Model Order Reduction Techniques for Uncertainty Quantification Problems

THÈSE N^{O} 6118 (2014)

PRÉSENTÉE LE 15 MAI 2014

À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE DE MODÉLISATION ET CALCUL SCIENTIFIQUE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Peng CHEN

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. A. Quarteroni, Dr G. Rozza, directeurs de thèse
Prof. J. S. Hesthaven, rapporteur
Prof. Y. Maday, rapporteur
Prof. T. Yizhao Hou, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

To my beloved family

Acknowledgements

Almost three years have passed in the blink of an eye for my PhD life. Looking back at the unique and rich experience, I would like to express my sincerest gratitude to all the people who make this possible.

First of all, I am deeply indebted to my advisor Prof. Alfio Quarteroni for providing me the precious opportunity to work closely with him for both my Master and PhD theses. From the perspectives of research topics to the proper choice of a single word in articles, he does not only care about the great detail but more importantly respect and value my interest and opinion. I appreciate it very much that he always puts the development of his students in the first place under whatever uncertainties. His generous support, unconditional trust, constant patience and incredible enthusiasm have substantially cultivated my academic growth. Beyond that, in every progress of my life, he is always there as a close friend for sharing, supporting and celebrating. It is my biggest fortune to have him as my advisor.

Secondly, I would like to thank my co-advisor Dr. Gianluigi Rozza for constantly being a great reference point during the whole period of my PhD study. My knowledge in the current advancement of many interesting fields has been significantly broadened through various material provided by him from conferences, workshops and other activities. Moreover, I have learned from him to be very cautious about citations in writing articles. He has kept great confidence in my research and high tolerance for my mistakes here and there. Whenever I have some request, he is always ready to offer me his immediate help. I am very grateful to him for keeping me original, happy and interested in my research.

A sincere acknowledgement goes to those people who have contributed to this thesis. In particular, I would like to thank Prof. Fabio Nobile and Dr. Lorenzo Tamellini for many enlightening discussions about the stochastic collocation method, which resulted in the work of its comparison with the reduced basis method in chapter 1 and their efficient combination in other chapters. I am grateful to Prof. Anthony Patera and Prof. Yvon Maday as well as their collaborators who play the pivotal role in developing reduced basis methods that I have benefited a lot for the design and analysis of new algorithms throughout my thesis. A nice conversation with Dr. Jing Li on her work of failure probability evaluation (co-authored with Prof. Dongbin Xiu) has inspired the idea of the work on accurate and efficient risk analysis presented in chapter 4. In this work, we also studied a goal-oriented adaptive algorithm pioneered by Prof. J. Tinsley Oden, to whom I am thankful for actively managing the publication of this chapter. Our work on high-dimensional uncertainty quantification problems in chapter 5 has been improved through communication with Prof. Michael Griebel and Prof. Jan Hesthaven, to whom I am indebted. The stochastic optimal control problems recently investigated independently by Prof. Christoph Schwab (on which we had a discussion in Bern), Prof. Max Gunzburger (who kindly handled our publications presented in chapter 2 and 6) and their co-workers have greatly influenced our study on this subject. I appreciate their helpful insights.

I sincerely acknowledge the members of my thesis committee for their precious time: Prof. Jan S. Hesthaven, Prof. Yvon Maday and Prof. Thomas Yizhao Hou, the latter being very inspiring for my further research. My deep gratitude goes to the jury president Prof. Thomas Mountford, one of my favourite teachers at EPFL for the courses Probabilité Avancée and Applied Stochastic Processes.

Acknowledgements

I would like to thank all my colleagues in the Chair of Modelling and Scientific Computing at EPFL who provide me an enjoyable environment all the time for both research and life. Special thanks go to my previous officemates Matteo Lesinigo and Cristiano Malossi, who helped me a lot not only for work but also in adapting myself to the amazing Italian culture. It was because of them that I had so much fun and learned so many new things for my PhD life! I am grateful to Simone Deparis, whose integrity, kindness and encouragement to me is very much appreciated. Many thanks to Ricardo Ruiz-Baier who has been a generous friend to me. I thank a lot Luca Dede' for all his helpful messages and his care of my progress in work and life. I am indebted to Toni Lassila, Andrea Manzoni and Fedrico Negri for their stimulating insights and excellent work in reduced basis methods, which have been very useful for my research. I would like to thank Simone Rossi and Paolo Tricerri for many interesting talks over nights, Gwenol Grandperrin and Samuel Quinodoz for being my French teacher, Radu Popescu and Andrea Bartezzaghi for their great companion in office, Matteo Lombardi, Davide Forti and Aymen Laadhari for a lot of happy chatting. Special thanks go to Laura Iapichino and Claudia Colciago, two beautiful girls always nice to me. I can not help remembering Marco Discacciati (my teacher of numerical approximation of PDE) and Matteo Astorino (my co-advisor of Master thesis) for their guide and inspiration. I am very grateful to Madam Annick Gaudin, Madam Corinne Craman and Madam Risse Catherine, who have made my working in Lausanne and traveling around the world very convenient.

My graduate life in EPFL has been more enjoyable because of many great friends. A special gratitude goes to Le Chen, who has helped me over the years starting even before I arrived in Switzerland. His generosity, frankness, and cheering attitude remain the sources of "positive energy" to me. Many thanks to Yuan Zhang for his companion during the hard times before our PhD journey. Over the years, all the chatting, running, mountain-hiking, bicycling with him have made my spare time very rich and amusing. I still remember clearly many nights studying in the math library with my Master friends Yankai Shao, Yi Wang, Ting Zhang, Jiayin Liu, Andy Shek, and pleasant traveling days in Switzerland and Europe with Jianwen Sun, Yixing Chen. Rounds of weekend leisure nights with math folks Han Wu, Guodong Zhou, Yun Bai, Ping Xi, Meiyue Shao, Fei Pu and computer folks Xiuwei Zhang, Min Ye, Mingfu Shao have brought me lots of fun. Thank you all my friends – you made my life wonderful here.

It has been my greatest fortune to meet a kind Swiss couple Edith and Martin Gamper, who bought me a metro ticket for free on the first day I arrived in Lausanne. Since then, we spent so much happy time together, my first Christmas eve at their home, my first visit to the National Museum and to the beautiful Lac de Joux, my first amazing sight of Mont Blanc with them, so many first times that I can not remember all. My friendship with them extends to their three sons: Stephan, Christian and Phillipe. I am grateful for the close friendship with Stephan, Syliviane and their lovely daughter Lucy who also live close to our place. I enjoyed a lot Christian and Richel's wedding in Paris. The food of our chef Phillipe is truly delicious. I am more than lucky that the whole Gamper's family take me as one of their family members in many big events. After Martin, we name our new born baby for their love.

To my grandparents and parents: because of you I came to this world and have grown up healthy and happy; you taught me to be honest, strong, generous and independent; you have given me everything you have and always care about me wherever I am. I can never thank you enough for your love.

To my beloved wife Lu: with you I started my journey all the way to the west– four years at hometown seeding our love, four for cultivating in the beautiful city Xi'an, after another four of blooming in Europe came our harvest season here – I have grown from a young boy of dream to a happy dad of love! You are the one I appreciate beyond any words and hope to live with for another four hundred years.

To my dear baby Martin: you made daddy's thesis full of joy with your first kick, your first hiccup, your first cry after birth. Looking at you in the eye, I can't help trying my best to be a better man for you!

Lausanne, January 3rd, 2014 Peng Chen

Preface

This thesis presents the major research work that have been carried out within three years starting from March 2011 at the Chair of Modelling and Scientific Computing (CMCS), in the Mathematics Institute of Computational Science and Engineering (MATHICSE) of École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. This thesis has been greatly benefited from the graduate courses at EPFL, and several international conferences, workshops and seminars listed in the CV at the end.

The problems, techniques, algorithms, analyses and numerical experiments of this thesis have been collected as a series of papers listed in the CV that have been published or submitted for publication in international journals. A large effort has been devoted to present the whole matter in a unified and coherent way with a harmonization of the style and notations. In particular, the preliminary chapter provides the common notations, useful tools and basic settings of the problems.

For the convenience of the reader and to properly acknowledge the co-authors, the corresponding references are provided at the beginning of each chapter.

Lausanne, January 3rd, 2014 Peng Chen

Abstract

The last few years have witnessed a tremendous development of the computational field of uncertainty quantification (UQ), which includes statistical, sensitivity and reliability analyses, stochastic or robust optimal control/design/optimization, parameter estimation, data assimilation, to name just a few. In all these problems, the solution of stochastic partial differential equations (PDEs) is commonly faced, for which many computational methods have been proposed, such as the extensively used Monte Carlo method and its several variants, the fast convergent stochastic Galerkin projection method and the nonintrusive stochastic collocation method. The large advancement of these computational methods with sparse and adaptive techniques has enabled efficient solution of the aforementioned UQ problems that feature high dimensionality, low regularity and arbitrary probability measures. However, when it becomes very expensive to solve the underlying deterministic PDEs, e.g., only a few tens or hundreds of full solutions are affordable in practice, these computational methods can not be applied directly since they may need millions of full solutions, or even beyond, in order to achieve a certain accuracy.

In this thesis, we develop, analyze and demonstrate novel stochastic computational strategies and algorithms based on model order reduction techniques, in particular based on reduced basis methods, to tackle this challenge in solving several typical UQ problems. We first compare the convergence properties and computational costs of the reduced basis method and the sparse grid stochastic collocation method, and demonstrate that the former is much more efficient than the latter without loss of accuracy in solving large-scale and high-dimensional UQ problems. In dealing with arbitrary probability measures, we propose a weighted reduced basis method inspired by the generalized polynomial chaos, and establish explicitly a priori error estimates for both one-dimensional and multidimensional stochastic/parametrized problems. A weighted empirical interpolation method with improved convergence property is proposed in order to decompose nonaffine random fields, which paves the way for effective application of the reduced basis method in solving more general UQ problems. A hybrid and goal-oriented adaptive reduced basis method with certification is proposed to efficiently and accurately solve a large class of UQ problems, involving pointwise evaluation, in particular failure probability for reliability analysis. Moreover, taking advantage of the sparsity and reducibility of UQ problems, we develop an adaptive and reduced computational framework that enables precise detection of the distinctive importance and the interaction of different dimensions, as well as automatic construction of a generalized sparse grid and reduced basis approximation of the quantities of interest.

Besides the development and demonstration of the model order reduction techniques in solving various demanding forward UQ problems, a large effort of this thesis has been devoted to the analysis and the efficient solution of inverse UQ problems, in particular stochastic optimal control problems. We succeed in proving not only the existence but also the uniqueness of the optimal solution via a stochastic saddle point formulation in the case of elliptic and Stokes constraints. A detailed analysis is carried out for the stochastic regularity of the optimal solution w.r.t. the random input data under certain smoothness hypothesis. We tailor the main ingredients of the developed adaptive and reduced computational strategy to solve stochastic optimal control problems with several different PDE constraints. The efficiency and accuracy of this strategy demonstrate its potentials in solving more general large-scale and high-dimensional inverse UQ problems with arbitrary probability measures.

Abstract

Keywords: uncertainty quantification, stochastic partial different equations, sparse approximation, stochastic collocation method, reduced basis method, reliability analysis, stochastic optimal control

Résumé

Les dernières années ont vu un développement considérable du champ de calcul de la quantification des incertitudes (QI), qui comprend des analyses statistiques, de sensibilité et de fiabilité, des contrôle optimal/conception optimal/optimisation stochastique ou robuste, estimation de paramètres, l'assimilation de données, pour n'en nommer que quelques-uns. Dans tous ces problèmes, la résolution d'équations aux dérivées partielles (EDP) stochastiques est souvent confronté, pour la quelle de nombreuses méthodes de calcul ont été proposées, telles que la méthode polulaire de Monte-Carlo et ses différentes extensions, la méthode de projection de Galerkin stochastique de convergence rapide et la méthode de collocation stochastique non-intrusive. Le grand progrès de ces méthodes de calcul avec des techniques clairsemées et adaptatives a permis une solution efficace des problèmes QI mentionnés ci-dessus qui comportent une grande dimension, une régularité faible et des mesures de probabilité arbitraires. Cependant, quand il devient très coûteux de résoudre les EDP déterministes sous-jacentes, par exemple, que quelques dizaines ou centaines de solutions complètes sont réalisable dans la pratique, ces méthodes de calcul ne peuvent pas être appliquées directement car elles peuvent avoir besoin des millions de solutions complètes ou d'avantage afin d'atteindre certaine précision.

Dans cette thèse, nous développons, analysons et démontrons de nouvelles stratégies stochastiques de calcul et des algorithmes basés sur des techniques de réduction de l'ordre de modèle, en particulier les méthodes des bases réduites, pour relever ce défi dans la résolution de plusieurs problèmes typiques de l'QI. Nous comparons d'abord les propriétés de convergence et les coûts de calcul de la méthode de bases réduites et la méthode de collocation stochastique sur grilles clairsemées, et démontrons que la première est beaucoup plus efficace que la dernière sans perte de précision dans la résolution des problèmes QI de grande échelle et de grande dimension. En traitant des mesures de probabilité arbitraires, nous proposons une méthode de bases réduites pondérée inspirée par les polynômes de chaos généralisé, et établissons explicitement des estimations d'erreur a priori pour les problèmes stochastiques/paramétrés à la fois unidimensionnel et multidimensionnels. Une méthode d'interpolation empirique pondérée avec une meilleure propriété de convergence est développée afin de décomposer des champs aléatoires nonaffine, qui ouvre la voie à l'application effective de la méthode des bases réduites à résoudre des problèmes plus généraux de l'QI. Une méthode des bases réduites hybride et adaptative basée sur les objectifs avec certification est proposée pour résoudre efficacement et avec précision un grand nombre de problèmes QI concernant l'évaluation ponctuelle, en particulier probabilité de défaillance pour l'analyse de fiabilité. Par ailleurs, profitant de la rareté et de réductibilité des problèmes QI, nous développons un cadre d'adaptation et de calcul réduit qui permet une détection précise de l'importance particulière et de l'interaction des différentes dimensions, ainsi que la construction automatique d'une approximation des quantités d'intérêt de grilles clairsemées généralisée et de bases réduites.

Outre le développement et la démonstration des techniques de réduction d'ordre de modèle dans la résolution d'une multiplicité de problèmes direct de QI exigeants, un grand effort a été consacrée dan cette thèse à l'analyse et la résolution efficace des problèmes inverses de QI, en particulier des problèmes de contrôle optimal stochastique. Nous réussissons à prouver non seulement l'existence, mais aussi l'unicité de la solution via une formulation de point selle stochastique dans le cas de

contraintes elliptiques et Stokes. Une analyse détaillée est effectuée pour la régularité de la solution stochastique optimal en ce qui concerne les données d'entrées aléatoires sous certaines hypothèses de régularité. Nous adaptons ces ingrédients principaux de la stratégie de calcul adaptative et réduite pour résoudre les problèmes de contrôle optimal stochastique avec plusieurs différentes contraintes de EDP. L'efficacité et la précision de cette stratégie démontrent son potentiel à résoudre des problèmes inverses de QI plus généraux, de grande échelle et de grande dimension avec mesures de probabilité arbitraires.

Mots-clés : quantification des incertitudes, équations aux dérivées partielles stochastiques, approximation creuse, méthode de collocation stochastique, méthode de bases réduites, analyse de fiabilité, contrôle optimal stochastique

Contents

Acknowledgements v

Preface vii

Abstract (English/Français) ix

Introduction 1

Preliminary 12

I Forward Uncertainty Quantification Problems: Challenges and Solutions 21

1 Comparison of stochastic collocation and reduced basis methods 23

1.1 Benchmark model . . . 24

1.2 Stochastic collocation method . . . 24

1.2.1 Univariate interpolation . . . 25

1.2.2 Multivariate tensor product interpolation . . . 25

1.2.3 Sparse grid interpolation . . . 26

1.3 Reduced basis method . . . 28

1.3.1 Training set . . . 28

1.3.2 Greedy algorithm . . . 28

1.3.3 A posteriori error estimate . . . 29

1.3.4 Offline-online computational decomposition . . . 29

1.4 Comparison of convergence analysis . . . 30

1.4.1 Preliminary comparison of convergence results . . . 30

1.4.2 Direct comparison of approximation errors . . . 34

1.5 Comparison of computational costs . . . 36

1.6 Numerical experiments . . . 38
1.6.1 Numerical experiments for a univariate problem . . . 38
1.6.2 Numerical experiments for multivariate problems . . . 41
1.6.3 Numerical experiments for higher dimensional problems . . . 44
1.7 Summary . . . 45

2 A weighted reduced basis method for arbitrary probability measures 47
2.1 A weighted reduced basis method . . . 48
2.2 Regularity and a priori error estimates . . . 49
2.2.1 Regularity results . . . 49
2.2.2 A priori convergence analysis . . . 52
2.3 Numerical examples . . . 56
2.3.1 One-dimensional problem . . . 57
2.3.2 Multidimensional problem . . . 58
2.4 Summary . . . 59

3 Decomposition of nonaffine fields – a weighted empirical interpolation 63
3.1 Weighted empirical interpolation method (wEIM) . . . 64
3.2 A priori convergence analysis . . . 65
3.3 Numerical experiments . . . 69
3.3.1 Parametric function in one dimension – geometric Brownian motion . . . 69
3.3.2 Parametric function in multiple dimensions – Karhunen–Loève expansion . . . 70
3.3.3 Parametric PDEs – application to the reduced basis method . . . 71
3.4 Summary . . . 73

4 Hybrid and goal-oriented adaptive reduced basis methods for risk analysis 75
4.1 Problem statement . . . 76
4.2 Reduced basis methods for evaluation of failure probability . . . 77
4.2.1 The reduced basis method . . . 77
4.2.2 A hybrid reduced basis method . . . 78
4.2.3 A goal-oriented adaptive reduced basis method . . . 79
4.2.4 Remarks on approximation error and computational cost . . . 80
4.3 Extension to more general PDE models . . . 80

4.3.1 Noncompliant problems . . . 81
4.3.2 Unsteady problems . . . 83
4.3.3 Nonaffine problems . . . 86
4.4 Numerical experiments . . . 91
4.4.1 Benchmark models . . . 91
4.4.2 Noncompliant problems . . . 95
4.4.3 Unsteady problems . . . 97
4.4.4 Nonaffine problems . . . 99
4.5 Summary . . . 101

5 Breaking the curse of dimensionality – sparsity and reducibility 103

5.1 High-dimensional uncertainty quantification . . . 105
5.1.1 Formulation of UQ problems . . . 105
5.1.2 Computational challenges . . . 106
5.2 Verified dimension adaptive hierarchical approximation . . . 107
5.2.1 Hierarchical interpolation and integration in one dimension . . . 107
5.2.2 Hierarchical Smolyak sparse grid in multiple dimensions . . . 110
5.2.3 Dimension adaptation for high-dimensional problems . . . 112
5.2.4 Comparison remarks . . . 114
5.3 Adaptive and weighted reduced basis method . . . 116
5.4 Numerical experiments . . . 119
5.4.1 Hierarchical construction with verification . . . 119
5.4.2 Sobol functions featuring strong interaction . . . 121
5.4.3 Approximation with arbitrary probability measure . . . 123
5.4.4 High-dimensional functions featuring sparsity . . . 124
5.4.5 Heat diffusion in thermal blocks . . . 126
5.4.6 Groundwater flow through porous medium . . . 130
5.5 Summary . . . 133

II Analyses and Fast Solvers for Stochastic Optimal Control Problems 135

6 Stochastic elliptic optimal boundary control with random advection field 137

6.1 Stochastic Robin boundary control problem . . . 138

6.1.1 Problem definition . . . 138
6.1.2 Stochastic saddle point formulation . . . 139
6.2 Stochastic regularity . . . 142
6.3 Approximation and error estimates . . . 147
6.3.1 Finite element approximation in physical space . . . 147
6.3.2 Collocation approximation in stochastic space . . . 148
6.3.3 Convergence for approximating stochastic optimal control problem . . . 149
6.4 Numerical results . . . 150
6.5 Summary . . . 153

7 Reduced basis method for stochastic elliptic optimal control problems 155
7.1 Problem statement . . . 156
7.1.1 Constrained optimal control problems . . . 157
7.2 Saddle point formulation . . . 157
7.2.1 Stochastic optimality system . . . 157
7.2.2 Saddle point formulation . . . 158
7.3 Numerical approximation . . . 158
7.3.1 Finite element method . . . 159
7.3.2 Weighted reduced basis method . . . 160
7.4 Numerical tests . . . 164
7.4.1 One-dimensional problems . . . 164
7.4.2 Moderate-dimensional problems . . . 167
7.4.3 High-dimensional problems . . . 168
7.5 Summary . . . 169

8 Stochastic optimal control problem constrained by Stokes equations 171
8.1 Problem statement . . . 172
8.1.1 Stochastic Stokes equations . . . 172
8.1.2 Finite dimensional assumption . . . 173
8.1.3 Constrained optimal control problem . . . 174
8.2 Saddle point formulation . . . 175
8.2.1 Optimality system . . . 175
8.2.2 Saddle point formulation . . . 176

8.2.3 Equivalence, uniqueness and stability estimates . . . 177
8.3 Stochastic regularity . . . 178
8.4 Numerical approximation . . . 183
8.5 Multilevel and weighted reduced basis method . . . 184
8.5.1 Reduced basis approximation . . . 185
8.5.2 A multilevel greedy algorithm . . . 186
8.5.3 A weighted a posteriori error bound . . . 187
8.5.4 Offline-online decomposition . . . 188
8.6 Error estimates . . . 190
8.6.1 Finite element approximation error . . . 190
8.6.2 Global error estimate . . . 191
8.7 Numerical experiments . . . 192
8.7.1 Isotropic case . . . 192
8.7.2 Anisotropic case . . . 193
8.8 Summary . . . 195

Conclusions and Perspectives **196**

Bibliography **213**

Curriculum Vitae **215**

Introduction

This thesis is devoted to developing, analyzing and verifying novel stochastic computational strategies and algorithms based on model order reduction techniques, in particular a reduced basis method, for the solution of several specific uncertainty quantification problems. We start by illustrating uncertainty quantification problems; then we provide a short survey of the state of the art in stochastic computational methods. We close this introduction by a presentation of the main contributions of this thesis and its organization.

