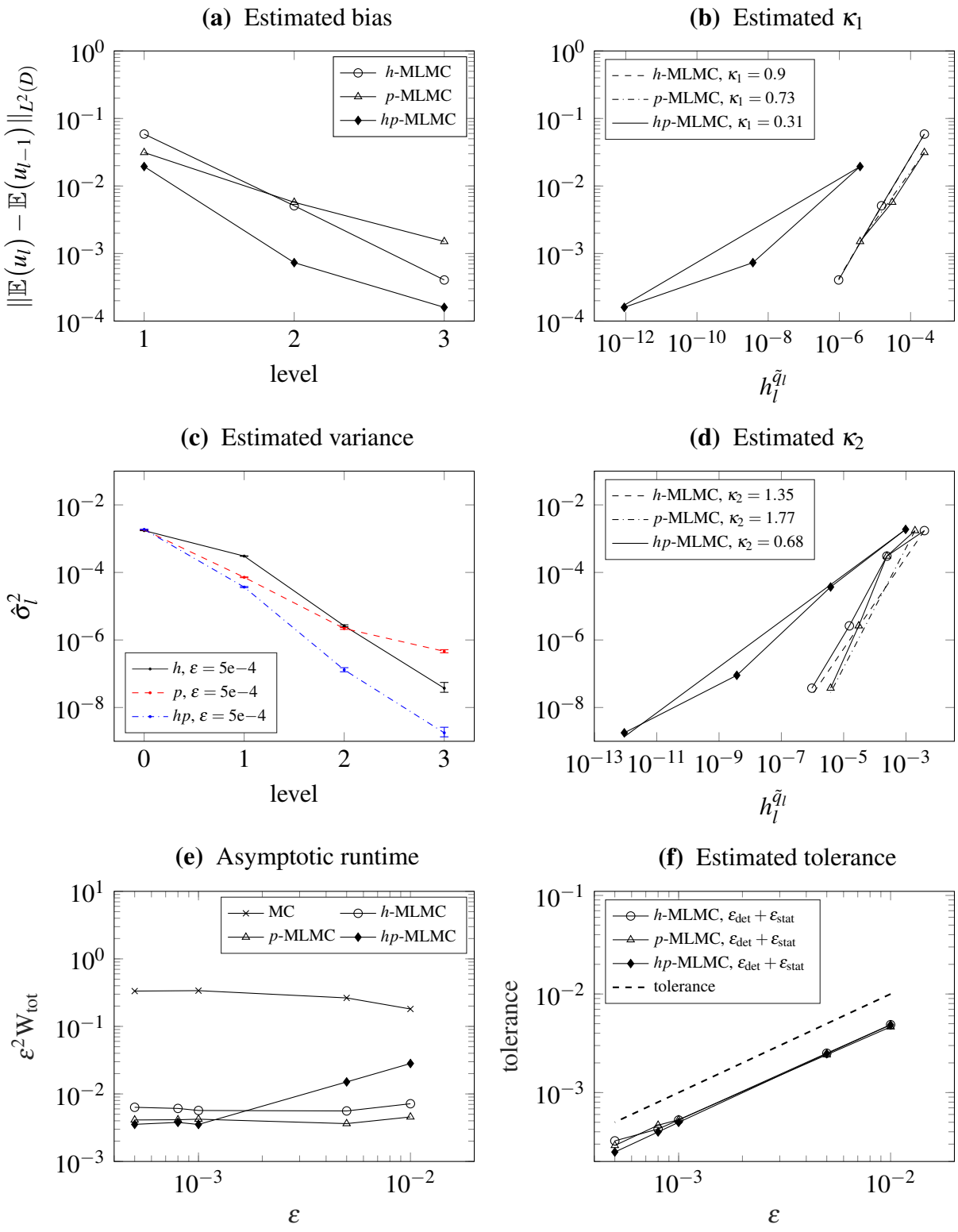
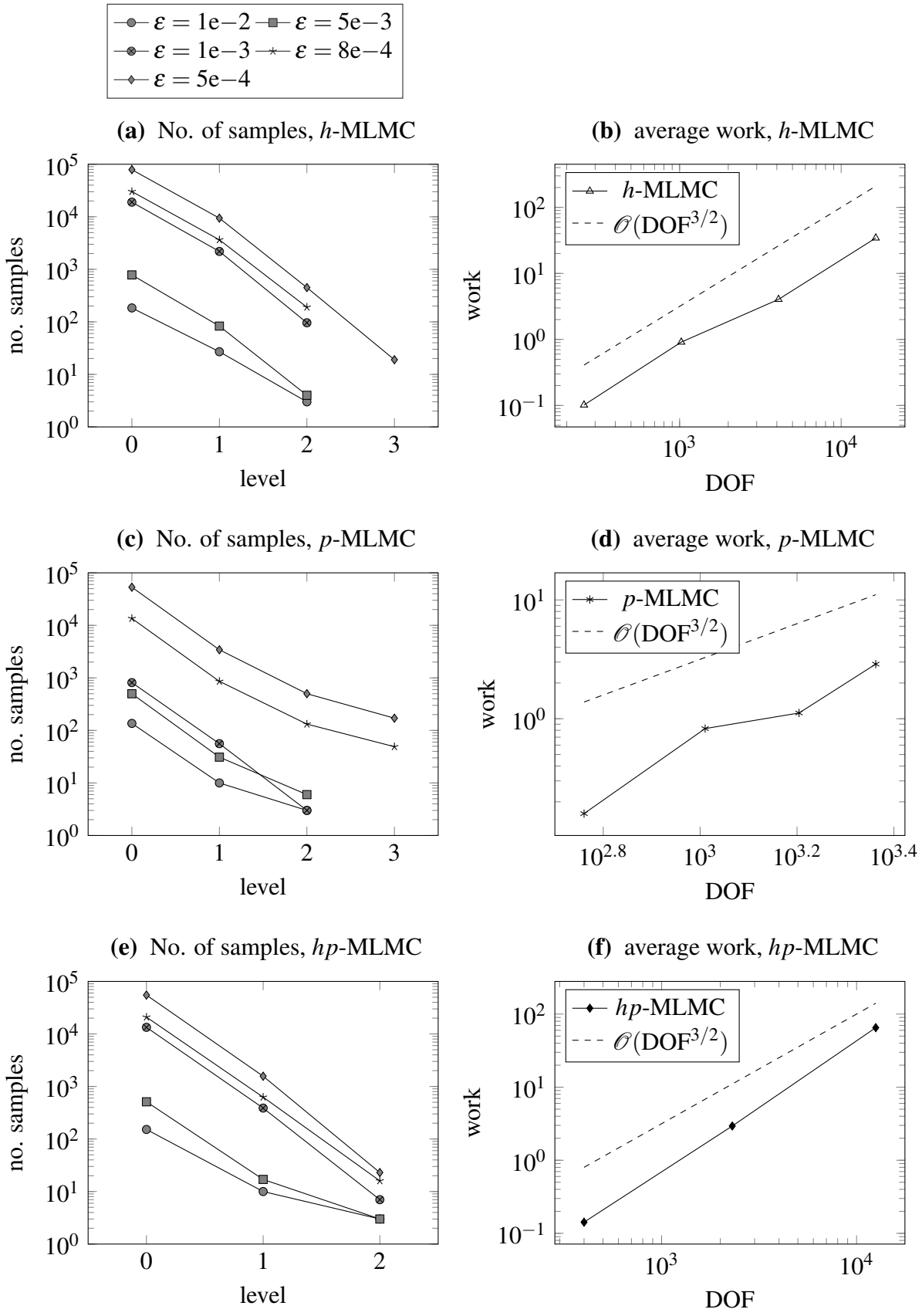



Figure 6.1: Estimated bias, variance, tolerance and asymptotic runtime. For $\hat{\sigma}_l^2$ we also plot the 95% confidence interval. Example 6.4.1.