Uncertainty quantification problems – motivations and scopes

Thanks to the fast development of computing hardware and numerical algorithms, the last few decades have witnessed tremendous growth of mathematical modelling and computational simulation, which nowadays become a routine as the third pillar in many scientific research and practical engineering applications besides theoretical investigation and experimental exploration. For example, the modelling and simulation of blood flow in human cardiovascular system have undergone large advancement because of better understanding of the morphology and functionality of the system, the availability of abundant clinical data as well as fast growing of computational resources and algorithms [75]. Specifically, various mathematical models targeted for part of cardiovascular system such as a portion of artery where diseases locate, or the entire arterial tree that delivers oxygen and nutrients to the whole body, have been established by fluid conservation laws and structural deformation theories. In particular, fluid (e.g., Navier-Stokes) equations and elastic (or viscoelastic) equations are coupled together to characterize the fluid structure interaction property of blood flow; geometrical multiscale models are established for large and small arteries; models for tissue perfusion, mass transfer, bypass design, to name just a few, have also been developed with specific objectives. Meanwhile, the development of associated computational techniques have greatly improved the applicability of cardiovascular modelling and simulation to conduct physiology and pathology investigations and perform clinical and/or surgical optimizations.

However, for all these mathematical models with different objectives, the corresponding simulation results may differ from reality or observations due to various uncertainties that are inevitably encountered in the modelling and simulation processes. First of all, uncertainties may arise from the model inputs [46, 135], including physical parameters representing material properties such as Young modulus of the blood vessel, initial or boundary conditions prescribed as blood flow rates or stresses, external loadings such as surrounding tissue pressure and working effort, computational geometries that are extracted and reconstructed from patients by MRI or some other techniques. These input uncertainties can be hardly determined as deterministic quantities, either because they may possess intrinsic randomness or because the measurements are not sufficient to produce precise inputs. Another source of discrepancy between the simulation results and reality comes from approximation errors, e.g., discretization errors in numerical approximation of the fluid structure interaction equations, and/or potential flaws during the computational implementation of the numerical algorithms. A more

fundamental source of uncertainties may be the inappropriate construction or oversimplification of the models themselves, e.g., models neglecting viscoelastic effect or more complex local fluid field, because of limited understanding of the complexity, diversity and variability of the underlying physical processes of the blood flow in multiple time and space scales. In order to incorporate these different and influential uncertainties and conduct more realistic and robust modelling and simulation tasks, three interrelated research branches in computational science and engineering have been developed in recent years [57]: (i) *uncertainty quantification*, which deals with propagation of various uncertainties from inputs to outputs of given physical systems that are described by mathematical models, and inversion of available data, experimental measurements or objectives for the outputs with the aim of reducing the uncertainties in the inputs; (ii) *verification*, that aims at verifying the accuracy of the approximation of the original mathematical models; (iii) *validation*, that studies how to validate the efficacy of the mathematical models against the underlying physical processes.

This thesis focuses on the first branch of solving uncertainty quantification problems in both forward and inverse settings, as illustrated in Figure 1, where the few topics of this thesis are highlighted with bold font. In short, the object under study can be assembled in the following three components. At the core, the underlying physical processes are described by some appropriate mathematical models, for instance elliptic or parabolic equations for heat or mass transfer, elastic or viscoelastic equations for structure deformation, Stokes or Navier-Stokes equations for fluid flow. In order to set up a mathematical model, we need to provide the necessary model inputs, such as physical parameters, boundary or initial conditions, computational geometries. However, due to lack of knowledge or intrinsic randomness, as mentioned before, the inputs may not be prescribed with deterministic values due to various uncertainties. Therefore, the model outputs, e.g., the stochastic solution itself, functional of the solution or multiple quantities of interest related to the solution, may only be evaluated or measured statistically. According to the available data and the objectives of the study, two different kinds of problems can be faced: one kind is known as forward problems, i.e., given random inputs of the model, we are interested in evaluating some statistical outputs; the other kind is known as inverse problems, i.e., provided observed or measured outputs, we aim at reducing the uncertainties of inputs or updating the state variables.

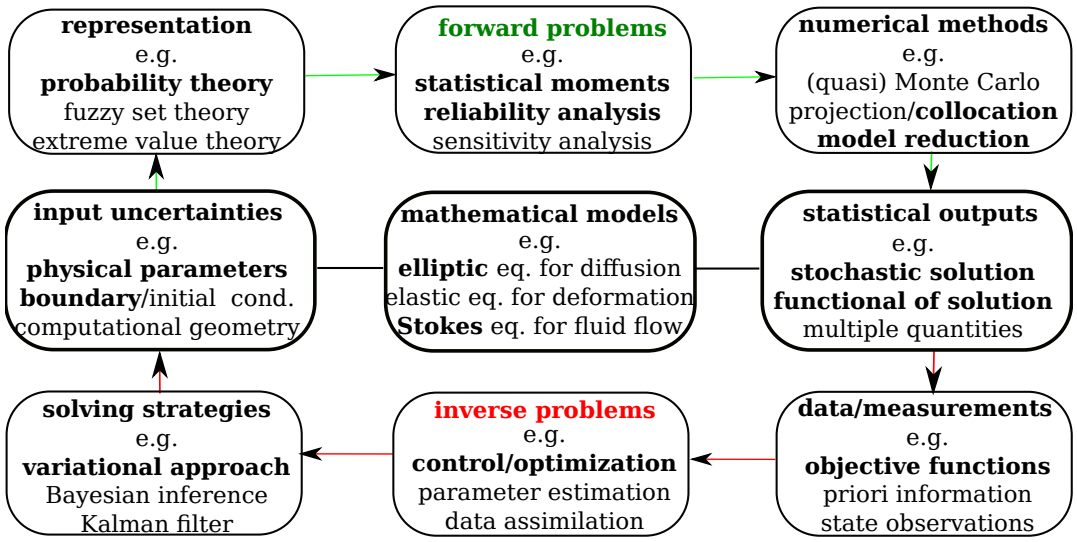

Figure 1: Schematic representation of solving forward and inverse uncertainty quantification problems

To solve a forward uncertainty quantification problem, the first step is to identify and represent the input uncertainties in a concrete mathematical structure, e.g., in the framework of probability, fuzzy set

or extreme value theories. Secondly, before solving the mathematical models with random inputs, the objectives (quantities of interest) should be clearly defined, which may involve computing statistical moments, e.g., expectation or variance, performing risk or reliability analysis via evaluation of failure probability, conducting sensitivity analysis to seek the most influential uncertainties. Based on different objectives of the forward problems, suitable numerical methods need to be used to solve the models.

On the other hand, in solving a inverse uncertainty quantification problem, we are provided with some data or measurements on the statistical outputs, such as desirable objective functions of the outputs, some a priori information on the quantities of interest or observations of the state variables. With these data at hand, our aims could be assimilating them in the model to update the state variables, controlling, optimizing or estimating certain input variables, e.g., force, geometry or physical parameters, in order to drive the outputs as close as possible to the objective functions or observations. Depending on different aims, the inverse problems can be classified as optimal control, optimization, parameter estimation, data assimilation, etc., which are generally ill-posed from the mathematical viewpoint. To facilitate the solution of these ill-posed problems, different computational strategies can be adopted accordingly, for example, penalized variational approach based on *Lagrange multipliers*, regularized Bayesian inference by Markov chain Monte Carlo methods or optimal maps, Kalman filters, etc.

Various mathematical and computational challenges are encountered in solving the forward and inverse uncertainty quantification problems. The first challenge is how to effectively identify and represent the input uncertainties or statistical observations of different types, especially how to compress high-dimensional uncertainties into a low dimensional space, while capturing the important properties of the uncertainties. Many advanced statistical methods come into play, including linear or nonlinear regression, principle component analysis (Karhunen–Loève expansion), maximum entropy, etc.. Provided that the uncertainties are well represented with appropriate mathematical structure, a further challenge is to study whether the models with random inputs, most often formulated as stochastic partial differential equations, are well-posed or not in terms of existence, uniqueness and regularity of the stochastic solution. This challenge is more involving in stochastic and functional analyses, for which the theories are far from mature to deal with nonlinear, multiscale and multiphysics models. Nowadays, a more significant and relevant challenge for computational science and engineering is how the models, even if their well-posedness is not fully understood, can be solved efficiently and accurately by a computer. In particular, this challenge is naturally confronted when one has to efficiently solve the stochastic models with arbitrary probability measures when the uncertainties are represented in the probability framework, to practically harness the total computational burden when the underlying model at one stochastic realization becomes very expensive to solve, to accurately approximate the solution in a high-dimensional stochastic space facing the common difficulty well-known as the “curse of dimensionality". *This thesis is mainly devoted to developing, analyzing and verifying stochastic computational strategies and algorithms specific to several different uncertainty quantification problems with the aim of reducing their computational complexity.*

Stochastic computational methods – state of the art

In order to solve both the forward and the inverse uncertainty quantification (hereafter abbreviated as UQ) problems in an efficient and reliable way, the essential task is to design and analyze efficient and accurate stochastic computational methods, which has been the main topic of the UQ research community in recent years. As a matter of fact, various stochastic computational methods have been developed and analyzed depending on the structures and properties of the UQ problems, including –(far from being extensive and complete)– perturbation method, Neumann expansion methods, Monte Carlo methods, stochastic Galerkin methods, *stochastic collocation methods*, and more recently methods based on *model order reduction techniques* [72, 80, 190, 208, 10, 207, 8, 137, 136, 16, 151, 89, 31, 24, 194, 18].

The perturbation method [109] based on Taylor expansion of random functions was developed for random functions featuring only small fluctuation around a deterministic expectation, while Neumann expansion method [210] uses Neumann expansion of the inverse of stochastic operator around a deterministic operator. Both methods are applicable only to deal with small uncertainties, thus suffer from inevitable errors and extremely involved for high order expansions. The most commonly used "brute-force" Monte Carlo method [72] basically samples points in probability space and simplifies a stochastic system to a deterministic one at the sampling points, so that only a deterministic system needs to be solved and statistical information can be easily obtained by taking moments. However, it converges very slowly with a convergence rate of $1/\sqrt{N}$ for N samples and becomes prohibitive for achieving accurate results, especially for those stochastic systems that are already quite computationally intensive in their deterministic settings. In order to accelerate the convergence, several improvements have been proposed such as quasi Monte Carlo [147, 62], Latin hypercube sampling [125, 103], multi-level Monte Carlo [95, 81]. Sampling the most representative points in order to accelerate Monte Carlo methods become critical in practical applications.

The stochastic Galerkin method has recently received increasing attention in plenty of applications [193, 205, 83]. It relies on spectral expansion of the random functions on some polynomial chaos, for instance Hermite polynomials of independent random variables, and the Galerkin approach to approximate the expansion in deterministic space [80, 7]. By adopting the techniques of the deterministic Galerkin approximation, both a priori and a posteriori error estimations can be derived [10]. Moreover, it enjoys fast convergence if the solution is sufficiently regular [55, 54]. The stochastic Galerkin method has also been extensively used for practical applications using generalized polynomial chaos [208] for uncertainties with more general distributions inspired by the structural coherence of different types of orthogonal polynomials and stochastic processes [186]. However, a very large algebraic system is typically associated to the stochastic Galerkin approach, which requires the availability of efficient solvers [67], such as Krylov iterative solvers with appropriate preconditioners.

The stochastic collocation method was developed from the non-intrusive deterministic collocation method [164] and sparse grid techniques [33]. It finds its application in a variety of fields, for instance chemical and environmental engineering [139] in the early years. Nevertheless, its numerical properties such as error convergence analysis, computational cost, as well as various extensions has been discovered and developed only in the recent years [207, 8]. In principle, the stochastic collocation method employs multivariate polynomial interpolations for the integral in the variational formulation of the stochastic system with respect to probability space rather than the Galerkin approximation in the spectral polynomial space. Due to the heavy computation of a deterministic system at each collocation point in high-dimensional space, isotropic or anisotropic sparse grids with suitable cubature rules [148, 149, 9] were successfully applied and analyzed for the stochastic collocation method to reduce the computation. Moreover, hierarchical construction of a generalized sparse grid [79, 110] have also been developed for the application of the stochastic collocation method. This method is preferred for more practical applications because it entails the advantages of both direct computation as Monte Carlo method and fast convergence as stochastic Galerkin method [12].

Despite the great development of the sparse techniques for stochastic Galerkin and collocation methods, there are still several common and major challenges to face, for instance low regularity of the solution in stochastic space and the already mentioned curse of dimensionality, which require a large number of stochastic realizations of the input uncertainties for accurately capturing the random outputs, resulting in prohibitive computations by directly using these stochastic computational methods. In addition to selecting the appropriate spectral basis for the expansion of different probability distributions of random inputs, two other approaches to capture local behaviour of the solution and to alleviate high-dimensional computational cost have been developed: the first one is to generalize the polynomials from globally smooth functions to piecewise polynomial basis [190], wavelet basis [119] and multi element polynomial chaos [203]; the second one is to efficiently reduce the computational cost by adaptivity - adaptive choice of polynomial basis [198], adaptive element selection in

multielement polynomial chaos [74], adaptive spectral decomposition [151], as well as adaptive sparse grid collocation [127]. More efforts are still in large demand for considerable reduction of stochastic computation loads in UQ problems featuring low regularity and high dimensionality.

Rather different from the methods and approaches presented above, especially the stochastic Galerkin and collocation methods based on dictionary bases, another type of stochastic computational method based on model order reduction techniques can also be applied and characterized by a large potential to be able to accelerate the solution of UQ problems, which include methods based on proper orthogonal decomposition [16, 204] or generalized spectral decomposition [151, 154], balanced truncation method [141, 89], Krylov-based method [70, 181], reduced basis method [178, 20]. The basic idea behind these model order reduction techniques is to project the associated large algebraic system, by using a combination of techniques such as singular value decomposition and/or in combination of greedy algorithms, to a small system that can effectively capture almost all the information carried by the original model. Among all of these reduction techniques, one of the most appealing in solving UQ problems is the reduced basis method. Briefly speaking, it seeks to parametrize the random inputs and select the most representative points in the parameter space by means of a greedy algorithm endowed with an a posteriori error estimate [178, 158, 24]. The essential idea for harnessing heavy computational burden is to separate the whole computational procedure into an offline stage and an online stage. During the former, the most computationally demanding elements (sampling parameters, assembling matrices and vectors, solving and collecting snapshots of solutions, etc.) are computed and stored once and for all. While during the online stage, only the parameter related elements are left to be computed and a small and very inexpensive Galerkin approximation problem to be solved. The reduced basis method was initially introduced for structure analysis [150] and recently has undergone vast development in theory [131, 178, 158, 87, 86, 49] and applied to many engineering problems [162, 51, 163, 117, 175, 41, 45]. This method is similar to the stochastic collocation method in terms of sampling and differs from the latter method because it uses a posteriori error estimate, and thus providing a great potential in the reduction effort for the total number of full solves of the original large-scale model; consequently, it helps in breaking the curse of dimensionality of solving high dimensional UQ problems whenever the stochastic solution manifold lies in a low dimensional probability space.

In solving inverse UQ problems, in particular the stochastic optimal control problems considered in this thesis, a very large, naturally coupled and ill-conditioned system is obtained from variational approach with Lagrange multiplier [200], which involves the forward model, an adjoint model and an additional system-closing model (e.g., the equation representing an optimality condition). Besides the difficulties in suitable application of the aforementioned stochastic computational methods, an additional computational challenges comes from the necessity of solving many times the forward and the adjoint stochastic model by iterative methods, e.g., steepest gradient method [161, 200], or solving once the coupled optimality system by “one-shot" method [185, 169]. To alleviate the computational effort by iterative method, sequential quadratic programming [197] and trust-region algorithm [112] have been applied and proved to be efficient, while to reduce the computational cost in solving the ill-conditioned system by one-shot method, efficient preconditioning techniques have been developed [185, 169]. However, when solving the underlying stochastic model becomes too expensive, it is only affordable for tens or hundreds of full solve in practice, making the methods introduced above impossible to be directly applied since the number of samples needed easily goes beyond the amount could be handled, especially for high-dimensional problems. Since quantities of interest usually reside in low dimensional manifold, model order reduction techniques may be applied using proper orthogonal decomposition or reduced basis approximation for parametrized optimal control problems [113, 132, 107]. As recently found in [57], an important area of future work is the use of model reduction for optimization under uncertainty.

Thesis contributions – methodologies, algorithms and theories

Methodological contributions

We carry out a detailed comparison between the reduced basis method and the most advanced sparse grid stochastic collocation method in solving a benchmark uncertainty quantification problem [50]. Conclusions are drawn from the comparison that model order reduction techniques, specifically the reduced basis method, are both more efficient and accurate in solving large-scale and high-dimensional uncertainty quantification problems.

Moreover, we develop a verified adaptive and reduced computational framework in solving high-dimensional UQ problems by taking advantage of the computational opportunities of sparsity in stochastic dimension and reducibility in the model order [42]. This framework can not only facilitate automatic detection of the sparsity – distinct importance and interaction of different stochastic dimensions but also extract and reconstruct the feature of stochasticity of the quantities of interest with affordable computational effort.

Furthermore, we tailor and apply this computational framework in solving stochastic optimal control problems constrained by partial different equations, which feature typical difficulties of solving inverse uncertainty quantification problems, and demonstrate that this framework can dramatically reduce the total computational cost without sacrificing numerical accuracy [47, 43, 45].

Algorithmic contributions

While the development of the generalized polynomial chaos [208] has brought remarkable progress and influences in solving uncertainty quantification problems that feature arbitrary probability measures, less attention has been paid in the model order reduction community in order to treat arbitrary probability measures. In this thesis, a weighted algorithm associated with the probability distribution of the input uncertainties for both the reduced basis method [49] and the empirical interpolation method [48] is proposed and demonstrated to be very efficient in handling random variables with arbitrary probability distribution. The algorithm is rather simple yet it can effectively capture the most representative feature of stochastic solution with less modes, and can compete with the most accurate Gauss quadrature formula for integration in terms of accuracy and nodes, even in a single dimension.

For several types of uncertainty quantification problems, in particular for risk analysis, surrogate or reduced model techniques have been criticized due to the flaw of accuracy [121], i.e., producing an inaccurate or even an erroneous quantity of interest. In this thesis, we develop a hybrid algorithm based on the reduced basis method, which not only eliminates the flaw of accuracy but also tremendously reduces the computational cost in Monte Carlo sampling [41]. In order to further reduce the number of reduced bases that require full solves of the underlying model, a goal-oriented adaptive algorithm is proposed and proved to achieve the same accuracy as the hybrid algorithm and needs much less full solves. The hybrid and goal-oriented adaptive algorithms are successfully extended to compute failure probability for a series of models based on different types of partial differential equations [41].

In solving stochastic optimal control problems constrained by partial differential equations, it is a common practice to resolve the corresponding stochastic optimality system coupling a state equation, an adjoint equation and an optimality condition. This requires not only the deterministic approximation of the optimal solution in the physical space by either iterative approach or one-shot approach, resulting in heavy computational effort in the first stage, but also the stochastic approximation in the probability space that leads to much heavier computational demand in a further stage. In this thesis, we take advantage of the fact that the stochastic optimal solution resides in a low-dimensional manifold to tailor an adaptive and multilevel algorithm based on the reduced basis method and the sparse grid stochastic collocation method [47, 43, 45]. This algorithm is proved to be able to considerably alleviate

the total computational effort for the global approximation, and thus it is suitable for more general large-scale stochastic and robust optimization problems.

Theoretical contributions

Since the beginning of the development of the reduced basis method, the main research efforts in the literature have been focused on deriving a posteriori error estimate for various models. A priori error estimate was only obtained in [131] for a particular elliptic equation in a single parametric dimension, and in [30, 20] in an indirect way by comparison to Kolmogorov N–width. In this thesis, we derive a priori convergence analysis for the reduced basis method, in particular with the weighted algorithm, in both a single parametric dimension and multiple dimensions based on Fourier analysis and the stochastic regularity of the model solution with respect to the input random variables [49]. The derivation is carried out for elliptic equation but can be readily applied to more general models as long as we can obtain the stochastic regularity of the solution, which depends on the regularity of the input data, by the same procedure as in this thesis. Moreover, we prove that the reduced basis method converges at least as fast as the stochastic collocation method [50].

The empirical interpolation method was originally proposed in [11] to decompose nonaffine fields for the purpose of efficient application of reduced basis method, where a rather crude a priori convergence result was obtained. Recently, the result was refined with comparison to Kolmogorov N–width in the context of multipropose interpolation procedure under more general settings in [129]. In this thesis, we succeed in improving the a priori convergence result obtained in [129] by a factor of 2^N in the development of the weighted empirical interpolation method [48]. For its proof, we have adopted the ideas used in [20] for the prove of an indirect a priori error estimate of a weakly greedy algorithm. In fact, the construction of an efficient empirical interpolation operator also employs a weakly greedy algorithm.

The stochastic optimal control problems constrained by partial differential equations are ill-posed problems, which pose typical difficulties for other inverse uncertainty quantification problems. Even when appropriate regularization is applied, which guarantees the existence of the stochastic optimal solution by Lions' argument [123] as proved in [91], the uniqueness of the solution has not been proved. In this thesis, we prove both existence and uniqueness of the stochastic optimal solution in a tensor product stochastic Hilbert space for several different models, including the linear diffusion [43], advection-diffusion [47] and Stokes equations [45], with both boundary and distributed control functions. We also obtain the stochastic regularity of the stochastic optimal solution via a saddle point formulation and using Brezzi's theorem [27]. Moreover, a priori error analysis for the finite element - stochastic collocation - reduced basis approximation of the optimal solution in both physical and stochastic space is carried out in detail and verified by numerical experiments in different settings.

Limitations and potentials

The mathematical models we have considered are admissibly simple, such as those based on linear elliptic equations, parabolic equations, and Stokes equations. Consequently, immediate applications of the theories and algorithms are limited to these types of equations. Moreover, the scopes of the uncertainty quantification problems we have studied are also limited to several typical requests, e.g., evaluation of statistical moments, computational of failure probability for risk analysis, solving stochastic optimal control problems with linear constraints. Particularly, more general inverse problems such as nonlinear optimal control, parameter identification and data assimilation, which bear further computational challenges, are not addressed in this thesis. More specific limitations will be provided at the end of each chapter and in the conclusions of this thesis.

Fortunately, model order reduction techniques, in particular the reduced basis method, are not re-

stricted to the models here addressed and have been applied to nonlinear elliptic equations [86], Navier–Stokes equations [60], Maxwell equations [51], and even multiscale [1] and multiphysics [117] models. Meanwhile, model reduction techniques have also been developed and applied to solve shape optimization, parameter identification and data assimilation problems [133, 122, 37]. Therefore, we hope that the concrete demonstration of the superior performance of the stochastic computational strategies and algorithms in tackling several common computational challenges developed in this thesis can be profitably applied in solving more practical and complex uncertainty quantification problems.

Thesis organization

The thesis is presented through addressing a series of issues in developing model order reduction techniques, in particular the reduced basis method, to solve uncertainty quantification problems. We introduce the basic notations, function spaces, useful tools and common settings of uncertainty quantification problems in the preliminary chapter. The first part of this thesis (chapter 1–5) is mainly focused on forward problems and the second part (chapter 6–8) is specifically devoted to stochastic optimal control problems. In the last chapter, we summarize some general conclusions and list a few perspectives. More details of the main body of the thesis are provided as follows.

Chapter 1: Comparison of stochastic collocation and reduced basis methods

In this chapter, we compare the convergence property and the computational cost of the stochastic collocation method and the reduced basis method in solving a simple benchmark UQ problem. The convergence rates of the two methods are summarized and compared based on the results available in the literature. Moreover, we prove that the reduced basis approximation error is bounded from above by the stochastic collocation approximation error. The computational cost of the two methods is compared for both offline construction and online evaluation. Furthermore, an efficient combination of the reduced basis method and the stochastic collocation method is demonstrated to feature a fast evaluation of statistical moments of the stochastic solution. Conclusions are drawn from the comparison that the stochastic collocation method is preferred for small-scale and low-dimensional problems, while the reduced basis method performs better for large-scale and high-dimensional problems, as supported by our numerical experiments.

Chapter 2: A weighted reduced basis method for arbitrary probability measures

In this chapter, input uncertainties with arbitrary probability measures are first considered for more efficient application of the reduced basis method, which is currently only used for stochastic problems with uniformly distributed random inputs or parameter space with Lebesgue measure [24]. In order to deal with more general stochastic problems with other distributed random inputs, we propose and analyze a new version of the reduced basis method with a weighted a posteriori error bound and name it as “weighted reduced basis method". The basic idea is to suitably assign a larger weight to samples that are more important or have a higher probability to occur than the others according to either the probability distribution function or some other available weight function depending on the specific application at hand. The benefit is to lighten the reduced space construction using a smaller number of bases without lowering the numerical accuracy. Moreover, we carry out a priori convergence analysis of the proposed method based on the Fourier analysis for analytic functions.

Chapter 3: Decomposition of nonaffine fields – weighted empirical interpolation

In this chapter, we treat one of the crucial ingredients for achieving the main advantage of the reduced basis method - offline-online computational stages separation that critically relies on decomposition of nonaffine random fields. In the context of arbitrary probability measures, we propose a weighted empirical interpolation method by considering a weighted optimization problem. Moreover, a priori error estimates for the convergence property of the proposed method is obtained, which improves the result in [129]. To demonstrate numerically its effectiveness and efficiency, we apply the weighted empirical interpolation method in approximating nonlinear parametric functions, geometric Brownian motion in one dimension, exponential Karhunen–Loève expansion in multiple dimensions, as well as reduced basis approximation to nonaffine stochastic elliptic problems, and compare it with the conventional empirical interpolation method and the sparse grid stochastic collocation method.

Chapter 4: A hybrid and goal-oriented adaptive strategy for risk analysis

In this chapter, we address the question of how to accurately and efficiently apply the reduced basis method in the context of risk analysis, which is one important type of uncertainty quantification problems. We develop a hybrid and goal-oriented adaptive computational strategy based on the reduced basis method with accuracy certification. In dealing with high-dimensional random input problems, we propose and demonstrate that the reduced basis approximation space constructed by a goal-oriented greedy algorithm governed by an accurate and sharp a posteriori error bound for the output approximation error is quasi-optimal, resulting in low-dimensional approximation space when the stochastic solution and output live in a low-dimensional manifold. For an accurate evaluation of the failure probability when the limit state surface is not smooth, we design a hybrid computational approach with goal-oriented adaptation, which is proved to result in the same failure probability as Monte Carlo sampling for several models based on different types of partial differential equations.

Chapter 5: Breaking the curse of dimensionality – sparsity and reducibility

This chapter is devoted to the presentation of an adaptive and reduced computational framework in addressing the common computational challenge of curse of dimensionality faced in many uncertainty quantification problems. We take advantage of the sparsity, weak interaction and distinct importance of different dimensions, and develop a verified dimension adaptive algorithm based on hierarchical surpluses and generalized sparse grid construction. This algorithm can effectively detect the sparsity of the solution in stochastic space and successfully avoid the stagnation phenomenon that could be encountered in the adaptive approximation. In order to alleviate the heavy computational burden at many sampling points in high dimensions, we turn to the reducibility of the full system and use a reduced basis approximation according to the quantities of interest whenever new grid points are included in the hierarchical construction of the generalized sparse grid. This adaptive and reduced computational framework is then compared with other techniques, such as analysis of variance and anisotropic sparse grid, for a wide range of high-dimensional stochastic problems.

Chapter 6–8: Analyses and fast solvers for stochastic optimal control problems

The last three chapters are devoted to the analyses and the development of fast solvers for stochastic optimal control problems constrained by partial differential equations of different types. In the analysis of the well-posedness of the problems, including existence and uniqueness of the stochastic optimal solution, saddle point formulations are established and proved to be equivalent to the stochastic optimal control problems, as well as the Karush-Kuhn-Tucker optimality systems obtained by variational approach via Lagrange multipliers. The finite element method with quasi-optimal preconditioner

and the stochastic collocation method are commonly applied for the approximation of the solution in physical and stochastic spaces, respectively. In order to achieve fast solve of the stochastic optimality system, the weighted reduced basis method is employed by tailoring the saddle point problem as a weakly coercive elliptic problem. Stochastic regularity of the solution with respect to different random variables are studied in detail. Provided certain regularity on the input random data, analytical extension of the stochastic solution in a complex region is obtained for all the considered problems, which lead to explicit analysis of the global error analysis with contribution from finite element approximation error, stochastic collocation approximation error and reduced basis approximation error.

Chapter 6 deals with a stochastic optimal Robin boundary control problem constrained by an advection-dominated elliptic problem, where the advection field is prescribed as a random field under the assumption of finite dimensional noise assumption. Finite element approximation is adopted with SUPG (Streamline Upwind/Petrov–Galerkin) stabilization. In this chapter, we derive and demonstrate the error estimates for both physical and stochastic approximations.

Chapter 7 presents the results about a weighted reduced basis method for a stochastic optimal control problem with distributed control function constrained by an elliptic equation, where uncertainties are provided in the diffusion coefficient that may represent heat conductivity of heterogeneous material or permeability of porous media of groundwater flow. In particular, experiments are designed to test the efficiency and accuracy of the weighted reduced basis approximation.

Chapter 8 extends the methods developed in Chapter 7 to solve a stochastic optimal control problem constrained by Stokes equations, where we consider a distributed control function and that uncertainties are present in the viscosity field and the Neumann boundary conditions. Particular attention is paid to the construction of the reduced basis space due to the outer saddle point structure of the stochastic optimality system with an inner saddle point structure of the Stokes equations. A multilevel greedy algorithm associated with the construction of isotropic or anisotropic sparse grid is proposed, whose computational efficiency is illustrated by several numerical experiments.

The following two figures illustrate the organization of the thesis in two tracks.

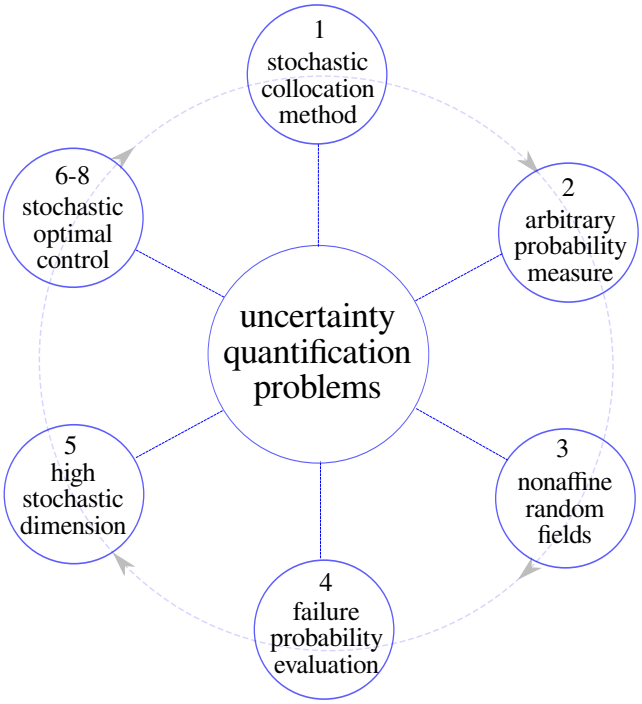

Figure 2: Thesis organization: schematic chart in the track of uncertainty quantification problems.

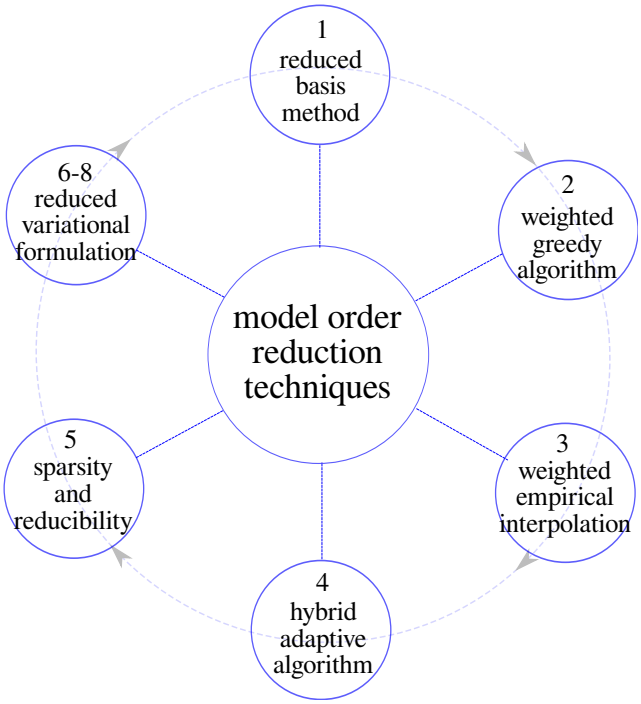

Figure 3: Thesis organization: schematic chart in the track of model order reduction techniques.

Preliminary

In this chapter we introduce some basic notations, function spaces, useful tools and the basic settings of uncertainty quantification problems which will be employed all along the rest of the thesis.

Basic notations and function spaces

Let D be an open and bounded physical domain in $\mathbb{R}^d$ ($d = 1,2,3$) with Lipschitz continuous boundary ∂D [165, 161]. Let $\bar{D} = D \cup \partial D$ be the closure of D; $x = (x_1, \dots, x_d) \in \bar{D}$ stands for the spatial coordinate. For $1 \le p \le \infty$, we consider the family of Banach spaces $L^p(D)$ which consist of the set of measurable functions (according to the Lebesgue measure) v defined in D, such that

$$\int_D |v(x)|^p dx < \infty, \quad 1 \le p < \infty, \tag{1}$$

or, when $p = \infty$,

$$\text{ess sup}_{x\in D}|v(x)| \equiv \{C \ge 0 \,|\, |v(x)| \le C \text{ almost everywhere in } D\} < \infty. \tag{2}$$

The associated norms of these spaces are given by

$$||v||_{L^p(D)} := \left(\int_D |v(x)|^p dx\right)^{1/p}, \quad 1 \le p < \infty, \tag{3}$$

or, when $p = \infty$,

$$||v||_{L^\infty(D)} = \text{ess sup}_{x\in D}|v(x)|. \tag{4}$$

When endowed with the scalar product

$$(w, v)_{L^2(D)} := \int_D w(x)v(x)dx \quad \forall w, v \in L^2(D), \tag{5}$$

the Banach space $L^2(D)$ becomes a Hilbert space. Indeed, it is a special case (when $s = 0$) of the Hilbert spaces in more general setting $H^s(D), s \ge 0$,

$$H^s(D) := \left\{v : D \to \mathbb{R} \,|\, v \text{ is measurable and } D^{\boldsymbol{\beta}} v \in L^2(D), |\boldsymbol{\beta}| \le s\right\}, \tag{6}$$

where the partial derivative is defined as $D^{\boldsymbol{\beta}} v := \partial^{|\boldsymbol{\beta}|} v / \partial x_1^{\beta_1} \cdots x_d^{\beta_d}$ for every nonnegative multi-index $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)$ and $|\boldsymbol{\beta}| = \beta_1 + \cdots + \beta_d$. The associated norms and seminorms are given by

$$||v||_{H^s(D)} := \left(\sum_{|\boldsymbol{\beta}|\le s} ||D^{\boldsymbol{\beta}} v||^s_{L^s(D)}\right)^{1/s} \text{ and } |v|_{H^s(D)} := \left(\sum_{|\boldsymbol{\beta}|= s} ||D^{\boldsymbol{\beta}} v||^s_{L^s(D)}\right)^{1/s}, \tag{7}$$

respectively. The endowed scalar product in $H^s(D)$ reads

$$(w,v)_{H^s(D)} = \sum_{|\boldsymbol{\beta}|\le s} \left(D^{\boldsymbol{\beta}} w, D^{\boldsymbol{\beta}} v\right)_{L^2(D)} \quad \forall\, w,v \in H^s(D). \tag{8}$$

Finally, we define the Hilbert space $H_0^1(D) := \{v \in H^1(D) : v = 0 \text{ on } \partial D\}$.

Let $(\Omega, \mathscr{F}, P)$ denote a complete probability space, where Ω is a set of outcomes $\omega \in \Omega$, $\mathscr{F}$ is a σ-algebra of events and $P : \mathscr{F} \to [0,1]$ with $P(\Omega) = 1$ is a probability measure [64]. A real-valued *random variable* is defined as a measurable function $Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$, being $\mathscr{B}$ the Borel σ-algebra on $\mathbb{R}$. The distribution function of a random variable $Y : \Omega \to \Gamma \subset \mathbb{R}$, being Γ the image of Y, is defined as $F_Y : \Gamma \to [0,1]$ such that $\forall\, y \in \Gamma, F_Y(y) = P(\omega \in \Omega : Y(\omega) \le y)$. Let $dF_Y(y)$ denote the distribution measure. Provided that $dF_Y(y)$ is absolutely continuous with respect to the Lebesgue measure dy, which we assume hereafter to be the case, there exists a probability density function $\rho : \Gamma \to \mathbb{R}$ such that $\rho(y)dy = dF_Y(y)$. For any positive integer $K \in \mathbb{N}_+$, we can define the real-valued K-dimensional random vector $Y = (Y_1, \ldots, Y_K) : (\Omega, \mathscr{F}) \to (\mathbb{R}^K, \mathscr{B}^K)$, where each of its element $Y_k, 1 \le k \le K$, is a random variable. The image of the random vector Y is denoted as $\Gamma = \Gamma_1 \otimes \cdots \otimes \Gamma_K \subset \mathbb{R}^K$, with $y = (y_1, \ldots, y_K) \in \Gamma$ representing its element. By the definition of the K-dimensional probability density function $\rho : \Gamma \to \mathbb{R}$, the probability measure is written as $\rho(y)dy \equiv \rho(y_1, \ldots, y_K)dy_1 \cdots dy_K$. Note that the new measure space $(\Gamma, \mathscr{B}(\Gamma), \rho(y)dy)$ is isometric to $(\Omega, \mathscr{F}, P)$ under the random variable/vector Y.

For any $1 \le p < \infty$ and any random variable v defined in the probability space $(\Omega, \mathscr{F}, P)$, we define the statistical p-th moment of v as

$$\mathbb{E}\left[v^p\right] := \int_\Omega v^p(\omega) dP(\omega), \tag{9}$$

and the associated stochastic Banach space as

$$L_P^p(\Omega) := \left\{v : \Omega \to \mathbb{R} \,|\, v \text{ is a random variable in } (\Omega, \mathscr{F}, P) \text{ such that } \mathbb{E}[|v|^p] < \infty\right\}, \tag{10}$$

whose norm is equipped as

$$||v||_{L_P^p(\Omega)} = \left(\int_\Omega |v(\omega)|^p dP(\omega)\right)^{1/p}. \tag{11}$$

When $p = 2$, $L_P^2(\Omega)$ is a stochastic Hilbert space with the scalar product defined as

$$(w,v)_{L_P^2(\Omega)} = \int_\Omega w(\omega) v(\omega) dP(\omega) \quad \forall\, w,v \in L_P^2(\Omega). \tag{12}$$

Suppose the random variable v is a function of the random variable/vector Y, then we have

$$\mathbb{E}\left[v^p\right] = \int_\Omega v^p(Y(\omega)) dP(\omega) \equiv \int_\Gamma v^p(y) dF_Y(y) \equiv \int_\Gamma v^p(y)\rho(y) dy, \tag{13}$$

so that $v \in L_\rho^p(\Gamma)$ and $||v||_{L_P^p(\Omega)} = ||v||_{L_\rho^p(\Gamma)}$, where

$$L_\rho^p(\Gamma) := \left\{w : \Gamma \to \mathbb{R} \,|\, w \text{ is a measurable function in } (\Gamma, \mathscr{B}(\Gamma), \rho(y)dy) \text{ such that } \mathbb{E}[|w|^p] < \infty\right\}, \tag{14}$$

and its norm is given by

$$||v||_{L_\rho^p(\Gamma)} = \left(\int_\Gamma |v(y)|^p \rho(y) dy\right)^{1/p}. \tag{15}$$

Correspondingly, when $p = 2$, $L_\rho^2(\Gamma)$ is a Hilbert space endowed with the scalar product

$$(w,v)_{L_\rho^2(\Gamma)} = \int_\Gamma w(y) v(y) \rho(y) dy \quad \forall\, w,v \in L_\rho^2(\Gamma). \tag{16}$$

For $p=\infty$, we define $L^\infty(\Gamma)$ similar to $L^\infty(D)$ with norm $||v||_{L^\infty(\Gamma)} = \text{ess sup}_{y\in\Gamma}|v(y)|$.

Let $v: D\times\Omega\to\mathbb{R}$ represent a real-valued *random field*, i.e., real-valued random variable defined in Ω for almost every $x\in D$, for which we define the following tensor-product Hilbert spaces:

$$L_P^p(\Omega)\otimes H^s(D) := \left\{v: D\times\Omega\to\mathbb{R}\,|\text{ for almost surely } w\in\Omega, v(\cdot,\omega)\in H^s(D) \text{ and } ||v||_{H^s(D)}\in L_P^p(\Omega)\right\}. \tag{17}$$

We denote $\mathscr{H}^s(D) = L_P^2(\Omega)\otimes H^s(D)$ for short and equip them with the following norms:

$$||v||_{\mathscr{H}^s(D)} := \left(\int_\Omega ||v(\cdot,\omega)||^2_{H^s(D)} dP(\omega)\right)^{1/2}. \tag{18}$$

When $s=0$, we employ the abbreviated notation $\mathscr{L}^2(D)$ alternative to $\mathscr{H}^0(D)$ by convention. Moreover, the scalar product in the tensor-product Hilbert spaces is defined as

$$(w,v)_{\mathscr{H}^s(D)} = \int_\Omega \sum_{|\boldsymbol{\beta}|\le s}\left(D^{\boldsymbol{\beta}}w, D^{\boldsymbol{\beta}}v\right)_{L^2(D)} dP(\omega) \quad \forall\, w,v\in\mathscr{H}^s(D). \tag{19}$$

When the random field v is a function of some random variable/vector Y at almost every $x\in D$, we may define the spaces $L_\rho^p(\Gamma)\otimes H^s(D)$ (which is isomorphic to the spaces $L_\rho^p(\Gamma; H^s(D))$), similar to $L_P^p(\Omega)\otimes H^s(D)$ and use the same shorthand notation $\mathscr{H}^s(D) = L_\rho^2(\Gamma)\otimes H^s(D)$ and $\mathscr{L}^2(D) = L_\rho^2(\Gamma)\otimes L^2(D)$.