**(a)** No. of samples, *h*-MLMC

**(b)** average work, *h*-MLMC

**(c)** No. of samples, *p*-MLMC

**(d)** average work, *p*-MLMC

**(e)** No. of samples, *hp*-MLMC

**(f)** average work, *hp*-MLMC

Figure 6.2: Computed number of samples and average work on each level. Example 6.4.1.

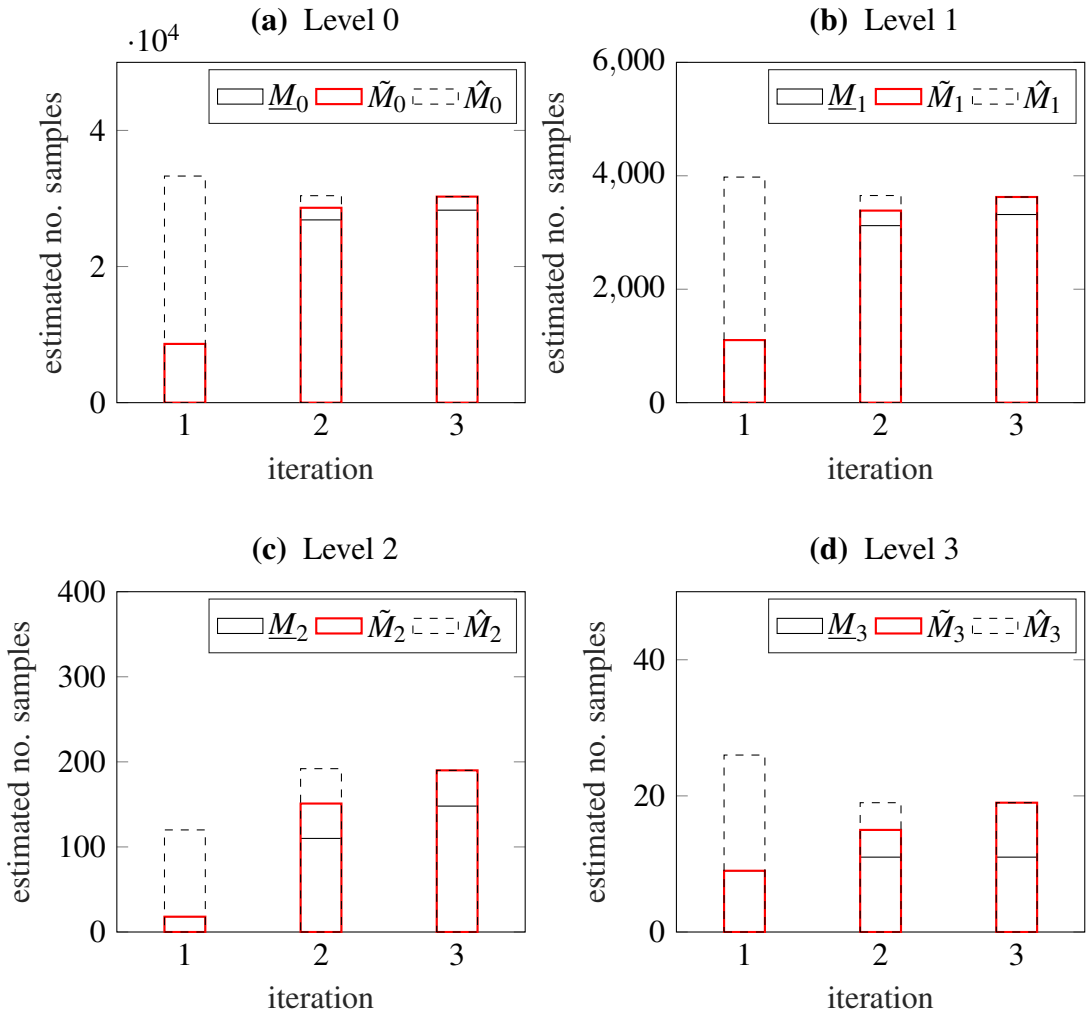



Figure 6.3: Estimated $\underline{M}_l$, $\hat{M}_l$ and chosen number of samples $\tilde{M}_l$ from (6.46) for different levels for the $h$-MLMC method. The tolerance in this example is $\varepsilon = 5\text{e}{-4}$. The number of warm-up samples was (100, 10, 3, 3) (Example 6.4.1).

### 6.4.2 Open Cavity

In this numerical example we investigate the influence of uncertain input parameters on the aeroacoustic feedback of cavity flows as in [74]. The prediction of aeroacoustic noise is an important branch of research for example in the automotive industry, however due to the large bandwidth of spatial and temporal scales, a high order numerical scheme with low dissipation and dispersion error is necessary to preserve important small scale information and hence it poses a very challenging numerical problem for UQ. We consider the flow over a two-dimensional open cavity, cf. Figure 6.4, using $h$-, $p$- and $hp$-MLMC.

For this flow problem we consider two uncertain parameters. The first uncertain parameter is the initial condition for pressure, i.e. we let $p^0 \sim \mathcal{N}(1.8, 0.01^2)$ be normally distributed. With this choice the Mach number $\mathrm{Ma} = \frac{v_1}{c}$, where $c = \sqrt{\gamma \frac{p}{\rho}}$ is the speed of sound, becomes uncertain. The initial condition in primitive variables then reads as follows

$$\left(\rho^0, v_1^0, v_2^0, p^0\right) = \left(1, 1, 0, p^0(y)\right).$$

As a second uncertain parameter, we let the boundary layer thickness in front of the cavity, $\delta_{99} \sim \mathcal{U}(0.28, 0.48)$, be uniformly distributed.

For the boundary conditions at the inlet we employ Dirichlet boundary conditions in combination with a precomputed Blasius boundary layer profile. All wall boundaries are modeled as isothermal no-slip walls where the temperature $\mathcal{T}$ is computed from the ideal gas law $p = \rho R \mathcal{T}$, where the gas constant for air satisfies $R = 287.058$. Downstream of the cavity we consider a pressure outflow boundary condition, where the pressure is specified by the initial pressure. Above the cavity we also consider a pressure outflow boundary condition. We augment the downstream and upper boundaries with a sponge zone (cf. Figure 6.4) to avoid that artificial reflections reenter the computational domain. Detailed information about the sponge zone can be found in [74].

For our QoI we record the pressure fluctuations $p(t, x, y)$ on top of the cavity at $x_0 = (x_1, x_2) = (1.57, 0)$ over time and then perform the discrete-time Fourier transform (DTFT) to obtain the sound pressure spectrum at $x_0$, i.e.

$$Q(u) = \mathrm{DTFT}\Big(p(\cdot, x_0, y)\Big).$$

The corresponding $L^2$-norm is then taken over frequency space. The mesh hierarchies for $h$-, $p$- and $hp$-MLMC can be found in Table 6.2. For the confidence intervals in (6.45) we set $\alpha$ to be 0.025.