The definitions can be generalized for a vector-valued *random field* $\boldsymbol{v} = (v_1,\dots,v_d): D\times\Omega\to\mathbb{R}^d$. We denote the Hilbert space $\mathscr{H}^{s,d}(D) = \left(L_P^2(\Omega)\otimes H^s(D)\right)^d$ (or $\mathscr{L}^{2,d}(D)$ for $s=0$), which is associated with the norm $||\boldsymbol{v}||_{\mathscr{H}^{s,d}(D)} = \sum_{i=1}^d ||v_i||_{\mathscr{H}^s(D)}$ and the scalar product $(\boldsymbol{v},\boldsymbol{w})_{\mathscr{H}^{s,d}(D)} = \sum_{i=1}^d (v_i,w_i)_{\mathscr{H}^s(D)}$.

When considering a spacial-temporal random field, or a stochastic process, $v:(0,T)\times D\times\Omega\to\mathbb{R}$, we may adopt the conventional manner [165, 161] for the definition of the space

$$L^q(0,T;\mathscr{H}^s(D)) := \left\{v:(0,T)\to\mathscr{H}^s(D)\,|\, v \text{ is measurable and } \int_0^T ||v(t)||^q_{\mathscr{H}^s(D)} dt<\infty\right\}, \tag{20}$$

$1\le q<\infty$, endowed with the norm

$$||v||_{L^q(0,T;\mathscr{H}^s(D))} := \left(\int_0^T ||v(t)||^q_{\mathscr{H}^s(D)} dt<\infty\right)^{1/q}. \tag{21}$$

Another most often used space for spatial-temporal random field, $C^0(0,T;\mathscr{H}^s(D))$, consists of $\mathscr{H}^s(D)$-valued continuous functions in $(0,T)$, endowed with the norm

$$||v||_{C^0(0,T;\mathscr{H}^s(D))} = \sup_{t\in(0,T)} ||v||_{\mathscr{H}^s(D)}. \tag{22}$$

Karhunen–Loève expansion

Random input data or uncertainties of a given physical system can be mathematically formulated as spatial, temporal or spatial-temporal random fields in general. Efficient and accurate representation and approximation of the random fields play a critical role in facilitating the development of advanced stochastic computational methods that feature faster convergence, higher accuracy and cheaper computational cost than the Monte Carlo method. By Parseval's identity formula in stochastic Hilbert space [124], any square-integrable random field can be represented by a (possibly infinite) linear combinations of orthogonal functions, in particular by any orthonormal basis of the stochastic Hilbert space

where the random field is defined. One special orthonormal basis is known as *Karhunen-Loève (KL) expansion* [108, 124], and can also be found in different contexts as *Proper Orthogonal Decomposition (POD)* or *Principle Component Analysis (PCA)*. A general presentation of the KL expansion is provided as follows.

Suppose $v : D \times \Omega \to \mathbb{R}$ is a square-integrable spatial random field, i.e. $v \in \mathscr{L}^2(D)$. Suppose also that the random field v has a continuous and bounded covariance function defined as

$$\mathbb{C}[v](x,x') := \mathbb{E}[(v(x,\cdot) - \mathbb{E}[v](x)))(v(x',\cdot) - \mathbb{E}[v](x')))], \quad \forall x, x' \in D, \tag{23}$$

where the expectation function $\mathbb{E}[v]$ of the random field v is given by

$$\mathbb{E}[v](x) := \int_\Omega v(x,\omega) dP(\omega), \quad \forall x \in D. \tag{24}$$

Let us define an integral operator T_v associated with the covariance function $\mathbb{C}$ as

$$T_v[w](x) := \int_D \mathbb{C}[v](x,x') w(x') dx'. \tag{25}$$

Then T_v is compact, positive and self-adjoint [124]. By spectral theorem [170], we have that T_v possesses the eigenpairs $(\lambda_k, v_k)_{k=1}^\infty$, i.e.

$$T_v[v_k] = \lambda_k v_k, \tag{26}$$

where the eigenfunctions $v_k, k = 1,2,\ldots$, form an orthonormal basis of $L^2(D)$, and the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Provided that the covariance function $\mathbb{C}$ is smooth, the eigenvalues decay exponentially fast to zero [189]. Moreover, by Mercer's theorem [170], for almost every (a.e.) $\omega \in \Omega$, the following Karhunen–Loève expansion holds

$$v(x,\omega) = \mathbb{E}[v](x) + \sum_{k=1}^\infty \sqrt{\lambda_k} v_k(x) Y_k(\omega), \quad \forall x \in D, \tag{27}$$

where $Y_k, k = 1,2,\ldots$, are uncorrelated random variables with zero mean and unit variance, defined by

$$Y_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D (v(x,\omega) - \mathbb{E}[v](x))\, v_k(x) dx. \tag{28}$$

We truncate the Karhunen–Loève expansion of the random field v with K terms as

$$v_K(x,\omega) := \mathbb{E}[v](x) + \sum_{k=1}^K \sqrt{\lambda_k} v_k(x) Y_k(\omega), \quad \forall x \in D, \tag{29}$$

then v_K represents the best K-term approximation of the random field v in $\mathscr{L}^2(D)$ and the associated approximation/truncation error reads

$$||v - v_K||_{\mathscr{L}^2(D)} = \left(\sum_{k=K+1}^\infty \lambda_k \right)^{1/2}. \tag{30}$$

The superiority of the KL expansion compared to other spectral expansion is that it provides the optimal orthonormal basis in the sense that the best K-term approximation is achieved when considering the total mean squared error (30).

Throughout the thesis, we make the following basic assumption for a random field of interest.

Assumption 0.1 *The random field $v : D \times \Omega \to \mathbb{R}$ depends on K-dimensional ($K = 1,2,\ldots,$) independent random variables $Y := (Y_1,\ldots,Y_K) : \Omega \to \mathbb{R}^K$ with the image $\Gamma := \prod_{k=1}^K \Gamma_k \subset \mathbb{R}^K$ and the joint probability*

density function $\rho := \prod_{k=1}^{K} \rho_k : \Gamma \to \mathbb{R}$. *Moreover, for the sake of simplicity, we assume*

$$v(x, Y(\omega)) = v_0(x) + \sum_{k=1}^{K} v_k(x) Y_k(\omega) \Longrightarrow v(x, y) = v_0(x) + \sum_{k=1}^{K} v_k(x) y_k, \tag{31}$$

where we can identify $v_0 = \mathbb{E}[v]$ *and* $v_k = \sqrt{\lambda_k} v_k$ *or* $y_k = \sqrt{\lambda_k} y_k$, $k = 1, \ldots, K$ *for a random field with the truncated Karhunen–Loève expansion (29).*

This assumption is not necessarily satisfied in many practical applications. For instance, a slightly more general representation of the random field is given by the affine expansion

$$v(x, Y(\omega)) = v_0(x) + \sum_{k=1}^{K} v_k(x) \Theta_k(Y(\omega)) \Longrightarrow v(x, y) = v_0(x) + \sum_{k=1}^{K} v_k(x) \Theta_k(y), \tag{32}$$

where $\Theta_k, 1 \le k \le K$, are functions of the random vector Y. The affine expansion of the random field is a crucial condition for efficient offline-online computational decomposition of the reduced basis method. However, in more general situations, one may face nonlinear, nonaffine random field, such as

$$v(x, Y(\omega)) = v_0(x) + \exp\left(\sum_{k=1}^{K} v_k(x) Y_k(\omega) \right) \Longrightarrow v(x, y) = v_0(x) + \exp\left(\sum_{k=1}^{K} v_k(x) y_k \right), \tag{33}$$

which is widely used for approximating positive random field. In this circumstance, suitable affine decomposition techniques may be employed in order to achieve the computational efficiency of the reduced basis method, e.g. *empirical interpolation method* as introduced in chapter 3.

We remark that for a square-integrable vector-valued random field $\boldsymbol{v} = (v_1, \ldots, v_d) : D \times \Omega \to \mathbb{R}^d$, i.e. $\boldsymbol{v} \in \mathscr{L}^{2,d}(D)$, the Karhunen-Loève expansion is defined similar to the scalar case as

$$\boldsymbol{v}(x, \omega) = \mathbb{E}[\boldsymbol{v}](x) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \boldsymbol{v}_k(x) Y_k(\omega), \tag{34}$$

where the random variables $Y_k, k = 1, 2, \ldots$, are given by

$$Y_k(\omega) = \sum_{j=1}^{d} \int_D (v_j(x, \omega) - \mathbb{E}[v_j](x)) (\boldsymbol{v}_k)_j(x) dx, \tag{35}$$

and $(\lambda_k, \boldsymbol{v}_k)_{k=1}^{\infty}$ are the eigenpairs of the integral operator $\boldsymbol{T}_{\boldsymbol{v}}$, which is defined as

$$(\boldsymbol{T}_{\boldsymbol{v}}[\boldsymbol{w}](x))_i := \sum_{j=1}^{d} \int_D \mathbb{C}^{i,j}[\boldsymbol{v}](x, x') w_j(x') dx', \quad 1 \le i \le d. \tag{36}$$

Here, the covariance matrix function $\mathbb{C}$ is given by

$$\mathbb{C}^{i,j}[\boldsymbol{v}](x, x') = \mathbb{E}[(v_i(x, \cdot) - \mathbb{E}[v_i](x))(v_j(x', \cdot) - \mathbb{E}[v_j](x'))], \quad 1 \le i, j \le d. \tag{37}$$

Similar to Assumption 0.1, we also assume finite dimensional noise for a vector-valued random field.

Stochastic partial differential equations

The solution of uncertainty quantification problems depends on the underlying mathematical models, which are usually formulated as partial differential equations with random inputs: these are known as stochastic partial differential equations (PDEs). We remark that stochastic PDEs are conventionally

referred to those PDEs with the random inputs given by Brownian motion or Wiener process [202]. In this thesis, we adopt a broader meaning of stochastic PDEs and specifically deal with the PDEs with random inputs under Assumption 0.1 of finite dimensional noise.

A general formulation of a stochastic PDE reads: find $u : (0,T) \times D \times \Omega \to \mathbb{R}$, such that

$$\begin{cases} \mathfrak{L}(u(t,x,\omega);\omega) &= f(t,x,\omega) \quad \forall (t,x,\omega) \in (0,T) \times D \times \Omega, \\ \mathfrak{B}(u(t,x,\omega);\omega) &= g(t,x,\omega) \quad \forall (t,x,\omega) \in (0,T) \times \partial D \times \Omega, \\ u(0,x,\omega) &= h(x,\omega) \quad \forall (x,\omega) \in D \times \Omega, \end{cases} \tag{38}$$

where u is the stochastic solution; f, g, h are the random fields representing forcing term, boundary and initial conditions, respectively; $\mathfrak{L}$ and $\mathfrak{B}$ are the stochastic differential operators defined in the entire domain D and on the boundary ∂D, respectively. A special case of a linear and coercive stochastic elliptic PDE has been widely considered as the benchmark model for the development of stochastic computational methods to solve more general stochastic problems formulated as PDEs with random inputs [206, 80, 10, 138, 76, 207, 8, 151, 50, 49]. For demonstration and illustration of the efficiency and the accuracy of novel stochastic methods and algorithms, we also employ this model throughout the thesis besides considering more general problems as (38).

A linear stochastic elliptic problem is formulated as: find $u : D \times \Omega \to \mathbb{R}$ such that it holds almost surely

$$\begin{aligned} -\nabla \cdot (a(x,\omega)\nabla u(x,\omega)) = f(x,\omega) \quad \forall (x,\omega) \in D \times \Omega, \\ u(x,\omega) = 0 \quad \forall (x,\omega) \in \partial D \times \Omega, \end{aligned} \tag{39}$$

where the divergence $\nabla\cdot$ and the gradient ∇ are taken with respect to x; the homogeneous Dirichlet boundary condition is prescribed on the whole boundary ∂D for simplicity. For the random forcing term f and the coefficient field a, we consider the following assumptions.

Assumption 0.2 *The random forcing term f is square integrable, i.e.*

$$||f||_{\mathscr{L}^2(D)} = \left(\int_\Omega \int_D |f(x,\omega)|^2 dx dP(\omega) \right)^{1/2} < \infty. \tag{40}$$

The random coefficient a is assumed to be uniformly bounded from below and from above, i.e., there exist two constants a_{min} and a_{max} with $0 < a_{min} < a_{max} < \infty$ such that

$$P(\omega \in \Omega : a_{min} < a(x,\omega) < a_{max} \quad \forall x \in \bar{D}) = 1. \tag{41}$$

Assumption 0.3 *Inherited from Assumption 0.1, f depends on finite dimensional noise, i.e.,*

$$f(x,\omega) = f(x,Y(\omega)) = f_0(x) + \sum_{k=1}^{K} f_k(x) Y_k(\omega) \Longrightarrow f(x,y) = f_0(x) + \sum_{k=1}^{K} f_k(x) y_k, \tag{42}$$

where $Y_1, \ldots, Y_K$ are independent random variables; $f_k \in L^2(D), 1 \le k \le K$. Similar finite dimensional noise assumption is made for the random coefficient field a, i.e.,

$$a(x,\omega) = a(x,Y(\omega)) = a_0(x) + \sum_{k=1}^{K} a_k(x) Y_k(\omega) \Longrightarrow a(x,y) = a_0(x) + \sum_{k=1}^{K} a_k(x) y_k, \tag{43}$$

where the leading term is assumed to be dominating and uniformly bounded away from 0, i.e.,

$$\exists \delta > 0, a_{\min} \text{ the same as in (41) s.t. } a_0(x) \ge \delta \; \forall x \in D, \text{ and } ||a_k||_{L^\infty(D)} < 2a_{\min}, 1 \le k \le K. \tag{44}$$

Remark 0.0.1 *When the random variables $Y_k^a, 1 \le k \le K_a$, for a and $Y_k^f, 1 \le k \le K_f$, for f are not the same, we collect them as $Y = (Y_1^a, \ldots, Y_{K_a}^a, Y_1^f, \ldots, Y_{K_f}^f)$ and reorder them as $(Y_1, \ldots, Y_K)$ with $K = K_a + K_f$.*

Under the above assumptions, the stochastic elliptic problem is transformed to a parametric elliptic problem, which reads: find $u : D \times \Gamma \to \mathbb{R}$ such that the following equations hold:

$$\begin{aligned} -\nabla \cdot (a(x,y)\nabla u(x,y)) = f(x,y) \quad \forall (x,y) \in D \times \Gamma, \\ u(x,y) = 0 \quad \forall (x,y) \in \partial D \times \Gamma. \end{aligned} \tag{45}$$

For the solution of the elliptic problem (45), we provide two weak formulations facilitating different stochastic computational methods. The first is called D-weak/Γ-strong formulation, or semi-weak formulation, which reads: $\forall\, y \in \Gamma$, find $u(y) \in H_0^1(D)$ such that

$$A(u, v; y) = F(v; y) \quad \forall\, v \in H_0^1(D), \tag{46}$$

where $A(\cdot, \cdot; y)$ and $F(\cdot; y)$ are parametrized bilinear and linear forms written as

$$A(u, v; y) = A_0(u, v) + \sum_{k=1}^{K} A_k(u, v) y_k \quad \text{and} \quad F(v; y) = (f_0, v) + \sum_{k=1}^{K} (f_k, v) y_k, \tag{47}$$

with the deterministic bilinear forms $A_k(u, v)$ given by $A_k(u, v) := (a_k \nabla u, \nabla v), k = 0, 1, \ldots, K$.

The D-weak/Γ-strong formulation (46) is suitable for nonintrusive methods, i.e. the methods that can directly use the deterministic solver of the underlying PDE at any given sample $y \in \Gamma$, such as the stochastic collocation method. As for the application of the stochastic spectral Galerkin projection method, where the solution is projected onto the basis in both physical space and stochastic space, we use the D-weak/Γ-weak formulation: find $u \in \mathscr{H}_0^1(D)$ such that

$$\mathscr{A}(u, v) = \mathscr{F}(v) \quad \forall\, v \in \mathscr{H}_0^1(D), \tag{48}$$

where the bilinear and linear forms are given by

$$\mathscr{A}(u, v) = \int_\Gamma \int_D a_0(x) \nabla u \cdot \nabla v \rho(y) dx dy + \sum_{k=1}^{K} \int_\Gamma \int_D a_k(x) y_k \nabla u \cdot \nabla v \rho(y) dx dy, \tag{49}$$

and

$$\mathscr{F}(v) = \int_\Gamma \int_D f_0(x) v \rho(y) dx dy + \sum_{k=1}^{K} \int_\Gamma \int_D f_k(x) y_k v \rho(y) dx dy. \tag{50}$$

To study the well-posedness of the semi-weak solution of problem (46) and the weak solution of problem (49), we need the following theorem [68].

Theorem 0.0.1 *(Lax-Milgram Theorem). Assume that X is a real Hilbert space, with norm $||\cdot||_X$ and inner product $(\cdot, \cdot)$. Let $< \cdot, \cdot >$ denote the pairing of X with its dual space. Suppose the bilinear mapping $B : X \times X \to \mathbb{R}$ satisfies the conditions*

$$|B(u, v)| \le \gamma ||u||_X ||v||_X \quad \forall\, u, v \in X, \tag{51}$$

and

$$\alpha ||u||_X^2 \le B(u, u) \quad \forall\, u \in X, \tag{52}$$

for some positive constants α, γ. Suppose also that $f : X \to \mathbb{R}$ is a bounded linear functional on X. Then

there exists a unique element $u \in X$ such that

$$B(u,v) =< f, v > \quad \forall v \in X. \tag{53}$$

Because of Assumption 0.2, the bilinear and linear forms in (47) satisfy the conditions (51) and (52) in $H_0^1(D)$, and thus the existence of a unique parametric solution $u(y) \in H_0^1(D)\ \forall y \in \Gamma$ to problem (46) is guaranteed by the Lax–Milgram theorem. Moreover, we have the a priori estimate for the solution

$$||u(y)||_{H_0^1(D)} \leq \frac{C_P}{a_{min}(y)} ||f(y)||_{L^2(D)} \quad \forall y \in \Gamma, \tag{54}$$

where $a_{min}(y) = \min_{x \in D} a(x,y)$ and C_P is the constant of the Poincaré inequality $||v||_{L^2(D)} \leq C_P ||\nabla v||_{L^2(D)}$, $\forall v \in H_0^1(D)$. Existence and uniqueness of the solution of problem (49) is also guaranteed by this theorem in the stochastic Hilbert space $\mathscr{H}_0^1(D)$, and the following a priori estimate holds

$$||u||_{\mathscr{H}_0^1(D)} \leq \frac{C_P}{a_{min}} ||f||_{\mathscr{L}^2(D)}, \tag{55}$$

where a_{min} is defined in Assumption 0.1; C_P satisfies $||v||_{\mathscr{L}^2(D)} \leq C_P ||\nabla v||_{\mathscr{L}^2(D)}$, $\forall v \in \mathscr{H}_0^1(D)$.

More often, we are not interested in the solution u itself but on a functional $s(u;y)$ of the solution as model output, e.g., the compliant outout $s(u;y) = F(u;y)$, as well as its statistics, e.g., the expectation

$$\mathbb{E}[s] = \int_\Gamma s(u;y)\rho(y)\,dy. \tag{56}$$

This chapter provides a common background for the thesis. More general stochastic PDEs such as parabolic equations, Stokes equations, and those involving noncompliant outputs, nonaffine random inputs as well as specific assumptions, definitions and notations will be introduced when in need.

Part I

Forward Uncertainty Quantification Problems: Challenges and Solutions

1 Comparison of stochastic collocation and reduced basis methods

Among many stochastic computational methods to solve uncertainty quantification problems, the stochastic collocation (SC) method based on sparse grid techniques [33, 207, 8, 149] is one of the most popular and widely used methods. It features fast convergence properties comparable to the stochastic Galerkin method and simple implementation typical of the Monte Carlo method. However, this method is computationally prohibitive for high-dimensional problems because of the heavy computational cost involved by the solution of the underlying model at one random realization. This constraint has prompted the development of various model order reduction techniques, in particular the reduced basis (RB) method [130, 178, 158].

In this first chapter, our target is the comparison of the stochastic collocation method and the reduced basis method focusing on a rather simple benchmark, a stochastic elliptic problem. Two important comparison criteria are considered: 1), convergence results of the approximation error; 2), computational costs for both offline construction and online evaluation. Numerical experiments are performed for test problems with stochastic dimensions (the number of independent random variables) from low (in magnitude $O(1)$) to moderate ($O(10)$), and to high ($O(100)$). The main result stemming from our comparison is that the reduced basis method converges better in theory and faster in practice than the stochastic collocation method for smooth problems, and is more suitable for large scale and high-dimensional stochastic problems when considering computational costs.

This chapter is organized as follows. In section 1.1, we set up the benchmark model and provide a regularity result for the comparison of the two methods. The general formulation for the stochastic collocation method and the reduced basis method are introduced in section 1.3 and 1.2, respectively. A theoretical comparison of convergence results in both univariate case and multivariate case, and a direct comparison of the approximation error are carried out in section 1.4. A detailed comparison of the computational costs for the two methods is provided by evaluating the cost of each step of the algorithms in section 1.5. In section 1.6, we perform a series of numerical experiments to assess the convergence rates and computational costs of the two methods. Finally, remarks about the possible limits of this comparison and some extensions to more general stochastic problems are given in the last section 1.7.

Reference for this chapter:

P. Chen, A. Quarteroni, and G. Rozza. Comparison between reduced basis and stochastic collocation methods for elliptic problems. Journal of Scientific Computing, 59:187–216, 2014.

1.1 Benchmark model

The linear and coercive stochastic elliptic PDE introduced in the preliminary chapter is adopted as our benchmark model. For simplicity, only the diffusion coefficient is considered as the source of uncertainties. Since any approach for the approximation of the solution in the stochastic space depends on the regularity of the solution with respect to the random vector $y \in \Gamma$, we summarize briefly the regularity results in Lemma 1.1.1 following [54] for infinite dimensional problems ($K = \infty$).

Lemma 1.1.1 *The following estimate for the solution of the problem (46) holds*

$$||\partial_y^\nu u||_{L^\infty(\Gamma;X)} \leq B|\nu|!b^\nu, \tag{1.1}$$

where $\nu = (\nu_1, \ldots, \nu_K) \in \mathbb{N}^K$, $|\nu| = \nu_1 + \cdots + \nu_K$, $H_0^1(D) \subset X \subset H^1(D)$, $B = ||u||_{L^\infty(\Gamma;X)}$, *and*

$$b_k = \frac{||a_k||_{L^\infty(D)}}{a_{min}} \text{ and } b^\nu = \prod_{k=1}^{K} b_k^{\nu_k}. \tag{1.2}$$

Furthermore, Lemma 1.1.1 implies thanks to the Taylor expansion the following regularity result which represents a generalization of the result in [13] from $\mathbb{R}^K$ to $\mathbb{C}^K$.

Corollary 1.1.2 *The solution* $u : \Gamma \to X$ *is analytic and can be analytically extended to the set*

$$\Sigma = \left\{ z \in \mathbb{C}^K : \sum_{k=1}^{K} |z_k - y_k| b_k < 1 \; \forall \, y \in \Gamma \right\}. \tag{1.3}$$

We may also write for $\tau_k \leq 1/(Kb_k), 1 \leq k \leq K$,

$$\Sigma_\tau = \left\{ z \in \mathbb{C}^K : dist(z_k, \Gamma_k) \leq \tau_k \, \forall 1 \leq k \leq K \right\}. \tag{1.4}$$

Remark 1.1.1 *The comparison of the two methods depends essentially on the following factors: the regularity of the stochastic solution in the stochastic space, the dimension of the stochastic space, and the complexity involved by the solution of a deterministic system at one stochastic realization. The conclusions drawn from the comparison results for the linear elliptic problem at hand are supposed to hold similarly for more general problems as long as the above factors are concerned.*

1.2 Stochastic collocation method

Given any realization $y \in \Gamma$, the stochastic collocation method [8] essentially adopts the Lagrange interpolation to approximate the solution $u(y)$ based on a set of deterministic solutions at the collocation points chosen according to the probability distribution function of the random variables. Therefore, we have to solve one deterministic problem at each collocation point. In order to achieve accurate and inexpensive collocation approximation of the stochastic solution as well as its statistics, it is important to select efficient collocation points. Let us introduce the univariate stochastic collocation at first.

1.2.1 Univariate interpolation

Given the collocation points in Γ, e.g., $y^0 < y^1 < y^2 < \cdots < y^N$, as well as the corresponding solutions $u(y^n), 0 \le n \le N$, we define the univariate N-th order Lagrange interpolation operator as

$$\mathscr{U}_N u(y) = \sum_{n=0}^{N} u(y^n) l^n(y), \tag{1.5}$$

where $l^n(y), 0 \le n \le N$, are the Lagrange characteristic polynomials of order N given in the form

$$l^n(y) = \prod_{m \neq n} \frac{y - y^m}{y^n - y^m}, \quad 0 \le n \le N. \tag{1.6}$$

One evaluation of $\mathscr{U}_N u(y)$ at a new realization $y \in \Gamma$ requires $O(N^2)$ operations by formula (1.5). In order to obtain efficient and stable polynomial interpolation, we use barycentric formula [164] and rewrite the characteristic polynomials as

$$l^n(y) = \underbrace{\frac{1}{\prod_{m \neq n}(y^n - y^m)}}_{\bar{w}^n} \cdot \frac{1}{y - y^n} \underbrace{\prod_{m=0}^{N} (y - y^m)}_{l(y)} = l(y) \frac{\bar{w}^n}{y - y^n}, \quad 0 \le n \le N, \tag{1.7}$$

where $\bar{w}^n, 0 \le n \le N$, are barycentric weights, so that the interpolation operator (1.5) becomes

$$\mathscr{U}_N u(y) = \sum_{n=0}^{N} \frac{\bar{w}^n}{y - y^n} u(y^n) \Big/ \sum_{n=0}^{N} \frac{\bar{w}^n}{y - y^n}, \text{ where } l(y) = \sum_{n=0}^{N} \frac{\bar{w}^n}{y - y^n}, \tag{1.8}$$

which instead needs only $O(N)$ operations for one evaluation provided that the barycentric weights are precomputed and stored. The expectation of the solution can therefore be approximated by

$$\mathbb{E}[u] \approx \mathbb{E}[\mathscr{U}_N u] = \sum_{n=0}^{N} \left(\int_\Gamma \left(\frac{\bar{w}^n}{y - y^n} \Big/ \sum_{n=0}^{N} \frac{\bar{w}^n}{y - y^n} \right) \rho(y) dy \right) u(y^n) = \sum_{n=0}^{N} w^n u(y^n), \tag{1.9}$$

where $w^n, 0 \le n \le N$, are quadrature weights. In order to improve the accuracy of the numerical integral in (1.9) and the numerical interpolation in (1.8), it is more suitable to select the collocation points as the quadrature abscissas. Available quadrature rules include *Clenshaw–Curtis* quadrature, Gaussian quadrature based on various orthogonal polynomials; see [164] for details.

1.2.2 Multivariate tensor product interpolation

Let us rewrite the univariate interpolation formula (1.5) with the index k for the k-th dimension as

$$\mathscr{U}_{N_k} u(y_k) = \sum_{y_k^{n_k} \in \Theta^k} u(y_k^{n_k}) l_k^{n_k}(y_k), \text{ where } \Theta^k = \{y_k^{n_k} \in \Gamma_k, n_k = 0, \ldots, N_k\} \text{ for some } N_k \ge 1; \tag{1.10}$$

then the multivariate interpolation is given as the tensor product of the univariate interpolation

$$\left(\mathscr{U}_{N_1} \otimes \cdots \otimes \mathscr{U}_{N_K}\right) u(y) = \sum_{y_1^{n_1} \in \Theta^1} \cdots \sum_{y_K^{n_K} \in \Theta^K} u(y_1^{n_1}, \ldots, y_K^{n_K}) \left(l_1^{n_1}(y_1) \otimes \cdots \otimes l_K^{n_K}(y_K)\right). \tag{1.11}$$

The corresponding barycentric formula for the multivariate interpolation is given as

$$\left(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_K}\right)u(y)=\sum_{y_1^{n_1}\in\Theta^1}\frac{b_{n_1}^1(y_1)}{\sum\limits_{y_1^{n_1}\in\Theta^1}b_{n_1}^1(y_1)}\cdots\sum_{y_K^{n_K}\in\Theta^K}\frac{b_{n_K}^K(y_K)}{\sum\limits_{y_K^{n_K}\in\Theta^K}b_{n_K}^K(y_K)}u(y_1^{n_1},\ldots,y_K^{n_K}),\tag{1.12}$$

where $b_{n_k}^k(y_k)=\bar{w}_{n_k}^k/(y_k-y_k^{n_k})$ with barycentric weights $\bar{w}_{n_k}^k, 1\le k\le K$, precomputed and stored. The multivariate barycentric formula reduces the tensor product interpolation from $O(N_1^2\times\cdots\times N_K^2)$ operations by (1.10) to $O(N_1\times\cdots\times N_K)$ operations by (1.12). Corresponding to the univariate interpolation, the expectation of the solution by the multivariate interpolation is given as

$$\mathbb{E}[u]\approx\mathbb{E}[\left(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_K}\right)u]=\sum_{y_1^{n_1}\in\Theta^1}\cdots\sum_{y_K^{n_K}\in\Theta^K}u(y_1^{n_1},\ldots,y_K^{n_K})\left(w_1^{n_1}\times\cdots\times w_K^{n_K}\right),\tag{1.13}$$

where the quadrature weights $w_k^{n_k}, 1\le k\le K$, can be precomputed and stored by

$$w_k^{n_k}=\int_{\Gamma_k}\left(b_{n_k}^k(y^k)\Big/\sum_{y_k^{n_k}\in\Theta^k}b_{n_k}^k(y^k)\right)\rho(y_k)dy_k.\tag{1.14}$$

We remark that the number of the collocation points or quadrature abscissas grows exponentially fast as $(N_1+1)\times\cdots\times(N_K+1)$, or $(N_1+1)^K$ if $N_1=\cdots=N_K$, which prohibits the application of the multivariate tensor product interpolation for high-dimensional stochastic problems, i.e., when K becomes large.

1.2.3 Sparse grid interpolation

In order to alleviate the *curse of dimensionality* in the interpolation on the full tensor product grid, various sparse grid techniques [33] have been developed, among which the *Smolyak* type [149] is one of the most popular constructions. For isotropic interpolation with the same degree $q\ge K$ for one-dimensional polynomial space in each direction, we have the *Smolyak* interpolation operator

$$\mathscr{S}_q u(y)=\sum_{q-K+1\le|i|\le q}(-1)^{q-|i|}\begin{pmatrix}K-1\\ q-|i|\end{pmatrix}\left(\mathscr{U}^{i_1}\otimes\cdots\otimes\mathscr{U}^{i_K}\right)u(y),\tag{1.15}$$

where $|i|=i_1+\cdots+i_K$ with the multivariate index $i=(i_1,\ldots,i_K)$ defined via the index set

$$X(q,K):=\left\{i\in N_+^K:\forall\, i_k\ge 1,\sum_{k=1}^K i_k\le q\right\},\tag{1.16}$$

and the set of collocation nodes for the sparse grid (see the middle of Figure 1.1) is thus collected as

$$H(q,K)=\bigcup_{q-K+1\le|i|\le q}\left(\Theta^{i_1}\times\cdots\times\Theta^{i_K}\right),\tag{1.17}$$

where the number of collocation nodes $\#\Theta^{i_k}=1$ if $i_k=1$, and $\#\Theta^{i_k}=2^{i_k-1}+1$ when $i_k>1$ in a nested structure. Note that we denote $\mathscr{U}^{i_k}\equiv\mathscr{U}_{N_k}$ defined in (1.10) for $N_k=2^{i_k-1}$. Define the differential operator $\Delta^{i_k}=\mathscr{U}^{i_k}-\mathscr{U}^{i_k-1}, k=1,\ldots,K$, with $\mathscr{U}^0=0$, we have an equivalent expression of the Smolyak

interpolation [8]

$$\begin{aligned}\mathscr{S}_q u(y) &= \sum_{i\in X(q,K)} \left(\Delta^{i_1}\otimes\cdots\otimes\Delta^{i_K}\right) u(y)\\ &= \mathscr{S}_{q-1}u(y) + \sum_{|i|=q} \left(\Delta^{i_1}\otimes\cdots\otimes\Delta^{i_K}\right) u(y).\end{aligned} \tag{1.18}$$

The above formula allows us to discretize the stochastic space in hierarchical structure based on nested collocation nodes, such as the extrema of Chebyshev polynomials or Gauss-Patterson nodes, leading to Clenshaw–Curtis cubature rule or Gauss-Patterson cubature rule, respectively [149, 110].

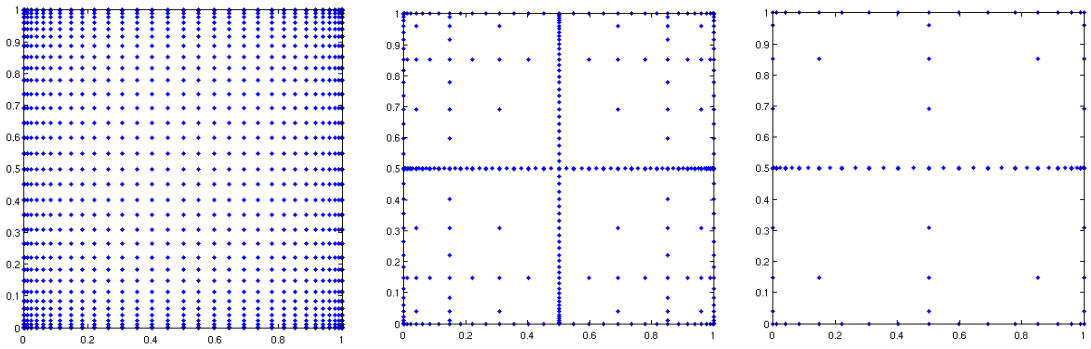

Figure 1.1: Two dimensional collocation nodes by Clenshaw–Curtis cubature rule in tensor product grid $q = 8$ (Left), sparse grid $q = 8$ (Middle), anisotropic sparse grid $q = 8$ and $\alpha = (1, 1.5)$(Right)

The Smolyak sparse grid [207] is originally developed as isotropic in every one-dimensional polynomial space. The convergence rate of the solution in each polynomial space may vary due to different importance of each random variable, which helps to reduce further the computational effort by interpolation based on anisotropic sparse grid [148] given by

$$\mathscr{S}_q^\alpha u(y) = \sum_{i\in X_\alpha(q,K)} \left(\Delta^{i_1}\otimes\cdots\otimes\Delta^{i_K}\right) u(y), \tag{1.19}$$

where the weighted index $X_\alpha(q,K)$ is defined as

$$X_\alpha(q,K) := \left\{ i\in N_+^K, i\geq 1 : \sum_{k=1}^K i_k\alpha_k \leq \min(\alpha)q \right\}. \tag{1.20}$$

Here, $\alpha = (\alpha_1,\ldots,\alpha_K)$ represents the weights in different dimensions, provided by either a priori or a posteriori error estimate; see [148]. Figure 1.1 displays the full tensor product grid, the sparse grid and the anisotropic sparse grid based on Clenshaw–Curtis cubature rule. We can observe that the isotropic and anisotropic sparse grids are much coarser than the full tensor product grid, leading to considerable reduction of the stochastic computation without much loss of accuracy, as we shall see in the convergence analysis and the numerical experiments in the following sections.

For certain specific problems, some other advanced techniques turn out to be more efficient than both the isotropic and the anisotropic Smolyak sparse grid techniques. For example, the quasi-optimal sparse grid [13] is assembled in a greedy manner to deal with the "accuracy-work" trade-off problem; the adaptive hierarchical sparse grid [127, 74] is referred to constructing the sparse grid adaptively in hierarchical levels with local refinement or domain decomposition in stochastic space, which is more suitable for low regularity problems; the combination of analysis of variance (ANOVA) and sparse grid

techniques [88, 100] to deal with high-dimensional problems, which will be studied in chapter 5.

1.3 Reduced basis method

Different from the interpolation approach used by the stochastic collocation method, the reduced basis method employs Galerkin projection in the reduced basis space spanned by a set of deterministic solutions [158, 178, 161]. Given any space X of dimension $\mathcal{N}$ for the approximation of the solution of problem (46) (for instance, finite element space), we hierarchically build the N dimensional reduced basis space X_N for $N = 1,\dots,N_{max}$, until satisfying tolerance requirement at $N_{max} \ll \mathcal{N}$, as

$$X_N = \text{span}\{u(y^n), n = 1,\dots,N\} \tag{1.21}$$

based on suitably chosen samples $S_N = \{y^1,\dots,y^N\}$ from a training set $\Xi_{train} \subset \Gamma$ (N is small under smoothness hypothesis [158]). The solutions $\{u(y^n), n = 1,\dots,N\}$ are called "snapshots" corresponding to the samples $\{y^n, n = 1,\dots,N\}$. Note that $X_1 \subset X_2 \subset \cdots \subset X_{N_{max}}$. Given any realization $y \in \Gamma$, we seek the solution $u_N(y)$ in the reduced basis space X_N by solving the following Galerkin projection problem

$$A(u_N, v; y) = F(v) \quad \forall\, v \in X_N. \tag{1.22}$$

Note we assume that f is deterministic for the comparison, so that $F(v)$ does not depend on y. With $u_N(y)$ we can evaluate the output $s_N(y) = s(u_N(y))$ as well as compute its statistics, e.g., expectation $\mathbb{E}[s_N]$, by using, e.g., Monte Carlo methods or quadrature formulas as used in stochastic collocation method. Four specific ingredients of the reduced basis method play a key role in selecting the most representative samples, hierarchically building the reduced basis space, and efficiently evaluating the outputs. They are the training set, the greedy algorithm, the a posteriori error estimate and the offline-online computational decomposition, which are addressed respectively follows.

1.3.1 Training set

Two criteria should be fulfilled in the choice of the training set: (i) we should minimize ineffectual samples in order to avoid unnecessary computation with limited gain, and (ii) sufficient to capture the most representative snapshots in order to build an accurate reduced basis space. In practice, the training set is usually chosen as randomly distributed or log-equidistantly distributed in the parameter space [178, 158]. As for stochastic problems with random variables obeying probability distributions other than uniform type, we propose to choose the samples in the training set according to the probability distributions. Furthermore, for the sake of comparison with the stochastic collocation method, we take the training set such that it contains all the collocation points used by the stochastic collocation method. Adaptive approaches for building the training set have also been explored starting from a small number of samples to a more significant set in the space Γ; see for example [212].

1.3.2 Greedy algorithm

Given a training set $\Xi_{train} \subset \Gamma$ and a first sample set $S_1 = \{y^1\}$ as well as its associated reduced basis space $X_1 = \text{span}\{u(y^1)\}$, we seek the sub-optimal solution to the $L^\infty(\Xi_{train}; X)$ optimization problem in a greedy way as [178]: for $N = 2,\dots,N_{max}$, find $y^N = \arg\max_{y\in\Xi_{train}} \triangle_{N-1}(y)$, where $\triangle_{N-1}$ is a sharp and inexpensive a posteriori error bound constructed in the current $N-1$ dimensional reduced basis space (it will be specified later). Subsequently, the sample set and the reduced basis space are enriched by $S_N = S_{N-1} \cup \{y^N\}$ and $X_N = X_{N-1} \oplus \text{span}\{u(y^N)\}$, respectively, to have an efficient hierarchical spaces and sample sets. For the sake of efficient computation of Galerkin projection and offline-online decomposition, we can normalize the snapshots by *Gram-Schmidt* process to get the

orthonormal basis of $X_N = \text{span}\{\zeta_1, \ldots, \zeta_N\}$ such that $(\zeta_m, \zeta_n)_X = \delta_{mn}$, $1 \le m, n \le N$. We remark that another algorithm that might be used for the sampling procedure is proper orthogonal decomposition (POD [178]), which is rather expensive in dealing with $L^2(\Xi_{train}; X)$ optimization and thus more suitable for low-dimensional problems.

1.3.3 A posteriori error estimate

The efficiency and reliability of the reduced basis approximation by the greedy algorithm relies critically on the availability of a cheap and sharp a posteriori error bound $\triangle_N$, which can be constructed as follows: for every $y \in \Gamma$, let $R(v; y) \in X'$ be the residual in the dual space of X, defined as

$$R(v; y) := F(v) - A(u_N(y), v; y) \quad \forall v \in X. \tag{1.23}$$

By the Riesz representation theorem [68], we have a unique function $\hat{e}(y) \in X$ such that $(\hat{e}(y), v)_X = R(v; y)\, \forall v \in X$ and $||\hat{e}(y)||_X = ||R(\cdot; y)||_{X'}$, where the X norm is defined as $||v||_X = A(v, v; \bar{y})$ at some reference value $\bar{y} \in \Gamma$ (we choose $\bar{y}$ as the center of Γ by convention). Define the error $e(y) := u(y) - u_N(y)$, by (46), (1.22), and (1.23) we have the following equation:

$$A(e(y), v; y) = R(v; y) \quad \forall v \in X. \tag{1.24}$$

By choosing $v = e(y)$ in (1.24), recalling the coercivity constant $\alpha(y)$ with the definition of its lower bound $\alpha_{LB}(y) \le \alpha(y)$ of the bilinear form $A(\cdot, \cdot; y)$, and using Cauchy–Schwarz inequality, we have

$$\alpha_{LB}(y)||e(y)||_X^2 \le A(e(y), e(y); y) = R(e(y); y) \le ||R(\cdot, y)||_{X'}||e(y)||_X = ||\hat{e}(y)||_X||e(y)||_X, \tag{1.25}$$

so that we can define the a posteriori error bound $\triangle_N$ for the solution u as

$$\triangle_N := ||\hat{e}(y)||_X / \alpha_{LB}(y), \tag{1.26}$$

which yields $||u(y) - u_N(y)||_X \le \triangle_N$ by (1.25). As for the output in the compliant case, i.e., $s = f$, we have the following error bound

$$|s(y) - s_N(y)| = |s(u(y)) - s(u_N(y))| = A(e(y), e(y); y) \le ||\hat{e}(y)||_X^2 / \alpha_{LB}(y) =: \triangle_N^s(y). \tag{1.27}$$

As for more general output where $s \neq f$, an adjoint problem of (46) can be employed to achieve a faster convergence of the approximation error $|s - s_N|$ [163]. The efficient computation of a sharp and reliable a posteriori error bound thus relies on the computation of a lower bound of the coercivity constant $\alpha_{LB}(y)$ as well as the value $||\hat{e}(y)||_X$ for any given $y \in \Gamma$. For the former, we can apply the *successive constraint linear optimization method* [102] to compute a lower bound $\alpha_{LB}(y)$ of $\alpha(y)$ or simply use a uniform lower bound $\alpha_{LB} \le \alpha(y) \forall y \in \Gamma$ provided that they are close to each other. For the latter, we turn to an efficient offline-online computational decomposition procedure.