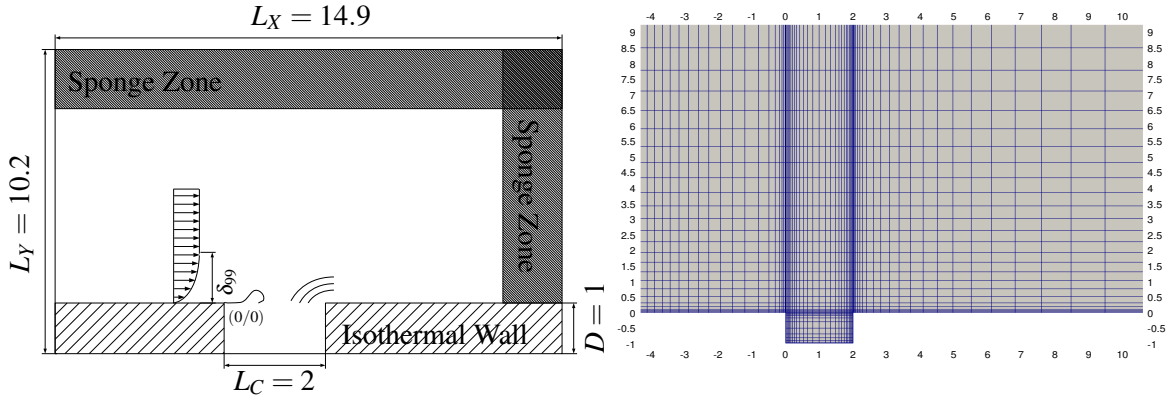



Figure 6.4: *Left:* Schematic sketch of the open cavity setup with a laminar inflow boundary layer. All geometric parameters are adopted from [74] and are non-dimensionalized by the cavity depth. *Right:* Computational mesh on the finest level. Example 6.4.2.

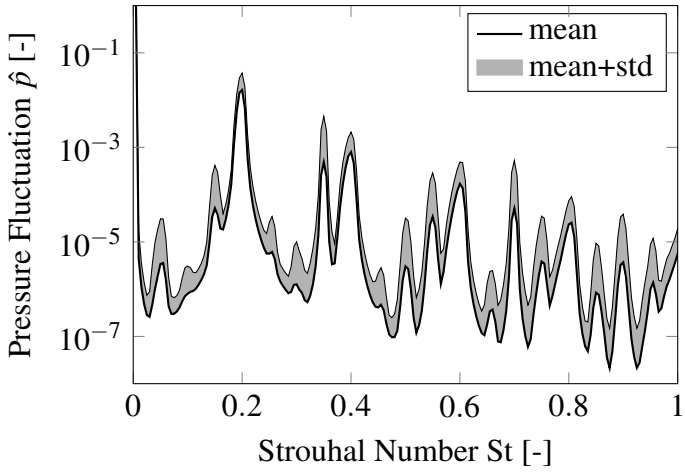



Figure 6.5: Mean frequency spectrum and standard deviation obtained with $p$-MLMC for $\varepsilon = 5\text{e}{-5}$. Example 6.4.2.

| level | $h$-MLMC | | $p$-MLMC | | $hp$-MLMC | |
|---|---|---|---|---|---|---|
| | $N_l$ | $q_l$ | $N_l$ | $q_l$ | $N_l$ | $q_l$ |
| 0 | 279 | 7 | 1987 | 4 | 279 | 4 |
| 1 | 423 | 7 | 1987 | 5 | 423 | 5 |
| 2 | 957 | 7 | 1987 | 6 | 957 | 6 |
| 3 | 1987 | 7 | 1987 | 7 | 1987 | 7 |

Table 6.2: Level setup for $h$-, $p$ and $hp$-MLMC. Example 6.4.2.

Because computing a MC reference solution for this example would be too expensive, we show in Figure 6.5 the resulting mean frequency spectrum and its standard deviation for $p$-MLMC with a prescribed tolerance $\varepsilon = 5\text{e}{-5}$. For $U_\infty$ being the free-stream velocity, $f$ the frequency and $L_c$ the length of the cavity we define the Strouhal number $\text{St}=\frac{fL}{U_\infty}$, which is a dimensionless frequency and an important fluid mechanical parameter. The dominant peaks, which correspond to the so-called Rossiter modes (cf. [74, 73]) are clearly observable in the mean spectrum. From the high standard deviation of $\hat{p}$ at $\text{St}\approx 0.3$ and $\text{St}\approx 0.4$, which correspond to the positions of the second and third Rossiter mode, we deduce that a switching between both modes takes place, which depends on the Mach number of the flow. The same phenomenon has been observed for an uncertain cavity depth in [74].

Considering the bias error we can see that in Figure 6.6(a) for $p$-MLMC the bias estimate $\|\mathbb{E}\big(Q(u_l)\big) - \mathbb{E}\big(Q(u_{l-1})\big)\|_{L^2}$ for $p$-MLMC is smaller than $2.5\text{e}{-5}$. As all three methods share the same finest level, we can assume that the bias error from (6.6) is satisfied for all three methods under consideration.

Figure 6.6(b) shows the estimated level variance $\hat{\sigma}_l^2$ across different levels. All three methods yield a very good variance reduction, especially $p$-MLMC has already a very small variance on level two. However, the computation of samples on the coarse grids is extremely costly for $p$-MLMC. Taking a closer look at the variance on level zero, we see that $hp$-MLMC achieves the same variance as $p$-MLMC but with much less DOFs. This yields the computational advantage of $h$- and $hp$-MLMC compared to $p$-MLMC (see Figure 6.6(c)) for this open cavity problem. The asymptotic work is still optimal for all three methods, which can be seen in Figure 6.6(c).

Finally, in Figure 6.7 we plot the number of computed samples for different tolerances and the average work that is needed to compute one sample on the corresponding level. Again, most of the computations are performed on the coarse levels as expected. In contrast to the smooth benchmark problem from Section 6.4.1 the average work does not scale as $\mathscr{O}(\text{DOF}_l^{3/2})$ but more like $\mathscr{O}(\text{DOF}_l^2)$, indicating that $\gamma_1 \approx 4$. The main reason for this behavior is that the uncertain Mach number influences the time-step size because the speed of sound is a part of the eigenvalues of the Jacobian of the advective fluxes. The bigger the pressure, the smaller the time-step size, which results in a bigger work.