1.3.4 Offline-online computational decomposition

The evaluation of the expectation $\mathbb{E}[s_N]$ and the a posteriori error bound $\triangle_N$ requires to compute the output s_N and the solution u_N many times. Similar situations can be encountered for other applications in the context of many query (optimal design, control) and real time computational problems. One of the key ingredients that make reduced basis method stand out in this ground is the offline-online computational decomposition, which becomes possible due to the affine assumption such as that made in (43). To start, we express the reduced basis solution in the form [178]

$$u_N(y) = \sum_{m=1}^{N} u_{Nm}(y)\zeta_m. \tag{1.28}$$

Upon replacing it in (1.22) and choosing $v = \zeta_n, 1 \leq n \leq N$, we obtain the problem of finding $u_{Nm}(y), 1 \leq m \leq N$, such that

$$\sum_{m=1}^{N} \left(A_0(\zeta_m, \zeta_n) + \sum_{k=1}^{K} y_k A_k(\zeta_m, \zeta_n) \right) u_{Nm}(y) = F(\zeta_n), \quad 1 \leq n \leq N. \tag{1.29}$$

From (1.29) we can see that the values $A_k(\zeta_m, \zeta_n), k = 0, 1, \ldots, K, 1 \leq m, n \leq N_{max}$, and $F(\zeta_n), 1 \leq n \leq N_{max}$, are independent of y, we may thus precompute and store them in the offline procedure. In the online procedure, we only need to assemble the stiffness matrix in (1.29) and solve the resulting $N \times N$ stiffness system with much less computational effort compared to solving the original $\mathcal{N} \times \mathcal{N}$ stiffness system. As for the computation of the error bound $\triangle_N(y)$, we need to compute $||\hat{e}(y)||_X$ at the selected sample y in the course of sampling procedure. We expand the residual (1.23) as

$$R(v; y) = F(v) - A(u_N, v; y) = F(v) - \sum_{n=1}^{N} u_{Nn} \left(\sum_{k=0}^{K} y_k A_k(\zeta_n, v) \right), \text{ where } y_0 = 1. \tag{1.30}$$

Set $(\mathscr{C}, v)_X = F(v)$ and $(\mathscr{L}_n^k, v)_X = -A_k(\zeta_n, v), \forall v \in X, 1 \leq n \leq N, 0 \leq k \leq K$, where $\mathscr{C}$ and $\mathscr{L}_n^k$ are the representatives of F and A_k^n (defined as $A_k^n(v) = -A_k(\zeta_n, v), \forall v \in X$) in X whose existence is guaranteed by the Riesz representation theorem. By recalling $(\hat{e}(y), v)_X = R(v; y)$, we obtain

$$||\hat{e}(y)||_X^2 = (\mathscr{C}, \mathscr{C})_X + \sum_{k=0}^{K} \sum_{n=1}^{N} y_k u_{Nn}(y) \left(2(\mathscr{C}, \mathscr{L}_n^k)_X + \sum_{k'=0}^{K} \sum_{n'=1}^{N} y_{k'} u_{Nn'}(y) (\mathscr{L}_n^k, \mathscr{L}_{n'}^{k'})_X \right). \tag{1.31}$$

Therefore, we can compute and store $(\mathscr{C}, \mathscr{C})_X, (\mathscr{C}, \mathscr{L}_n^k)_X, (\mathscr{L}_n^k, \mathscr{L}_{n'}^{k'})_X, 1 \leq n, n' \leq N_{max}, 0 \leq k, k' \leq K$, in the offline procedure, and evaluate $||\hat{e}(y)||_X$ in the online procedure by assembling (1.31).

Remark 1.3.1 *Different from the stochastic collocation method that was presented regardless of the underlying system, the reduced basis method is introduced for a linear, coercive and affine elliptic problem. In fact, the same approach presented above can be extended to more general problems [178, 163], e.g., time dependent, nonlinear, noncoercive and nonaffine problems , as long as an a posteriori error bound is cheap to obtain and the offline construction and the online evaluation can be efficiently decomposed using proper techniques [11, 129, 87, 86]; see, e.g., [162, 51, 163, 117, 175] for many different kind of recent applications.*

1.4 Comparison of convergence analysis

In this section, we provide a comparison of the theoretical convergence results between the stochastic collocation method and the reduced basis method. In the first part, a preliminary comparison is carried out based on the available convergence results in the literature at the best of our knowledge on the state of the art. Then we perform a direct comparison between the approximation errors of the two methods.

1.4.1 Preliminary comparison of convergence results

Let us first consider a priori error estimate for one-dimensional Lagrange interpolation for $y \in \Gamma = [-1, 1]$ without loss of generality. In fact, we can map any bounded interval Γ into $[-1, 1]$ by shifting and rescaling. The convergence result for the univariate approximation error is given as follows.

Proposition 1.4.1 *Thanks to the analytic regularity in Corollary 1.1.2, we have the exponential conver-*

gence rate for the one-dimensional stochastic collocation approximation error in $L^\infty(\Gamma;X)$ norm

$$||u-\mathscr{U}_N u||_{L^\infty(\Gamma;X)} \le C_N r^{-N} = C_N e^{-(\ln r)N}, \tag{1.32}$$

where $r=\sqrt{1+\tau^2}+\tau \ge (\sqrt{5}+1)/2 \approx 1.6$ owing to (1.4) and assumption (44); see [8]. The constant C_N is bounded in a logarithmic rescaling $C_N \le C\ln(N+1)$, where C is a constant independent of N.

Remark 1.4.1 *The same result has been obtained in $L^2(\Gamma;X)$ norm in [8] except that the constant C_N in (1.32) is independent of N. For the sake of comparison with the convergence rate of the reduced basis method, we consider (1.32) in the norm of $L^\infty(\Gamma;X)$ with the constant C_N depending on N.*

Proof Firstly, we demonstrate that the operator $\mathscr{U}_N : C^0(\Gamma;X) \to L^\infty(\Gamma;X)$ is continuous. In fact, by the definition of $\mathscr{U}_N$ in (1.5), we have the following estimate

$$\begin{aligned} ||\mathscr{U}_N u||_{L^\infty(\Gamma;X)} &= \sup_{y\in\Gamma}\left|\left|\sum_{n=0}^{N} u(y^n)l^n(y)\right|\right|_X \\ &\le \sup_{y\in\Gamma}\left(\sum_{n=0}^{N}|l^n(y)|\right)\max_{n=0,1,\ldots,N}||u(y^n)||_X \le \Lambda(N)||u||_{C^0(\Gamma;X)}, \end{aligned} \tag{1.33}$$

where $\Lambda(N)$ is the optimal Lebesgue constant bounded by (see [162])

$$\Lambda(N) := \sup_{y\in\Gamma}\left(\sum_{n=0}^{N}|l^n(y)|\right) \le \frac{3}{4}+\frac{2}{\pi}\ln(N+1). \tag{1.34}$$

Therefore, by the fact $\mathscr{U}_N w = w \forall\, w \in \mathscr{P}_N(\Gamma)\otimes X$ (where $\mathscr{P}_N(\Gamma)$ is the space of polynomials of order less than or equal to N), we have that for every function $u \in C^0(\Gamma;X)$,

$$||u-\mathscr{U}_N u||_{L^\infty(\Gamma;X)} \le ||u-w||_{L^\infty(\Gamma;X)} + ||\mathscr{U}_N(w-u)||_{L^\infty(\Gamma;X)} \le (1+\Lambda(N))||u-w||_{C^0(\Gamma;X)}. \tag{1.35}$$

Moreover, the following approximation error estimate holds for every function $u \in C^0(\Gamma;X)$ (see [8])

$$\inf_{w\in\mathscr{P}_N(\Gamma)\otimes X}||u-w||_{C^0(\Gamma;X)} \le \frac{2}{r-1}r^{-N}\max_{z\in\Sigma}||u(z)||_X. \tag{1.36}$$

A combination of (1.33), (1.34), (1.35), and (1.36) leads to the result stated in (1.32) with the constant C_N such that $C_N \le C\ln(N+1)$, where C depends only on $\max_{z\in\Sigma}||u(z)||_X$ and r. □

For the same one-dimensional parametric problem, a priori error estimates have been established for the reduced basis approximation [130, 178]. Note that in the context of the reduced basis approximation, the result is based on the assumption that the parameter y is positive with $0 < y_{\min} \le y \le y_{max} < \infty$. For the sake of consistent comparison with the stochastic collocation method, we still take the same parameter range $\Gamma = [-1,1]$ and introduce a new parameter by $\mu = y + (1+\delta)$ with $\delta > 0$ so that $\mu \in [\delta, 2+\delta]$ with $\mu_{min} = \delta > 0$ and $\mu_{max} = 2+\delta$. Correspondingly, the coefficient becomes $a(x,y) = a_0(x) + a_1(x)y = (a_0(x) - (1+\delta)a_1(x)) + a_1(x)\mu$ and will be denoted as $\hat{a}_0(x) + a_1(x)\mu$ for convenience. We state the convergence result for one-dimensional reduced basis approximation given in [158, 178] in the following proposition.

Proposition 1.4.2 *Suppose that $\ln\mu_r = \ln(\mu_{max}/\mu_{min}) > 1/2e$ and $N \ge N_{crit} \equiv 1 + [2e\ln\mu_r]_+$ ($[s]_+$ is the maximum integer smaller than s); then*

$$||u-u_N||_{L^\infty(\Gamma;X)} \le Ce^{-(N-1)/(N_{crit}-1)}, \tag{1.37}$$

where u_N is the reduced basis approximation of the solution in the reduced basis space spanned by N snapshots, and C is independent of N. Note that the samples $\mu_1,\dots,\mu_N$ are taken as equidistant within $[\ln(\mu_{min}),\ln(\mu_{max})]$ in the way that $\ln(\mu_n)-\ln(\mu_{n-1})=\ln(\mu_r)/(N-1), 2\le n\le N$; see [158].

To our knowledge, the a priori error estimates in Proposition 1.4.1 for the stochastic collocation approximation and in Proposition 1.4.2 for the reduced basis approximation are the state of the art results currently available in the literature. Both of them show an exponential convergence rate for the approximation of the analytic solution with respect to the parameter $y\in\Gamma$. In order to guarantee the positiveness of $\hat{a}_0(x)$ in Proposition 1.4.2, we require $\delta\le 1/2$ by assumption (44). Therefore, the minimal value of N_{crit} is $1+[2e\ln(u_r)]_+=9$, so that the convergence rate in (1.37) becomes $e^{-(N-1)/8}\approx 1.13^{-(N-1)}$ for N dimensional reduced basis approximation, which is larger than $r^{-(N-1)}$ ($r>1.6$) in the stochastic collocation approximation (1.32) using N collocation nodes corresponding to $\mathscr{U}_{N-1}$. From this closer look, it would seem that the error bound of stochastic collocation approximation converges faster than that of the reduced basis approximation in the univariate case under the above specific assumptions.

In the multivariate case, the property of convergence rate inherits that of the univariate case thanks to the full tensor product structure of the multivariate Lagrange interpolation (1.10) in the stochastic collocation approximation. A priori error estimate is obtained in the following proposition.

Proposition 1.4.3 *Under assumption (44) and the analytic regularity of the solution in Corollary 1.1.2, being $\Gamma=[-1,1]^K$ for simplicity, the following convergence result holds:*

$$||u-\mathscr{U}_N u||_{L^\infty(\Gamma;X)}\le\sum_{k=1}^{K}C_{N_k}e^{-\ln(r_k)N_k}, \tag{1.38}$$

where $r_k=\sqrt{1+\tau_k^2}+\tau_k>1, 1\le k\le K$, from (1.4) and $N=(N_1,\dots,N_K)$ is the interpolation order corresponding to the interpolation operator $(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_K})$.

Proof We split the interpolation error in (1.38) into K pairs by adding and subtracting the same term

$$\begin{aligned}
||u-\mathscr{U}_N u||_{L^\infty(\Gamma;X)} &= ||u-(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_K})u||_{L^\infty(\Gamma;X)}\\
&\le ||u-(\mathscr{U}_{N_1}\otimes\mathscr{I}\otimes\cdots\otimes\mathscr{I})u||_{L^\infty(\Gamma;X)}\\
&\quad+||(\mathscr{U}_{N_1}\otimes\mathscr{I}\otimes\cdots\otimes\mathscr{I})u-(\mathscr{U}_{N_1}\otimes\mathscr{U}_{N_2}\otimes\mathscr{I}\otimes\cdots\otimes\mathscr{I})u||_{L^\infty(\Gamma;X)}\\
&\quad+\cdots\\
&\quad+||(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_{K-1}}\otimes\mathscr{I})u-(\mathscr{U}_{N_1}\otimes\cdots\otimes\mathscr{U}_{N_K}||_{L^\infty(\Gamma;X)}\\
&\le\sum_{k=1}^{K}C_{N_k}e^{-\ln(r_k)N_k},
\end{aligned}\tag{1.39}$$

where $\mathscr{I}$ is the identity operator and $C_{N_k}\le C\ln(N_k+1)$. The first inequality is due to a recursive application of triangular inequality, while the second is a direct consequence of Proposition 1.4.1 for univariate interpolation. We remark that more general results have been obtained for unbounded Γ and arbitrarily distributed random variables other than the uniform type in [8], with norm $L^2_\rho(\Gamma;X)$ instead of $L^\infty(\Gamma;X)$. □

Remark 1.4.2 *If $C_{N_k}=C_{N_1}, r_k=r>1, 1\le k\le K$, and $N_k=N_1, 2\le k\le K$. Then the total number of collocation nodes is $N=K^{N_1}$ and the error estimate in Proposition 1.4.3 becomes*

$$||u-\mathscr{U}_N u||_{L^\infty(\Gamma;X)}\le C_{N_1}KN^{-\frac{\ln(r)}{\ln(K)}}, \tag{1.40}$$

which decays very slowly when K is large and the region of analyticity r is small. For instance, when $K = 10$ and $r = 1.6$ as in Proposition 1.4.1, we need at least $N = 10^{10}$ in order to have $KN^{-\frac{\ln(r)}{\ln(K)}} \leq 0.1$.

The convergence analysis of the isotropic and anisotropic Smolyak sparse grids stochastic collocation methods have been studied in [149] and [148] in the norm $L^2_\rho(\Gamma; X)$. Using the same argument in the proof of Proposition 1.4.1, the following results in $L^\infty(\Gamma; X)$ norm are straightforward.

Proposition 1.4.4 *Suppose that the function u can be analytically extended to a complex domain $\Sigma(\Gamma; \tau)$. By using isotropic Smolyak sparse grid and Clenshaw–Curtis collocation nodes, we have*

$$||u - \mathscr{S}_q u||_{L^\infty(\Gamma;X)} \leq C_{q-K+1} N_q^{-r}, \tag{1.41}$$

where: C_{q-K+1} is a constant depending on $q-K+1$ and r s.t. $C_{q-K+1} \leq C(r)\ln(2^{q-K+1}+2)$; $N_q = \#H(q,K)$ is the number of collocation nodes; $r = \min(\ln(\sqrt{r_1}), \ldots, \ln(\sqrt{r_K}))/(1+\ln(2K))$ with $r_1, \ldots, r_K$ defined in (1.38). Using the anisotropic Smolyak sparse grid with Clenshw-Curtis collocation nodes, we have

$$||u - \mathscr{S}_q^\alpha u||_{L^\infty(\Gamma;X)} \leq C_{q-K+1} N_q^{-r(\alpha)}, \tag{1.42}$$

where $r(\alpha) = \min(\alpha)(\ln(2)e - 1/2)/\left(\ln(2) + \sum_{k=1}^K \min(\alpha)/\alpha_k\right)$ and $\alpha_k = \ln(\sqrt{r_k}), k = 1, \ldots, K$.

As for the reduced basis approximation in multivariate problems, there is unfortunately no direct a priori error estimate in the literature. However, there is indeed a comparison between the Kolmogorov N-width given by (slightly different from the notation in [20])

$$d_N(\Gamma; X) := \inf_{\dim(S_N)=N} \sup_{y\in\Gamma} \inf_{w_N \in X_N} ||u(y) - w_N||_X, \tag{1.43}$$

which defines the error of the optimal approximation, and the convergence rate of the N dimensional reduced basis approximation error by the greedy algorithm. In (1.43), the notations are the same as in section 1.3: S_N is a subset of samples with cardinality N; $X_N = \text{span}\{u(y), y \in S_N\}$ is a function space spanned by the snapshots. Essentially, the Kolmogorov N-width measures the error of the best or optimal N dimensional approximation over all possible N dimensional approximation. As for the reduced basis approximation with the reduced basis space X_N constructed from a greedy algorithm, we define its $L^\infty(\Gamma)$ error as

$$\sigma_N(\Gamma) = \sup_{y\in\Gamma} \inf_{w_N \in X_N} ||u(y) - w_N||_X. \tag{1.44}$$

In practice we use a posteriori error estimator $\triangle_N$ as introduced in section 1.3 instead of the true error $\inf_{w_N\in X_N} ||u(y) - w_N||_X$ for the greedy selection of quasi-optimal samples, which satisfies

$$c\triangle_N \leq \inf_{w_N\in X_N} ||u(y) - w_N||_X \leq C\triangle_N, \text{ where } 0 < \gamma \equiv \frac{c}{C} \leq 1. \tag{1.45}$$

A recent result [20] established a relation between the Kolmogorov width d_N and the reduced basis approximation error σ_N, which is summarized in the following proposition.

Proposition 1.4.5 *Suppose that $\exists M > 0$ s.t. $d_0(\Gamma; X) \leq M$. Moreover, assume that $\exists r > 0$,*

$$\text{if } d_N(\Gamma; X) \leq MN^{-r} \text{ then } \sigma_N(\Gamma) \leq CMN^{-r} \quad \forall N > 0, \tag{1.46}$$

where the constant C depends only on r and γ. Moreover, assume that $\exists a > 0$,

$$\text{if } d_N(\Gamma; X) \leq Me^{-aN^r} \text{ then } \sigma_N(\Gamma) \leq CMe^{-cN^s} \quad \forall N \geq 0, \tag{1.47}$$

where the constants $s = r/(r+1)$ and c, C depends only on a, r and γ.

This proposition basically states that whenever the Kolmogorov width decays at either an algebraic or exponential rate, the greedy algorithm will also generate a quasi-optimal approximation space with the error decaying in a similar way. By the definition of Kolmogorov width, which measures the error of the optimal approximation among all the possible approximations, we have that the stochastic collocation approximation error can not be smaller than or decay faster than the Kolmogorov width. In particular, we have that the Kolmogorov width is smaller than the isotropic and anisotropic sparse grid collocation approximation error, i.e.,

$$d_{N_q}(\Gamma;X) \leq \min\{||u-\mathscr{S}_q u||_{L^\infty(\Gamma;X)}, ||u-\mathscr{S}_q^\alpha u||_{L^\infty(\Gamma;X)}\}. \tag{1.48}$$

If $d_{N_q}(\Gamma;X) \leq MN_q^{-\tilde{r}}$ then $\tilde{r} \geq \max\{r, r(\alpha)\}$, where r and $r(\alpha)$ are the algebraic convergence rates in (1.41) and (1.42), respectively. Therefore, we have the following a priori error estimate: the reduced basis approximation error $\sigma_{N_q}(\Gamma) \leq CMN_q^{-\tilde{r}}$, which decays faster than the stochastic collocation approximation error. Moreover, if the stochastic solution is analytic in the probability space, as is the case for the elliptic problem (39) with analytic solution in Corollary 1.1.2, the Kolmogorov width can achieve exponential convergence rate for analytical problems in practice [20], so that the reduced basis approximation error also decays exponentially and much faster than the stochastic collocation approximation error, as demonstrated by numerical experiments presented and discussed in section 1.6.

Both the Kolmogorov width $d_N(\Gamma;X)$ and the greedy error $\sigma_N(\Gamma)$ are provided on the whole region Γ. However, in practice they are defined over the training set $\Xi_{train} \subset \Gamma$. When the latter is dense enough, i.e., $d_N(\Gamma;X)$ and $\sigma_N(\Gamma)$ are indistinguishable from $d_N(\Xi_{train};X)$ and $\sigma_N(\Xi_{train})$, the comparison above is valid. On the other hand, if the training set Ξ_{train} is rather sparse in Γ, which is usually the case in high-dimensional problem, the comparison might be invalid. In order to have more rigorous and fair comparison of the reduced basis approximation and the stochastic collocation approximation, we perform a direct comparison of their approximation errors in the next section.