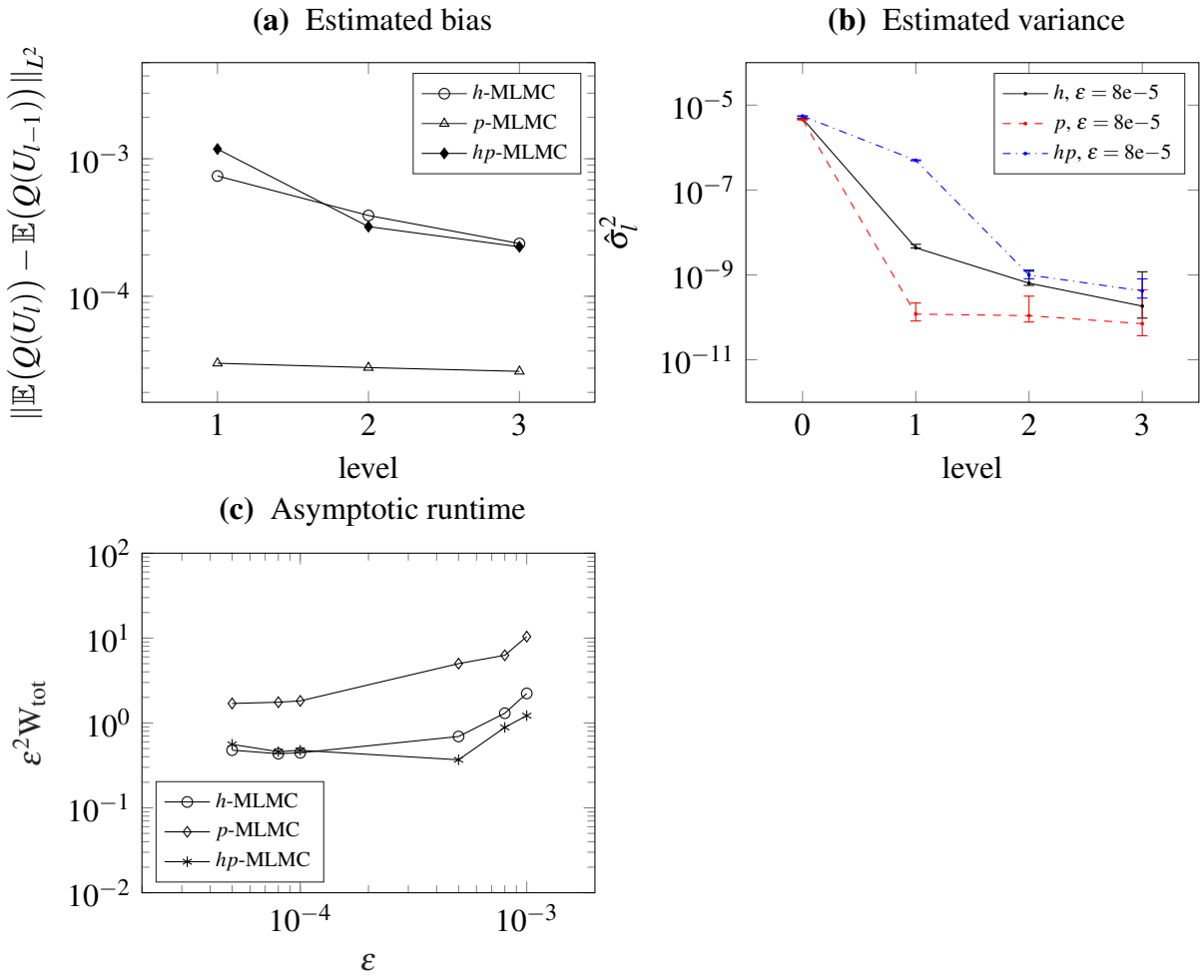



Figure 6.6: Estimated bias, variance and asymptotic work. For $\hat{\sigma}_l^2$ we also plot the 95% confidence interval. Example 6.4.2.

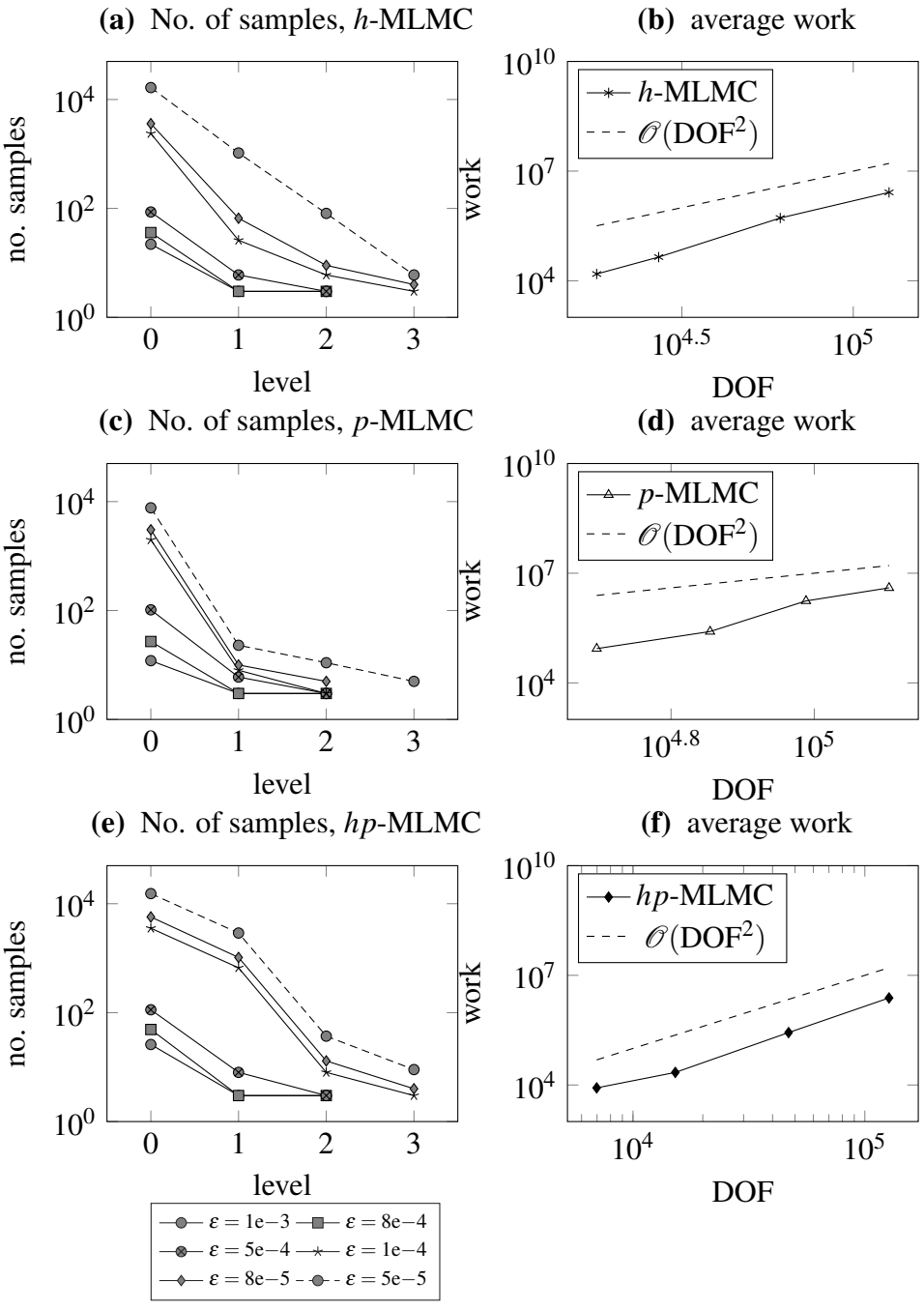



Figure 6.7: Computed number of samples and average work on each level. Example 6.4.2.

# 7 Conclusions and outlook

In this section we summarize the outcomes and achievements of this thesis. Moreover, we want to point out some possible directions for further research in the field of UQ of compressible flows.

## 7.1 Summary

High-resolution numerical schemes that account for uncertainties which may arise due to modeling or measurement errors, are becoming more and more important. Thanks to the continuous improvement of computer-processing capacities in the last few decades, it is now possible to combine high-fidelity numerical solvers with mathematical models, which allow for the incorporation and simulation of uncertainties in input parameters. In this thesis we investigated, developed and improved various methods for UQ of compressible flows.

In a first step we embedded deterministic hyperbolic conservation laws into a probabilistic framework and established existence and uniqueness of so-called random entropy admissible weak solution for random conservation laws with uncertain initial data and random flux function. We discretized the random conservation laws in space and time using the Runge–Kutta Discontinuous Galerkin method and for the stochastic discretization (UQ method respectively) we employed the Stochastic Galerkin (SG), Non-intrusive Spectral Projection (NISP), Stochastic Collocation (SC) and Multilevel Monte Carlo (MLMC) method.

While SG is a frequently and successfully used method for UQ of random pdes, it has been shown that for nonlinear hyperbolic problems it does not necessarily preserve hyperbolicity of the underlying hyperbolic system. To remedy this issue we developed a hyperbolicity-preserving numerical scheme for a SG-DG discretization of random hyperbolic conservation laws. We derived a modified CFL-condition, under which the space-stochastic cell-mean is admissible, i.e. generates a hyperbolic SG system, after one time-step of forward Euler. With the help of a slope-limiter, which limits the solution towards the admissible cell-mean, we en-

sure that the solution remains admissible after each time-step. Various numerical experiments have proven that the modified numerical scheme is a well-suited and robust method for intrusive UQ of complex compressible flow problems governed by the Euler equations. Beside the development of the hyperbolicty-preserving numerical scheme for the SG method, we extended the massively parallel solver FLEXI [63] to the SG and ME-SG method, resulting in the novel High-Performance Computing UQ software framework SG-FLEXI [12]. The numerical solver is capable of simulating random compressible Euler and random compressible Navier–Stokes equations up to three spatial dimensions and arbitrary random dimensions.

Another important issue that we addressed in this thesis is the derivation of error estimates for discretization errors arising from spatio-temporal and stochastic discretization. In this work we have derived the first rigorous a posteriori error analysis for random hyperbolic conservation laws. Our approach is based on the relative entropy framework of Dafermos and DiPerna [26], where we estimated the difference between the random entropy admissible weak solution and a reconstruction of the numerical approximation in the $L^2_w(\Xi;L^2(D))$-norm. Furthermore, we showed that the space-time-stochastic residual admits a splitting into a residual representing the space-time error and a residual representing the stochastic approximation error. Based on the decomposition of the residual we proposed for the SC method a novel residual-based, space-stochastic adaptive algorithm. Extensive numerical investigations showed that for smooth solutions both residuals exhibit the correct order of convergence and are indeed independent of each other. Moreover, the proposed adaptive numerical schemes provided a significant efficiency gain compared to uniform mesh refinements.

In the last chapter of this thesis we introduced the $hp$-MLMC method, an extension of the classical MLMC method to arbitrarily $hp$-refined meshes. We generalized the complexity analysis of MLMC to the $hp$-MLMC method and proved that in best case the asymptotic work of $hp$-MLMC scales quadratically with the prescribed tolerance. Another issue that we addressed in this thesis is the robust estimation of the number of additional samples after every iteration of $hp$-MLMC. The estimate of the number of additional samples per level is based on statistical estimates of average computational time and level variances and is therefore a stochastic quantity. As the efficiency of the MLMC method strongly depends on this quantity we provide a novel confidence interval for the optimal number of samples per level. The proposed confidence interval provides a robust and reliable lower estimate for the number of additional samples and helps to prevent overestimating the optimal number of samples per level. We applied the $hp$-MLMC to different problems governed by the compressible Navier–Stokes equations, in particular we considered a cavity flow problem from computational acoustics, demonstrating that the method is suitable to handle complex engineering problems.

## 7.2 Directions for further research

In this thesis we presented a first rigorous a posteriori error estimate for random conservation laws. However, a main issue of our approach is that we have only reliable error control as long as the solution is smooth. In the post-shock regime the error estimator blows-up under mesh refinement. For scalar conservation laws an a posteriori error estimator based on Kružkov estimates may circumvent this problem (see Theorem 3.2), however the estimator has a sub-optimal rate of convergence, cf. [71] and the discussion in Remark 5.14. For systems of hyperbolic conservation laws an appropriate a posteriori error analysis, which is also reliable in the post-shock regime is still not in sight. Therefore, our proposed approach corresponds to the current state of the art of the a posteriori error analysis for hyperbolic systems of conservation laws.

A novel approach to UQ for random conservation may be provided by so-called statistical solutions [41, 42], which are probability measures on $L^p$-spaces. This approach extends the notion of (spatially one-point) measure-valued solutions [32] to spatially $k$-point cross-corelations. The authors of [41] were able to prove existence and uniqueness of entropy admissible statistical solutions for scalar conservation laws. Recent results concerning the convergence of numerical approximations towards (dissipative) measure-valued [39, 44] and statistical solutions [43] suggest that measure-valued, resp. statistical solutions might be the correct framework to handle the issue of existence and uniquess of solutions of multi-dimensional hyperbolic conservation laws. In a recent publication [43] the authors were able to extend the relative entropy method to the setting of statistical solutions and prove weak-strong uniqueness for dissipative entropy admissible statistical solutions. We believe that it is possible to extend our a posteriori error framework to statistical solutions and obtain similar error estimates in the Wasserstein distance.

Another interesting problem for further research is the extension of the reconstruction to two- or three spatial domains. A first approach to extend the reconstruction to two spatial dimensions is presented in [50]. However, the proposed reconstruction does not yield an optimal decay of the corresponding residual and from our point of view, the correct reconstruction approach is far from clear. Moreover, the a posteriori error analysis should be extended to include more realistic boundary conditions and it would be interesting to apply the error estimator to problems described by the compressible Navier–Stokes equations.

Another open problem is the optimality of the stochastic residuals. Although we were not able to prove any convergence rates for the stochastic residual, our numerical experiments indicate that the stochastic residual exhibits an optimal (algebraic or spectral) error decay, hence further investigations in this direction should be conducted.

Concerning the $hp$-MLMC method an interesting question would be if it is possible to further increase the efficiency of the $hp$-MLMC method using $hp$-adaptive numerical schemes. This would also be a very good opportunity to further develop the a posteriori error estimator for the compressible Navier–Stokes equations and use the residual as indicator for refinement.

# Bibliography


[1] R. ABGRALL AND S. MISHRA, *Uncertainty quantification for hyperbolic systems of conservation laws*, in Handbook of numerical methods for hyperbolic problems, vol. 18 of Handb. Numer. Anal., Elsevier/North-Holland, Amsterdam, 2017, pp. 507–544.

[2] B. K. ALPERT, *A class of bases in $L^2$ for the sparse representation of integral operators*, SIAM J. Math. Anal., 24 (1993), pp. 246–262.

[3] I. BABUŠKA, F. NOBILE, AND R. TEMPONE, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Rev., 52 (2010), pp. 317–355.

[4] I. BACKUS, *Sod-shocktube*. `https://github.com/ibackus/sod-shocktube`, 2017.

[5] S. BANIK, A. N. ALBATINEH, M. O. A. ABU-SHAWIESH, AND B. G. KIBRIA, *Estimating the population standard deviation with confidence interval: A simulation study under skewed and symmetric conditions*, International Journal of Statistics in Medical Research, 3 (2014), pp. 356–367.