1.4.2 Direct comparison of approximation errors

As mentioned above, selecting an appropriate training set Ξ_{train} for the reduced basis approximation is crucial. For its effective comparison with the stochastic collocation approximation, we choose training set as the set represented by the collocation points used in the latter approximation, which we denote as Ξ_{sc} in general for both the full tensor product grid and the sparse grid. Let us denote also the interpolation formula on Ξ_{sc} as $\mathscr{I}_{sc} : C^0(\Gamma;X) \to L^\infty(\Gamma;X)$; then we have the following proposition for a direct comparison.

Proposition 1.4.6 *Provided that the training set Ξ_{train} for the reduced basis approximation is taken the same as the collocation set Ξ_{sc} for the stochastic collocation approximation, we have*

$$||u-u_N||_{L^\infty(\Gamma;X)} \leq C||u-\mathscr{I}_{sc}u||_{L^\infty(\Gamma;X)}, \tag{1.49}$$

where $C = 3a_{max}/a_{min}$ (with a_{max}, a_{min} defined in (41)) is a constant independent of N.

Proof By the definition of the reduced basis approximation u_N in (1.28), we have

$$\begin{aligned} ||u-u_N||_{L^\infty(\Gamma;X)} &= \sup_{y\in\Gamma} ||u(y)-u_N(y)||_X \\ &\leq \frac{C}{3} \sup_{y\in\Gamma} \inf_{w\in X_N} ||u(y)-w||_X \\ &\leq \frac{C}{3} \sup_{y\in\Xi_{sc}/S_N} \inf_{w\in X_N} ||u(y)-w||_X + \frac{C}{3} \sup_{y\in\Gamma/\Xi_{sc}} \inf_{w\in X_N} ||u(y)-w||_X, \end{aligned} \tag{1.50}$$

where $C = 3a_{max}/a_{min}$ is a constant independent of N according to the Céa lemma [165]; the first inequality is due to the property of Galerkin projection on the space X_N, the second one comes from the fact that $\Gamma = S_N \cup (\Xi_{sc}/S_N) \cup (\Gamma/\Xi_{sc})$, and the reduced basis approximation error vanishes for any $y \in S_N$ so that only the last two terms in (1.50) remain. For the second term of (1.50), we have

$$\sup_{y\in\Gamma/\Xi_{sc}} \inf_{w\in X_N} ||u(y)-w||_X \leq \sup_{y\in\Gamma/\Xi_{sc}} \inf_{v\in X_{sc}} ||u(y)-v||_X + \inf_{w\in X_N} ||v-w||_X, \tag{1.51}$$

where the function v is defined in the space X_{sc} that is spanned by the solutions at the collocations points in Ξ_{sc}. Therefore, the first term of (1.51) satisfies

$$\sup_{y\in\Gamma/\Xi_{sc}} \inf_{v\in X_{sc}} ||u(y)-v||_X \leq \sup_{y\in\Gamma/\Xi_{sc}} ||u(y)-\mathscr{I}_{sc}u(y)||_X = \sup_{y\in\Gamma} ||u(y)-\mathscr{I}_{sc}u(y)||_X, \tag{1.52}$$

and the second term of (1.51) can be bounded by noting that v is one solution at some $y \in \Xi_{sc}$ as

$$\inf_{w\in X_N} ||v-w||_X \leq \sup_{y\in\Xi_{sc}} \inf_{w\in X_N} ||u(y)-w||_X = \sup_{y\in\Xi_{sc}/S_N} \inf_{w\in X_N} ||u(y)-w||_X. \tag{1.53}$$

A combination of (1.50), (1.51), (1.52), and (1.53) leads to the following error bound:

$$||u-u_N||_{L^\infty(\Gamma;X)} \leq \frac{2C}{3} \sup_{y\in\Xi_{sc}/S_N} \inf_{w\in X_N} ||u(y)-w||_X + \frac{C}{3} \sup_{y\in\Gamma} ||u(y)-\mathscr{I}_{sc}u(y)||_X. \tag{1.54}$$

Moreover, we can construct the reduced basis space in such a way that the reduced basis approximation error in the collocation/training set Ξ_{sc} (the first term of (1.54)) is smaller than the stochastic collocation approximation error over Γ (the second term of (1.54)), i.e.,

$$\sup_{y\in\Xi_{sc}/S_N} \inf_{w\in X_N} ||u(y)-w||_X \leq \sup_{y\in\Gamma} ||u(y)-\mathscr{I}_{sc}u(y)||_X, \tag{1.55}$$

which is always viable and an extreme case is that all the collocation points are included in the sample set, i.e., $\Xi_{sc} = S_N$, so that the first term of (1.54) vanishes. Therefore, by substituting (1.55) into (1.54), we obtain (1.49). □

Since the evaluation of the statistics by the Monte Carlo algorithm converges very slowly, we propose the approach of evaluating the solution by the reduced basis method at all the collocation nodes first and then applying quadrature formula (56) to assess the statistics. To improve the accuracy of this approach, we also build the training set Ξ_{train} as the collocation/quadrature nodes $\Xi_{sc} = \Xi_{train}$. In fact, we have the error estimate between the expectation $\mathbb{E}[s]$ and the value $\mathbb{E}[s_{rb}]$ approximated by reduced basis method ($\mathbb{E}[s_{sc}]$ is the value approximated by the stochastic collocation method)

$$|\mathbb{E}[s]-\mathbb{E}[s_{rb}]| \leq |\mathbb{E}[s]-\mathbb{E}[s_{sc}]| + |\mathbb{E}[s_{sc}]-\mathbb{E}[s_{rb}]|, \tag{1.56}$$

where the first term represents the quadrature error, and the second term is bounded by (1.27) as

$$\begin{aligned} |\mathbb{E}[s_{sc}]-\mathbb{E}[s_{rb}]| &\leq \sum_{y^i\in\Xi_{sc}} w^i |s(y^i)-s_{rb}(y^i)| \\ &\leq \max_{y\in\Xi_{sc}} |s(y)-s_{rb}(y)| \\ &\leq \max_{y\in\Xi_{sc}} ||s||_{X'} ||u(y)-u_N(y)||_X, \end{aligned} \tag{1.57}$$

where $w^i > 0$ are quadrature weights. As long as the reduced basis approximation error is smaller than the quadrature error, (1.56) is dominated by the first term, i.e. the quadrature error.

1.5 Comparison of computational costs

In this section, we aim at comparing in detail the computational costs in terms of operations count and storage of the reduced basis method and the stochastic collocation method. Let us begin with the computational cost ($C(\cdot)$ stands for operations count and $S(\cdot)$ for storage) for the stochastic collocation method, which is listed along side Algorithm 1. The major computational cost for the reduced basis method is listed along side Algorithm 2.

A few notations are given in order: $N_{sc} = \#\Theta = (N_1+1)\times\cdots\times(N_K+1)$; $N_t = \#\Xi_{train}$; $N_{rb} = N_{max}$; W_α is the average work to evaluate the lower bound α_{LB} over the training set; W_s is the work to solve once the linear system arising from (46) with $C(\mathcal{N}^2) \le W_s \le C(\mathcal{N}^3)$, and W_m is the work to evaluate $(\mathscr{L},\mathscr{L})_X$ in (1.31) with $C(\mathcal{N}) \le W_m \le C(\mathcal{N}^2)$. The total computational costs (apart from that of the common initialization) for the reduced basis method and stochastic collocation method are calculated from Algorithms 1 and 2 and presented in Table 1.1.

computational cost	SC	RB
offline operations count	$C(N_{sc}(W_s+K))$	$C(N_tW_\alpha + N_{rb}W_s + KN_{rb}^2\mathcal{N} + K^2N_{rb}^2W_m + N_tK^2N_{rb}^3)$
online operations count	$C(N_{sc})$	$C(N_{rb}^3 + KN_{rb}^2 + K^2N_{rb}^2)$
total storage	$S(N_{sc}(\mathcal{N}+K))$	$S(N_{rb}\mathcal{N} + K^2N_{rb}^2 + KN_t)$

Table 1.1: Computational costs of the stochastic collocation (SC) and the reduced basis (RB) methods

More in detail, the offline cost for the stochastic collocation method is dominated by the solution of the problem (46) for N_{sc} times with total operations $C(N_{sc}(W_s+K))$. Its online cost scales as $C(N_{sc})$ by the multivariate barycentric formula or quadrature formula. The total storage is dominated by that for all the solutions $S(N_{sc}(\mathcal{N}))$. As for the reduced basis method, the offline cost is the sum of that for precomputing the lower bound $C(N_tW_\alpha)$, solving the system for N_{rb} times with total operations $C(N_{rb}W_s + KN_{rb}^2\mathcal{N})$, computing error bound with operations $C(K^2N_{rb}^2W_m)$, and searching in the training set with operations $C(N_tK^2N_{rb}^3)$. The online cost is the sum of that for assembling (1.29) with operations $C(KN_{rb}^2)$, solving it with operations $C(N_{rb}^3)$, and evaluating the error bound with operations $K^2N_{rb}^2$; as for statistics by quadrature formula, we need $C(N_{sc}(N_{rb}^3 + KN_{rb}^2))$ operations. The total storage for the reduced basis method takes $S(N_{rb}\mathcal{N} + K^2N_{rb}^2 + KN_t)$ for storing the solution, stiffness matrix as well as the training set.

From Table 1.1 we can observe that an explicit comparison of computational costs for the reduced basis method and the stochastic collocation method crucially depends on the number of collocation points N_{sc}, the size of the training set N_t, the dimension of the reduced basis N_{rb} and parameters K, and the work of computing the lower bound W_α. In general, provided that the problem is computationally consuming in the sense that $\mathcal{N}$ is very large, and provided that $N_{sc} \approx N_t$, we have $N_{rb} \ll N_{sc}$ so that the reduced basis method is much more efficient in the offline procedure under the condition that $W_\alpha \ll W_s$ by the successive constraint linear optimization algorithm. As for the online evaluation of the solution at a new $y \in \Gamma$, this advantage becomes even more evident especially in high dimensions since the online operations count for the reduced basis method is much smaller than that for the stochastic collocation method, i.e., $C(N_{rb}^3 + KN_{rb}^2 + K^2N_{rb}^2) \ll C(N_{sc})$. However, as for the evaluation of the statistics, e.g., expectation $\mathbb{E}[u]$, the online operations count $C(N_{sc}(N_{rb}^3 + KN_{rb}^2))$ is larger for the reduced basis method than the online operations count ($C(N_{sc})$) for the stochastic collocation method. Moreover, if we choose the size of the training set larger than the number of collocation points $N_t \gg N_{sc}$, which is usually the case in practice for low-dimensional problems ($K = 1,2,3$), or else the work W_α for the computation of the lower bound α_{LB} is not significantly smaller than W_s, the stochastic collocation method can perform as well as or even better than the reduced basis method when $N_t \gg N_{sc}$.

Algorithm 1 The stochastic collocation method

procedure OFFLINE CONSTRUCTION
  Initialization: mesh, parameters, finite element functions $\varphi_i, 1 \le i \le \mathcal{N}$, etc;
  Precompute and store stiffness matrix $(A_k)_{ij} = A_k(\varphi_i, \varphi_j), 0 \le k \le K$ and vector $(F)_i = F(\varphi_i)$;

  Precompute and store the collocation nodes $\Theta = \Theta^1 \times \cdots \times \Theta^K$; ▷ $C(N_{sc})/S(KN_{sc})$
  for $k = 1, \ldots, K$ **do**
    for $n_k = 0, \ldots, N_k$ **do**
      Precompute and store the barycentric weights $\bar{w}_k^{n_k}(y_k^{n_k}), y_k^{n_k} \in \Theta_k$; ▷ $C(N_k)/S(1)$
      Precompute and store quadrature weights $w_{n_k}^k$ by formula (1.14); ▷ $C(N_k)/S(1)$
    end for
  end for
  for $n = 1, \ldots, N_{sc}$ **do**
    Compute and store the solution $u(y^n), y^n \in \Theta$; ▷ $C(W_s)/S(\mathcal{N})$
  end for
end procedure

procedure ONLINE EVALUATION
  Given $y \in \Gamma$, compute the solution $u(y)$ by interpolation (1.12), (1.15), or (1.19); ▷ $C(N_{sc})$
  Evaluate the expectation $\mathbb{E}[u]$ by (1.13); ▷ $C(N_{sc})$
end procedure

Algorithm 2 The reduced basis method

procedure OFFLINE CONSTRUCTION
  Initialization: mesh, parameters, finite element functions $\varphi_i, 1 \le i \le \mathcal{N}$, tolerance ε, etc.;
  Precompute and store stiffness matrix $(A_k)_{ij} = A_k(\varphi_i, \varphi_j), 0 \le k \le K$ and vector $(F)_i = F(\varphi_i)$;

  Precompute and store Ξ_{train} and $\alpha_{LB}(y), y \in \Xi_{train}$ by SCM; ▷ $C(N_t W_\alpha)/S(N_t)$
  Initialize $y^1 \in \Xi_{train}$, $S_1 = \{y^1\}$, $X_1 = \{\zeta_1\}, \zeta_1 = u(y^1)/||u(y^1)||_X$; ▷ $C(W_s)/S(\mathcal{N})$
  Compute and store $A_k(\zeta_1, \zeta_1)$ and $F(\zeta_1), 0 \le k \le K$; ▷ $C(K\mathcal{N})/S(1)$
  Compute and store $(\mathscr{C}, \mathscr{C})_X, (\mathscr{C}, \mathscr{L}_1^k)_X, (\mathscr{L}_1^k, \mathscr{L}_1^{k'})_X, 0 \le k, k' \le K$; ▷ $C(K^2 W_m)/S(K^2)$
  for $N = 2, \ldots, \bar{N}_{max}$ **do**
    Compute $\triangle_{N-1}^u(y) = ||\hat{e}(y)||_X / \alpha_{LB}(y)$ by (1.31); ▷ $C(K^2N^2N_t)/S(N_t)$
    Choose $y^N = \arg\max_{y \in \Xi_{train}} \triangle_{N-1}^u(y)$; ▷ $C(N_t)/S(1)$
    if $\triangle_{N-1}^u(y^N) \le \varepsilon$ **then**
      $N_{max} = N - 1$; **Break**;
    end if
    Set $S_N = S_{N-1} \cup y^N$ and compute $u(y^N)$; ▷ $C(K + W_s)/S(\mathcal{N})$
    Orthogonalize $X_N = \text{span}\{\zeta_1, \ldots, \zeta_{N-1}, u(y^N)\}$; ▷ $C(\mathcal{N})/S(\mathcal{N})$
    Compute and store $A_k(\zeta_m, \zeta_n)$ and $F(\zeta_N)$ for (1.29); ▷ $C(KN\mathcal{N})/S(N^2)$
    Compute and store $(\mathscr{C}, \mathscr{L}_N^k)_X, (\mathscr{L}_n^k, \mathscr{L}_{n'}^{k'})_X$ for (1.31); ▷ $C(K^2NW_m)/S(K^2N)$
  end for
end procedure

procedure ONLINE EVALUATION
  Given $y \in \Gamma$, assemble and solve (1.29) and compute $\triangle_N(y)$; ▷ $C(N_{rb}^3 + KN_{rb}^2 + K^2N_{rb}^2)$
  Evaluate statistics by quadrature formula with N_{sc} abscissas; ▷ $C(N_{sc}(N_{rb}^3 + KN_{rb}^2))$
end procedure

1.6 Numerical experiments

In this section, we provide numerical substantiation to our previous analysis on the convergence rate and on the computational costs for the comparison of the reduced basis method and the stochastic collocation method. More precisely, we consider a stochastic elliptic problem in a two dimensional unit square $D = (0,1)^2$ with element $x = (x_1, x_2)$. The deterministic forcing term $f = 1$ is fixed. The coefficient $a(x,\omega)$ is a random field (depending only on x_1) with finite second moment, whose expectation and correlation are given as

$$\mathbb{E}[a](x) = \frac{c}{100}, \text{ for a suitable } c > 0; \quad \mathbb{C}[a](x,x') = \frac{1}{100^2}\exp\left(-\frac{(x_1 - x_1')^2}{L^2}\right), \quad x, x' \in D, \tag{1.58}$$

where L is the correlation length. The Karhunen-Loève expansion of the random field a is

$$a(x,\omega) = \frac{1}{100}\left(c + \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1(\omega) + \sum_{n=1}^{\infty}\sqrt{\lambda_n}\left(\sin(n\pi x_1)y_{2n}(\omega) + \cos(n\pi x_1)y_{2n+1}(\omega)\right)\right), \tag{1.59}$$

where the uncorrelated random variables $y_n, n \geq 1$, have zero mean and unit variance, and the eigenvalues $\lambda_n, n \geq 1$, have the following expression

$$\sqrt{\lambda_n} = \left(\sqrt{\pi}L\right)^{1/2}\exp\left(-\frac{(n\pi L)^2}{8}\right), \quad n \geq 1. \tag{1.60}$$

The random field $a(x,\omega)$ will be chosen as in (1.61) and (1.64) below. All the numerical computation is performed in MATLAB on an Intel Core i7-2620M Processor of 2.70 GHz.

1.6.1 Numerical experiments for a univariate problem

For the test of a univariate stochastic problem, we take

$$a(x,\omega) = \frac{1}{100}\left(1 + \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2}\sin(2\pi x_1)y_1(\omega)\right), \tag{1.61}$$

where $y_1(\omega)$ obeys uniform distribution with zero mean and unit variance $y_1(\omega) \sim \mathcal{U}(-\sqrt{3},\sqrt{3})$. We implement Algorithm 1 for the stochastic collocation approximation with Clenshaw–Curtis nodes (the same as Chebyshev-Gauss-Lobatto nodes [35, 199]), defined for $y_1 \in \Gamma_1 = [-\sqrt{3},\sqrt{3}]$ as

$$y_1^n = -\sqrt{3}\cos\left(\frac{n\pi}{N}\right), \quad n = 0,\ldots,N. \tag{1.62}$$

We also implement Algorithm 2 for the reduced basis approximation with equidistant training set Ξ_{train} with cardinality $N_t = 1000$, which is rather dense in the interval $[-\sqrt{3},\sqrt{3}]$. We take randomly the testing set Ξ_{test} with $N_{test} = 1000$ samples and define the $L^\infty(\Gamma)$ error between the "true" solution u (here finite element solution) and approximate solution u_{approx} as

$$||u - u_{approx}||_{L^\infty(\Gamma;X)} \approx \max_{y\in\Xi_{test}} ||u(y) - u_{approx}(y)||_X. \tag{1.63}$$

We also compute the statistical error $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$ with the expectation defined in (1.9).

Figure 1.2 illustrates the convergence of the error with respect to the number of collocation nodes and reduced bases for the stochastic collocation approximation and reduced basis approximation, respectively. From the left side of Figure 1.2, we observe that both approximations achieve the ex-

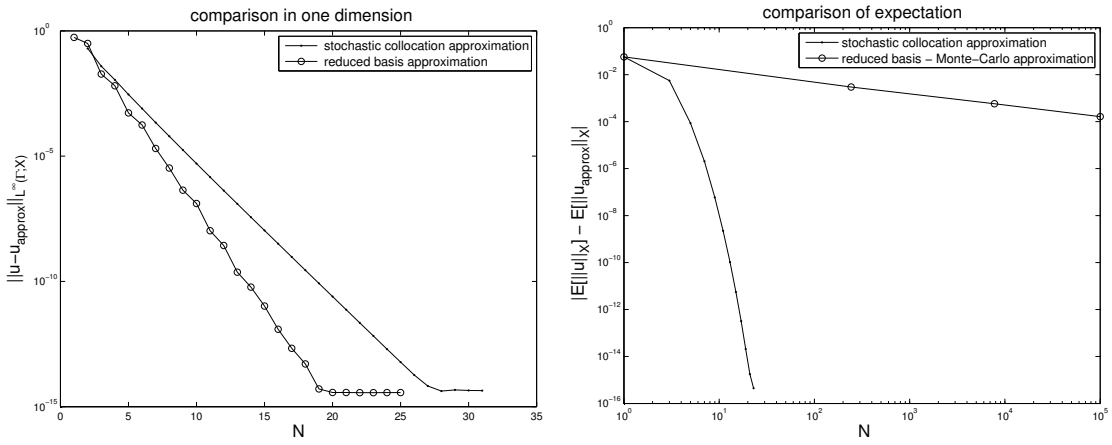

Figure 1.2: Comparison for convergence rate of the error $||u-u_{approx}||_{L^\infty(\Gamma;X)}$ (left) and the expectation $|\mathbb{E}[||u||_X]-\mathbb{E}[||u_{approx}||_X]|$ (right) between the true and the approximate solutions in 1D

ponential convergence rates in accordance with Propositions 1.4.1 and 1.4.2. The reduced basis approximation (with convergence rate $\approx \exp(-1.8N)$) turns out to be slightly better than the stochastic collocation approximation (with convergence rate $\approx \exp(-1.3N)$). As for the computation of the expectation $\mathbb{E}[||u-u_{approx}||_X]$, we apply Clenshaw–Curtis rule [199] for the stochastic approximation and Monte-Carlo algorithm for the reduced basis approximation. The right side of Figure 1.2 shows that the quadrature rule with exponential convergence rate $\approx \exp(-1.6N)$ is apparently superior to Monte-Carlo algorithm with algebraic convergence rate $\approx N^{-1/2}$ for the univariate problem.