[6] A. BARTH AND F. G. FUCHS, *Uncertainty quantification for linear hyperbolic equations with stochastic process or random field coefficients*, Appl. Numer. Math., 121 (2017), pp. 38–51.

[7] A. BARTH, C. SCHWAB, AND J. ŠUKYS, *Multilevel Monte Carlo simulation of statistical solutions to the Navier-Stokes equations*, in Monte Carlo and quasi-Monte Carlo methods, vol. 163 of Springer Proc. Math. Stat., Springer, [Cham], 2016, pp. 209–227.

[8] A. BARTH, C. SCHWAB, AND N. ZOLLINGER, *Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients*, Numer. Math., 119 (2011), pp. 123–161.

[9] A. BARTH AND A. STEIN, *A study of elliptic partial differential equations with jump diffusion coefficients*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1707–1743.

[10] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[11] A. BECK, J. DÜRRWÄCHTER, T. KUHN, F. MEYER, C.-D. MUNZ, AND C. ROHDE, *hp-Multilevel Monte Carlo Methods for Uncertainty Quantification of Compressible Flows*, submitted to SIAM Journal on Scientific Computing, arXiv:1808.10626, (2018).

[12] ——, *A High-Order Stochastic Galerkin Code for the Compressible Euler and Navier-Stokes Equations*, submitted to Computers and Fluids, (2019).

[13] A. BESPALOV, D. PRAETORIUS, L. ROCCHI, AND M. RUGGERI, *Goal-oriented error estimation and adaptivity for elliptic PDEs with parametric or uncertain inputs*, Comput. Methods Appl. Mech. Engrg., 345 (2019), pp. 951–982.

[14] S. BIANCHINI AND R. M. COLOMBO, *On the stability of the standard Riemann semigroup*, Proc. Amer. Math. Soc., 130 (2002), pp. 1961–1973.

[15] A. BRESSAN, *Uniqueness and stability for one dimensional hyperbolic systems of conservation laws*, in XIIIth International Congress on Mathematical Physics (London, 2000), Int. Press, Boston, MA, 2001, pp. 311–317.

[16] A. BRESSAN AND P. LEFLOCH, *Uniqueness of weak solutions to systems of conservation laws*, Arch. Rational Mech. Anal., 140 (1997), pp. 301–317.

[17] A. BRESSAN AND M. LEWICKA, *A uniqueness condition for hyperbolic systems of conservation laws*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 673–682.

[18] C. M. BRYANT, S. PRUDHOMME, AND T. WILDEY, *Error decomposition and adaptivity for response surface approximations from PDEs with parametric uncertainty*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1020–1045.

[19] H.-J. BUNGARTZ AND M. GRIEBEL, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.

[20] A. CHERTOCK, A. KURGANOV, M. LUKÁCOVÁ-MEDVIDOVÁ, P. SPICHTINGER, AND B. WIEBE, *Stochastic Galerkin method for cloud simulation*, Mathematics of Climate and Weather Forecasting, 5 (2019), pp. 65–106.

[21] E. CHIODAROLI, *A counterexample to well-posedness of entropy solutions to the compressible Euler system*, J. Hyperbolic Differ. Equ., 11 (2014), pp. 493–519.

[22] K. A. CLIFFE, M. B. GILES, R. SCHEICHL, AND A. L. TECKENTRUP, *Multilevel*

*Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15.

[23] B. COCKBURN AND C.-W. SHU, *The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[24] ——, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[25] P. G. CONSTANTINE, M. S. ELDRED, AND E. T. PHIPPS, *Sparse pseudospectral approximation method*, Comput. Methods Appl. Mech. Engrg., 229/232 (2012), pp. 1–12.

[26] C. M. DAFERMOS, *Hyperbolic conservation laws in continuum physics*, vol. 325 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, fourth ed., 2016.

[27] C. DE LELLIS AND L. SZÉKELYHIDI, JR., *On admissibility criteria for weak solutions of the Euler equations*, Arch. Ration. Mech. Anal., 195 (2010), pp. 225–260.

[28] A. DEDNER AND J. GIESSELMANN, *A posteriori analysis of fully discrete method of lines discontinuous Galerkin schemes for systems of conservation laws*, SIAM J. Numer. Anal., 54 (2016), pp. 3523–3549.

[29] A. DEDNER, C. MAKRIDAKIS, AND M. OHLBERGER, *Error control for a class of Runge-Kutta discontinuous Galerkin methods for nonlinear conservation laws*, SIAM J. Numer. Anal., 45 (2007), pp. 514–538.

[30] A. DEDNER AND M. OHLBERGER, *A new hp-adaptive DG scheme for conservation laws based on error control*, in Hyperbolic problems: theory, numerics, applications, Springer, Berlin, 2008, pp. 187–198.

[31] B. DESPRÉS, G. POËTTE, AND D. LUCOR, *Robust uncertainty propagation in systems of conservation laws with the entropy closure method*, in Uncertainty quantification in computational fluid dynamics, vol. 92 of Lect. Notes Comput. Sci. Eng., Springer, Heidelberg, 2013, pp. 105–149.

[32] R. J. DIPERNA, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[33] V. DOLEJŠÍ AND M. FEISTAUER, *Discontinuous Galerkin method*, vol. 48 of Springer

Series in Computational Mathematics, Springer, Cham, 2015. Analysis and applications to compressible flow.

[34] J. DÜRRWÄCHTER, T. KUHN, F. MEYER, L. SCHLACHTER, AND F. SCHNEIDER, *A hyperbolicity-preserving discontinuous stochastic galerkin scheme for uncertain hyperbolic systems of equations*, J. Comput. Appl. Math., (2019), p. 112602.

[35] M. EIGEL, C. J. GITTELSON, C. SCHWAB, AND E. ZANDER, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg., 270 (2014), pp. 247–269.

[36] B. EINFELDT, C.-D. MUNZ, P. L. ROE, AND B. SJÖGREEN, *On Godunov-type methods near low densities*, J. Comput. Phys., 92 (1991), pp. 273–295.

[37] O. G. ERNST, A. MUGLER, H.-J. STARKLOFF, AND E. ULLMANN, *On the convergence of generalized polynomial chaos expansions*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 317–339.

[38] J. FEINGERG AND H. P. LANGTANGEN, *Chaospy: An open source tool for designing methods of uncertainty quantification*, Journal of Computational Science, 11 (2015), pp. 46–57.

[39] E. FEIREISL, M. LUKÁCOVÁ-MEDVIDOVÁ, AND H. MIZEROVÁ, *Convergence of Finite Volume Schemes for the Euler Equations via Dissipative Measure-Valued Solutions*, Foundations of Computational Mathematics, (2019), pp. 1–44.

[40] U. S. FJORDHOLM, R. KÄPPELI, S. MISHRA, AND E. TADMOR, *Construction of approximate entropy measure-valued solutions for hyperbolic systems of conservation laws*, Found. Comput. Math., 17 (2017), pp. 763–827.

[41] U. S. FJORDHOLM, S. LANTHALER, AND S. MISHRA, *Statistical solutions of hyperbolic conservation laws: foundations*, Arch. Ration. Mech. Anal., 226 (2017), pp. 809–849.

[42] U. S. FJORDHOLM, K. LYE, AND S. MISHRA, *Numerical approximation of statistical solutions of scalar conservation laws*, SIAM J. Numer. Anal., 56 (2018), pp. 2989–3009.

[43] U. S. FJORDHOLM, K. LYE, S. MISHRA, AND F. WEBER, *Statistical solutions of hyperbolic systems of conservation laws: numerical approximation*, arXiv:1906.02536, (2019).

[44] U. S. FJORDHOLM, S. MISHRA, AND E. TADMOR, *On the computation of measure-valued solutions*, Acta Numer., 25 (2016), pp. 567–679.

[45] R. G. GHANEM AND P. D. SPANOS, *Stochastic finite elements: a spectral approach*, Springer-Verlag, New York, 1991.

[46] J. GIESSELMANN, C. MAKRIDAKIS, AND T. PRYER, *A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws*, SIAM J. Numer. Anal., 53 (2015), pp. 1280–1303.

[47] J. GIESSELMANN, F. MEYER, AND C. ROHDE, *A posteriori error analysis and adaptive non-intrusive numerical schemes for systems of random conservation laws*, accepted for publication in BIT Numerical Mathematics, arXiv:1902.05375, (2019).

[48] ——, *An a posteriori error analysis based on non-intrusive spectral projections for systems of random conservation laws*, accepted for publication in Proceedings of HYP2018, arXiv:1902.05375, (2019).

[49] ——, *A posteriori error analysis for random scalar conservation laws using the Stochastic Galerkin method*, IMA J. Numer. Anal., URL: `https://dx.doi.org/10.1093/imanum/drz004`, (2019).

[50] J. GIESSELMANN AND T. PRYER, *A posteriori analysis for dynamic model adaptation in convection-dominated problems*, Math. Models Methods Appl. Sci., 27 (2017), pp. 2381–2423.

[51] M. B. GILES, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617.

[52] C. J. GITTELSON, R. ANDREEV, AND C. SCHWAB, *Optimality of adaptive Galerkin methods for random parabolic partial differential equations*, J. Comput. Appl. Math., 263 (2014), pp. 189–201.

[53] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[54] L. GOSSE AND C. MAKRIDAKIS, *Two a posteriori error estimates for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 38 (2000), pp. 964–988.

[55] D. GOTTLIEB AND D. XIU, *Galerkin method for wave equations with uncertain coefficients*, Commun. Comput. Phys., 3 (2008), pp. 505–518.

[56] D. GUIGNARD AND F. NOBILE, *A posteriori error estimation for the stochastic collocation finite element method*, SIAM J. Numer. Anal., 56 (2018), pp. 3121–3143.

[57] M. GUNZBURGER, C. G. WEBSTER, AND G. ZHANG, *An adaptive wavelet stochastic*

*collocation method for irregular solutions of partial differential equations with random input data*, in Sparse grids and applications—Munich 2012, vol. 97 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2014, pp. 137–170.

[58] B. GUSTAFSSON, H.-O. KREISS, AND J. OLIGER, *Time-dependent problems and difference methods*, Pure and Applied Mathematics (Hoboken), John Wiley & Sons, Inc., Hoboken, NJ, second ed., 2013.

[59] A. HARTEN AND J. M. HYMAN, *Self-adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comput. Phys., 50 (1983), pp. 235–269.

[60] R. HARTMANN AND P. HOUSTON, *Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 979–1004.

[61] S. HEINRICH, *Multilevel monte carlo methods*, in Large-Scale Scientific Computing, S. Margenov, J. Waśniewski, and P. Yalamov, eds., Berlin, Heidelberg, 2001, pp. 58–67.

[62] J. S. HESTHAVEN AND T. WARBURTON, *Nodal discontinuous Galerkin methods*, vol. 54 of Texts in Applied Mathematics, Springer, New York, 2008. Algorithms, analysis, and applications.

[63] F. HINDENLANG, G. J. GASSNER, C. ALTMANN, A. BECK, M. STAUDENMAIER, AND C.-D. MUNZ, *Explicit discontinuous Galerkin methods for unsteady problems*, Comput. & Fluids, 61 (2012), pp. 86–93.

[64] H. HOLDEN AND N. H. RISEBRO, *Front tracking for hyperbolic conservation laws*, vol. 152 of Applied Mathematical Sciences, Springer, Heidelberg, second ed., 2015.

[65] D. I. KETCHESON, *Highly efficient strong stability-preserving Runge-Kutta methods with low-storage implementations*, SIAM J. Sci. Comput., 30 (2008), pp. 2113–2136.

[66] M. KÖPPEL, F. FRANZELIN, I. KRÖKER, S. OLADYSHKIN, G. SANTIN, D. WITTWAR, A. BARTH, B. HAASDONK, W. NOWAK, D. PFLÜGER, AND ET AL., *Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario*, Comput. Geosci., 23 (2019), pp. 339–354.

[67] M. KÖPPEL, I. KRÖKER, AND C. ROHDE, *Intrusive uncertainty quantification for hyperbolic-elliptic systems governing two-phase flow in heterogeneous porous media*, Comput. Geosci., 21 (2017), pp. 807–832.

[68] D. A. KOPRIVA, *Implementing spectral methods for partial differential equations*, Scientific Computation, Springer, Berlin, 2009. Algorithms for scientists and engineers.

[69] I. KRÖKER, W. NOWAK, AND C. ROHDE, *A stochastically and spatially adaptive parallel scheme for uncertain and nonlinear two-phase flow problems*, Comput. Geosci., 19 (2015), pp. 269–284.

[70] D. KRÖNER, *Numerical schemes for conservation laws*, Wiley-Teubner Series Advances in Numerical Mathematics, John Wiley & Sons, Ltd., Chichester; B. G. Teubner, Stuttgart, 1997.

[71] D. KRÖNER AND M. OHLBERGER, *A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions*, Math. Comp., 69 (2000), pp. 25–39.

[72] S. N. KRUŽKOV, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (123) (1970), pp. 228–255.

[73] T. KUHN, J. DÜRRWÄCHTER, A. BECK, AND C.-D. MUNZ, *Zonal large eddy simulation of active open cavity noise using a high order discontinuous galerkin method*, AIAA Conference Paper, (2019).

[74] T. KUHN, J. DÜRRWÄCHTER, F. MEYER, A. BECK, C. ROHDE, AND C.-D. MUNZ, *Uncertainty Quantification for Direct Aeroacoustic Simulations of Cavity Flows*, Journal of Theoretical and Computational Acoustics, 27 (2019), p. 1850044.

[75] J. KUSCH, R. G. MCCLARREN, AND M. FRANK, *Filtered Stochastic Galerkin Methods For Hyperbolic Equations*, Journal of Computational Physics, (2019), p. 109073.

[76] O. P. LE MAÎTRE AND O. M. KNIO, *Spectral methods for uncertainty quantification*, Scientific Computation, Springer, New York, 2010. With applications to computational fluid dynamics.

[77] O. P. LE MAÎTRE, M. T. REAGAN, H. N. NAJM, R. G. GHANEM, AND O. M. KNIO, *A stochastic projection method for fluid flow. II. Random process*, J. Comput. Phys., 181 (2002), pp. 9–44.

[78] S. MISHRA, N. H. RISEBRO, C. SCHWAB, AND S. TOKAREVA, *Numerical solution of scalar conservation laws with random flux functions*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 552–591.

[79] S. MISHRA AND C. SCHWAB, *Sparse tensor multi-level Monte Carlo finite volume meth-*

*ods for hyperbolic conservation laws with random initial data*, Math. Comp., 81 (2012), pp. 1979–2018.

[80] S. MISHRA, C. SCHWAB, AND J. ŠUKYS, *Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions*, J. Comput. Phys., 231 (2012), pp. 3365–3388.

[81] M. MOTAMED AND D. APPELÖ, *A multiorder discontinuous Galerkin Monte Carlo method for hyperbolic problems with stochastic parameters*, SIAM J. Numer. Anal., 56 (2018), pp. 448–468.

[82] M. MOTAMED, F. NOBILE, AND R. TEMPONE, *A stochastic collocation method for the second order wave equation with a discontinuous random speed*, Numer. Math., 123 (2013), pp. 493–536.

[83] F. MÜLLER, P. JENNY, AND D. W. MEYER, *Multi-level Monte Carlo for two phase flow and Buckley-Leverett transport in random heterogeneous porous media*, J. Comput. Phys., 250 (2013), pp. 685–702.

[84] M. OHLBERGER, *A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws*, Internat. J. Numer. Methods Fluids, 59 (2009), pp. 333–354.

[85] M. P. PETTERSSON, G. IACCARINO, AND J. NORDSTRÖM, *Polynomial chaos methods for hyperbolic partial differential equations*, Mathematical Engineering, Springer, Cham, 2015. Numerical techniques for fluid dynamics problems in the presence of uncertainties.

[86] M. PISARONI, F. NOBILE, AND P. LEYLAND, *A continuation multi level Monte Carlo (C-MLMC) method for uncertainty quantification in compressible inviscid aerodynamics*, Comput. Methods Appl. Mech. Engrg., 326 (2017), pp. 20–50.

[87] G. POËTTE, B. DESPRÉS, AND D. LUCOR, *Uncertainty quantification for systems of conservation laws*, J. Comput. Phys., 228 (2009), pp. 2443–2467.

[88] G. PUPPO AND M. SEMPLICE, *Numerical entropy and adaptivity for finite volume schemes*, Commun. Comput. Phys., 10 (2011), pp. 1132–1160.

[89] N. H. RISEBRO, C. SCHWAB, AND F. WEBER, *Multilevel Monte Carlo front-tracking for random scalar conservation laws*, BIT, 56 (2016), pp. 263–292.

[90] B. SCHIECHE AND J. LANG, *Adjoint error estimation for stochastic collocation meth-*

*ods*, in Sparse grids and applications—Munich 2012, vol. 97 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2014, pp. 271–293.

[91] L. SCHLACHTER AND F. SCHNEIDER, *A hyperbolicity-preserving stochastic Galerkin approximation for uncertain hyperbolic systems of equations*, J. Comput. Phys., 375 (2018), pp. 80–98.

[92] C. W. SCHULZ-RINNE, *Classification of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Math. Anal., 24 (1993), pp. 76–88.

[93] D. SERRE AND A. F. VASSEUR, *About the relative entropy method for hyperbolic systems of conservation laws*, A panorama of mathematics: pure and applied, 658 (2015), pp. 237–248.

[94] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[95] M. SONNTAG AND C.-D. MUNZ, *Efficient parallelization of a shock capturing for discontinuous Galerkin methods using finite volume sub-cells*, J. Sci. Comput., 70 (2017), pp. 1262–1289.

[96] E. SÜLI, *A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems*, in An introduction to recent developments in theory and numerics for conservation laws (Freiburg/Littenweiler, 1997), vol. 5 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 1999, pp. 123–194.

[97] E. F. TORO, *Riemann solvers and numerical methods for fluid dynamics*, Springer-Verlag, Berlin, third ed., 2009. A practical introduction.

[98] J. TRYOEN, O. LE MAÎTRE, AND A. ERN, *Adaptive anisotropic spectral stochastic methods for uncertain scalar conservation laws*, SIAM J. Sci. Comput., 34 (2012), pp. A2459–A2481.

[99] J. TRYOEN, O. LE MAÎTRE, M. NDJINGA, AND A. ERN, *Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems*, J. Comput. Phys., 229 (2010), pp. 6485–6511.

[100] R. VERFÜRTH, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[101] X. WAN AND G. E. KARNIADAKIS, *An adaptive multi-element generalized polyno-*

*mial chaos method for stochastic differential equations*, J. Comput. Phys., 209 (2005), pp. 617–642.

[102] ——, *Multi-element generalized polynomial chaos for arbitrary probability measures*, SIAM J. Sci. Comput., 28 (2006), pp. 901–928.

[103] N. WIENER, *The Homogeneous Chaos*, Amer. J. Math., 60 (1938), pp. 897–936.

[104] J. A. S. WITTEVEEN AND G. IACCARINO, *Simplex stochastic collocation with ENO-type stencil selection for robust uncertainty quantification*, J. Comput. Phys., 239 (2013), pp. 1–21.

[105] P. WOODWARD AND P. COLELLA, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys., 54 (1984), pp. 115–173.

[106] K. WU, H. TANG, AND D. XIU, *A stochastic Galerkin method for first-order quasilinear hyperbolic systems with uncertainty*, J. Comput. Phys., 345 (2017), pp. 224–244.

[107] D. XIU, *Stochastic collocation methods: a survey*, in Handbook of uncertainty quantification. Vol. 1, 2, 3, Springer, Cham, 2017, pp. 699–716.

[108] D. XIU AND J. S. HESTHAVEN, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.

[109] D. XIU AND G. E. KARNIADAKIS, *The Wiener-Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

[110] Q. ZHANG AND C.-W. SHU, *Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 641–666.

[111] ——, *Stability analysis and a priori error estimates of the third order explicit Runge-Kutta discontinuous Galerkin method for scalar conservation laws*, SIAM J. Numer. Anal., 48 (2010), pp. 1038–1063.

[112] X. ZHANG AND C.-W. SHU, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.

# Acronyms

| | |
|---|---|
| CFL | Courant-Friedrichs-Lewy |
| DG | Discontinuous Galerkin |
| DOF | Degree(s) of Freedom |
| DTFT | Discrete Time Fourier Transform |
| eoc | Experimental order of convergence |
| FD | Finite Differences |
| FE | Finite Element |
| FV | Finite Volume |
| GDG | Generalized Discontinuous Galerkin |
| gPC | Generalized Polynomial Chaos |
| hDSG | Hyperbolicity-preserving Discontinuous Stochastic Galerkin |
| HLLE | Harten-Lax-van Leer-Einfeldt |
| LF | Lax-Friedrichs |
| LW | Lax-Wendroff |
| MC | Monte Carlo |
| ME | Multi-Element |
| MLMC | Multilevel Monte Carlo |
| MSE | Mean square error |
| NISP | Non-intrusive Spectral Projection |
| PC | Polynomial Chaos |
| PDE | Partial differential equation |
| QoI | Quantity of interest |
| Re | Reynolds Number |
| RKDG | Runge–Kutta Discontinuous Galerkin |
| SC | Stochastic Collocation |
| SRS | Standard Riemann Semigroup |
| SSP RK | Strong stability preserving Runge–Kutta |
| St | Strouhal number |
| Std | Standard Deviation |
| TVBM | Total Variation bounded in the mean |
| UQ | Uncertainty Quantification |