As for the computational costs, though the reduced basis approximation needs slightly fewer snapshots than the stochastic collocation approximation, it costs more for the computation of a posteriori error estimator by greedy sampling over a large training set in the offline construction. In Table 1.2 for the univariate problem, we observe that for small-scale problems, i.e., the mesh size h is large, the offline construction of the reduced basis approximation is apparently more expensive than the stochastic collocation approximation. When the problem becomes large-scale, i.e., the mesh size h is small, the computational cost is dominated by the cost required for the solution of the finite element problem; then the reduced basis approximation is as efficient as the stochastic collocation approximation or even better. Moreover, it takes $C(N_{SC}) = C(28)$ operations count for the online evaluation of the solution $u(y)$ for any given $y \in \Gamma$ by the stochastic collocation method while the reduced basis method needs more computation $C(N_{RB}^3) = C(8000) > C(N_{SC}) = C(28)$. From Table 1.2 we can see that the online computational cost of the reduced basis approximation increases with the scale of the problem and takes more cost than that of the stochastic collocation approximation, which depends only on the number of collocation points N_{SC}. In the computation of expectation, the reduced basis - Monte-Carlo approximation is much more expensive than the stochastic collocation approximation with corresponding quadrature rule for the univariate problem. In order to alleviate the computational costs, we can first evaluate the solution at the collocation nodes by the reduced basis method and then use the quadrature formula to compute the expectation. However, this is not so useful if the number of collocation nodes is comparable to the number of reduced bases, as in the univariate case. We will compare the proposed approach with the stochastic collocation method for multivariate case later. From the univariate experiment, we conclude that the stochastic collocation approximation is more efficient than the reduced basis approximation for small-scale problem in terms of computational costs and become less efficient as the problem becomes large-scale and expensive to solve.

Figure 1.3 depicts the procedure of the reduced basis construction by greedy sampling algorithm and

time t(s) | size h	1/8	1/16	1/32	1/64	1/128
t_{RB}(1D,$N_t = 10^3$)	4(0.0003)	7(0.0003)	12(0.002)	14(0.005)	33(0.02)
t_{SC}(1D,$N_{SC} = 28$)	0.04(0.0002)	0.1(0.0002)	1(0.0002)	6(0.0002)	31(0.0002)

Table 1.2: 1D offline (online in brackets) computational costs measured in CPU time by the reduced basis (RB) and the stochastic collocation (SC) methods achieving the same accuracy.

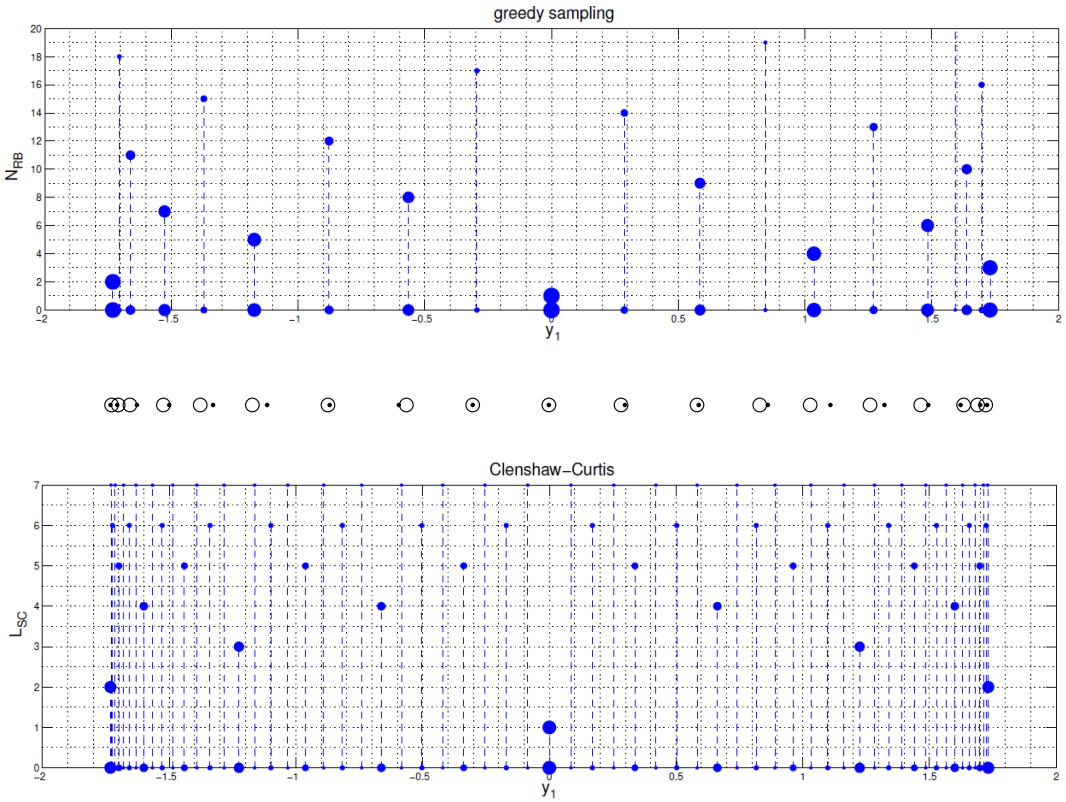

Figure 1.3: Comparison of greedy sampling (top) and hierarchical Clenshaw–Curtis rule (bottom). The bigger the size of the nodes, the earlier they are selected in the hierarchical approximation. Middle: the distribution of the greedy samples for reduced basis ($\circ$) and Clenshaw–Curtis nodes ($\cdot$).

the hierarchical stochastic collocation construction based on Clenshaw–Curtis nodes. At the top of Figure 1.3, we use larger size of dots to show earlier samples selected in the greedy algorithm, which is very similar to the hierarchical collocation construction shown at the bottom of Figure 1.3 in terms of the position and selected order of the nodes. This effect can be observed more closely in the middle figure, where the greedy samples is in full consistency with the Clenshaw–Curtis nodes. In fact, the maximum distance of the corresponding points between the greedy samples and the Clenshaw–Curtis nodes (CC) is 0.074, and the mean distance is 0.023. For comparison, we also test Chebyshev–Gauss nodes (CG), Legendre–Gauss nodes (LG) and Legendre–Gauss–Lobatto nodes (LGL) (see[35]), and the results are listed in Table 1.3, from which we can see that Clenshaw–Curtis nodes are the best choice, followed by Legendre–Gauss–Lobatto nodes. Note that the average distances of the samples in the training set are $2\sqrt{3}/1000 = 0.0035$ and $2\sqrt{3}/10000 = 0.00035$, which are much smaller than the quantities in Table 1.3, so that we are confident with the intrinsic difference between the samples selected by the greedy algorithm and the collocation nodes. This numerical coincidence has also been observed for empirical

interpolation method [11, 129], which is efficiently used in affinely approximation of nonlinear terms for nonlinear problems in the framework of the reduced basis approximation. This fact sheds light on the similarity of projection and interpolation in the common framework of nonlinear approximation, in the way that the greedy algorithm for the reduced basis projection tends to select the points on which the Lebesgue constant, arising in the stochastic collocation/interpolation, is minimized.

N_t	CC	CG	LG	LGL
1000	0.074(0.023)	0.108(0.033)	0.131(0.047)	0.082(0.024)
10000	0.076(0.022)	0.110(0.034)	0.134(0.049)	0.085(0.024)

Table 1.3: Comparison of the maximum distance (average distance in $(\cdot)$) between greedy samples in the reduced basis approximation and collocation nodes for the stochastic collocation approximation.

1.6.2 Numerical experiments for multivariate problems

For the test of a multivariate problem, we truncate the random field $a(x,\omega)$ from Karhunen-Loève expansion (1.59) with five uniformly distributed random variables $y = (y_1, \cdots, y_5) \in \Gamma = [-\sqrt{3}, \sqrt{3}]^5$, and the correlation length $L = 1/8$ so that the two eigenvalues $\lambda_1 \approx 0.2132, \lambda_2 \approx 0.1899$, written as

$$a(x,\omega) = \frac{1}{100}\left(4 + \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1(\omega) + \sum_{n=1}^{2} \sqrt{\lambda_n}\left(\sin(n\pi x_1) y_{2n}(\omega) + \cos(n\pi x_1) y_{2n+1}(\omega)\right)\right). \tag{1.64}$$

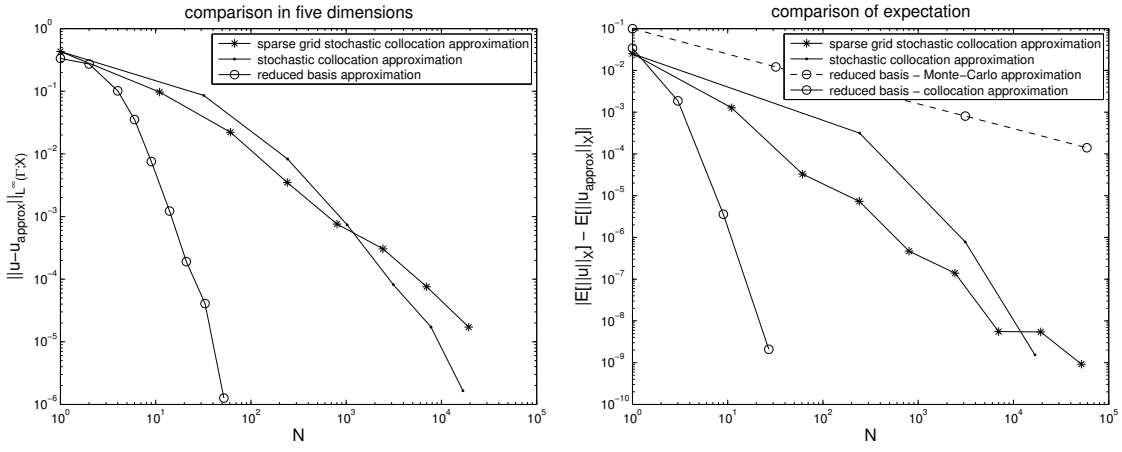

Figure 1.4: Comparison for the convergence rate of the error $||u - u_{approx}||_{L^\infty(\Gamma;X)}$ (left) and the expectation $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$ (right) between the true and the approximated solutions in 5D.

The tensor product of one-dimensional Clenshaw–Curtis nodes (1.62) for $N = 1,2,3,4,5,6,7$ as well as a single node $[0,0,0,0,0]$ are used for the stochastic collocation approximation, while the Smolyak sparse grid with level $q - 5 = 1,2,3,4,5,6,7$ are used for the stochastic sparse grid collocation approximation. For the reduced basis approximation, we select the same 7^5 samples as used in the tensor product stochastic collocation nodes. The convergence results for $L^\infty(\Gamma)$ error and the expectation error are displayed in Figure 1.4. From the left side of Figure 1.4, we observe obviously a larger convergence rate for the reduced basis approximation (still achieving an exponential convergence rate $\approx \exp(-0.2N)$) than the stochastic collocation approximation (only gaining a convergence rate $\approx \exp(0.0002N)$ or

time t(s) | size h	1/8	1/16	1/32	1/64	1/128
t_{RB}(5D,$N_t = 10^3$)	50(0.0008)	55(0.001)	57(0.002)	76(0.01)	159(0.05)
t_{RB}(5D,$N_t = 7^5$)	839(0.0005)	843(0.001)	846(0.002)	864(0.009)	949(0.05)
t_{SC}(5D,$N_{SC} = 7^5$)	17(0.02)	58(0.02)	755(0.02)	3619(0.02)	17252(0.02)

Table 1.4: 5D offline (online in brackets) computational costs measured in CPU time by the reduced basis (RB) and the stochastic collocation (SC) methods achieving the same accuracy.

rather an algebraic convergence rate $\approx N^{-1.5}$). The sparse grid collocation achieves more accurate approximation than the tensor product collocation at the beginning, and loses this advantage to the latter due to slower convergence for our specific experiment in five dimensions (5D).

As for the convergence of the expectation $E[||u||_X]$, as seen from the right side of Figure 1.4, the highest convergence rate (gaining an exponential convergence rate) is still achieved by the reduced basis - collocation approximation, essentially by constructing the reduced basis at first and then evaluating the solution at the collocation/quadrature points by the reduced basis approximation. Similar convergence behaviour can be observed for the tensor product and the sparse grid collocation approximations, which are still better than the reduced basis - Monte-Carlo approximation, though this advantage becomes less important than that in the univariate case.

For the comparison of computational costs, besides the same 7^5 training samples as used in the tensor product stochastic collocation nodes, we also use $N_t = 1000 \ll 7^5$ randomly generated samples as the training set and obtain the same number of reduced bases to achieve the same accuracy due to the smoothness of the solution in the parameter space. From Table 1.4, we may see that the offline computational cost for the stochastic collocation approximation grows exponentially fast as the complexity of the problem, while for the reduced basis approximation, it increases slightly and is dominated linearly by the cardinality of the training set Ξ_{train} from the comparison between $7^5 \approx 1.7 \times 10^4$ and 10^3, which is almost the same ratio of the CPU time $839/50 \approx 17$. In comparison, the reduced basis approximation becomes much more efficient than the stochastic collocation approximation in offline construction for large-scale problems while it loses moderately to the latter for the online computational cost. In the computation of the expectation, the reduced basis - collocation approximation is much faster than the stochastic collocation approximation: $949(\text{offline}) + 0.05 \times 7^5(\text{online}) \approx 1789 \ll 17252$ for large-scale problem ($h = 1/128$) while this becomes opposite for small-scale problem ($h = 1/8$) since $839 + 0.0005 \times 7^5 \approx 847 \gg 17$.

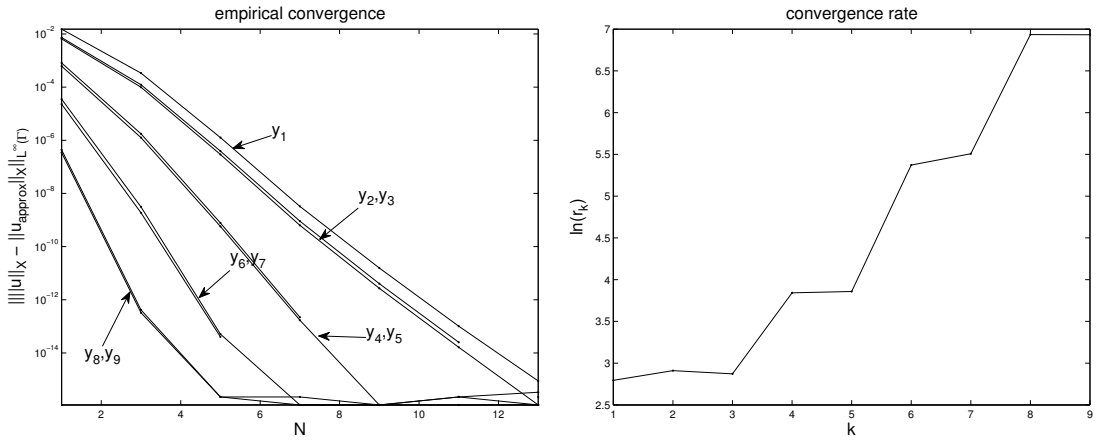

Figure 1.5: Empirical convergence (left) and fitted convergence (right) rates in dimension $1 \le k \le 9$.

In solving the five dimensional stochastic problems, we can see that both the stochastic collocation and the reduced basis methods achieve better convergence property than the Monte-Carlo algorithm. However, when the number of random variables or parameters becomes very large, the tensor product stochastic collocation approximation would need too many collocation points so that the quadrature formula losses its advantage over the Monte-Carlo algorithm. Meanwhile, the size of the training set for reduced basis construction also grows exponentially with the dimensions of the problem. Therefore, it is necessary to alleviate the computational cost. When the random variables $y_k, 1 \le k \le K$ have different importance for the stochastic problem, it would be worthless to put the same weight on the ones with little importance as on those with much larger influence. For instance, the first few eigenvalues $\lambda_1 \approx 0.4782, \lambda_2 \approx 0.0752, \lambda_3 \approx 0.0034, \lambda_4 \approx 0.000045$ decay so fast for a large correlation length ($L = 1/2$) in the Karhunen-Loève expansion (1.59) that the random variables have distinct weights in determining the value of the coefficient $a(x, y_1, \ldots, y_K)$.

The key idea behind the anisotropic sparse grid is that we take advantage of the anisotropic weights, placing more collocation points in the dimensions that has a slower convergence in order to balance and minimize the global error [148]. How to obtain a sharp estimate of the importance or the weight of different dimensions is crucial to use the anisotropic sparse grid. One way is to derive a priori error estimate with the convergence rate, e.g., $\exp(-\ln(r_k)N), 1 \le k \le K$ in (1.38), as accurate as possible. However, deriving an analytical estimation of the convergence rate for general problems is rather difficult. Alternatively, we may perform empirical estimation by fitting the convergence rate from the numerical evaluation for each dimension (see Figure 1.5), and use the estimated convergence rates as α in (1.19) for the anisotropic sparse grid construction [148]. For the test of the efficiency of the anisotropic grid, we take the correlation length $L = 1/2$, $c = 5$ for the coefficient $a(x, \omega)$ in (1.59) and truncate it with nine random variables $y = (y_1, \ldots, y_9) \in \Gamma = [-\sqrt{3}, \sqrt{3}]^9$. Instead of the norm $||u - u_{approx}||_{L^\infty(\Gamma,X)}$, we use $||||u||_X - ||u_{approx}||_X||_{L^\infty(\Gamma)}$ to reduce the evaluation cost.

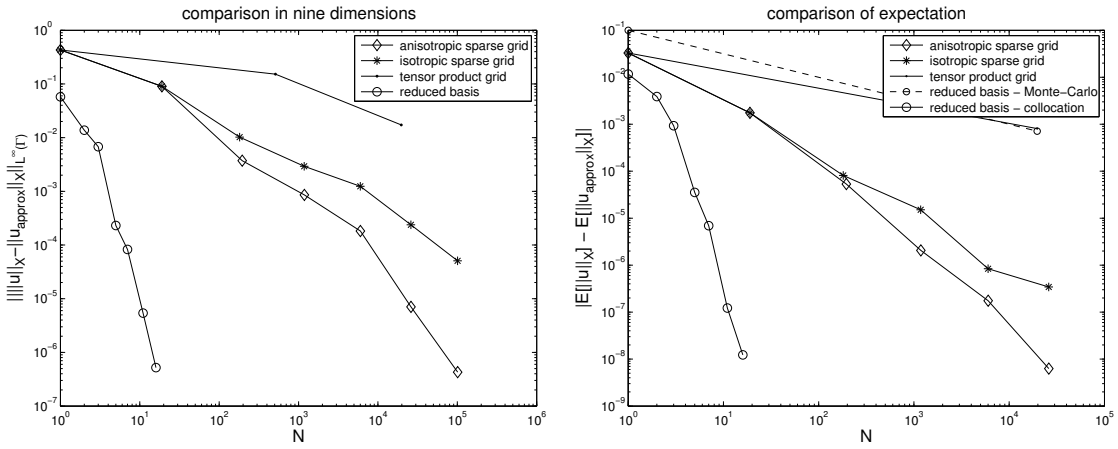

Figure 1.6: Comparison for the convergence rates of $||||u||_X - ||u_{approx}||_X||_{L^\infty(\Gamma)}$ (left) and the expectation $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$ (right) between the true and the approximate solutions in 9D.

We use the isotropic sparse grid and the anisotropic sparse grid at the interpolation level $q - 9 = 1, 2, 3, 4, 5, 6$ for the stochastic collocation approximation in (1.15) and (1.19), and choose the training samples as the collocation nodes in the sparse grid at the deepest interpolation level $q - 9 = 6$ ($100897 \approx 10^5$ nodes) for the reduced basis approximation. From Figure 1.6 we can see that the reduced basis approximation converges much faster than the stochastic collocation approximation in both $L^\infty(\Gamma)$ norm and the expectation norm. The offline computational cost of the reduced basis approximation for small-scale problems $h = 1/8, 1/16, 1/32$ is larger than that of the stochastic collocation approximation,

time t(s) | size h	1/8	1/16	1/32	1/64	1/128
t_{RB}(9D,$N_t = 10^3$)	85(0.0007)	91(0.001)	93(0.002)	121(0.01)	235(0.04)
t_{RB}(9D,$N_t \approx 10^5$)	8577(0.0008)	8582(0.001)	8585(0.002)	8610(0.01)	8722(0.04)
t_{SC}(9D,$N_{SC} \approx 10^5$)	154(0.13)	305(0.13)	4804(0.13)	23401(0.13)	101795(0.13)

Table 1.5: 9D offline (online in brackets) computational costs measured in CPU time by the reduced basis (RB) and the stochastic collocation (SC) methods achieving the same accuracy.

while for large-scale problems $h = 1/64, 1/128$ this becomes rather opposite; see Table 1.5. Besides, we also use 10^3 randomly generated training samples for the reduced basis approximation, and we still obtain the high accuracy in both $L^\infty(\Gamma)$ norm and the expectation norm because the solution is very smooth in the parameter space. We can see from Table 1.5 that the computational cost with 10^3 samples is far less than that of the sparse grid stochastic collocation approximation for both the offline construction and the online evaluation. In fact, the online construction of the reduced basis approximation stays the same as dominated by the number of reduced basis N_{rb} as $O(N_{rb}^3 + KN_{rb}^2 + K^2N_{rb}^2)$, while the online cost for the stochastic collocation approximation grows with the number of collocation points in an approximately linear way $O(N_{sc})(10^5/7^5 \approx 0.13/0.02)$. Figure 1.6 also brings us to the fact that the anisotropic sparse grid is more efficient than the isotropic sparse grid for anisotropic problems. Meanwhile, we can see that the stochastic collocation approximation based on tensor product grid starts to converge slower than $N^{-1/2}$, which is the typical convergence rate of the Monte-Carlo method.

1.6.3 Numerical experiments for higher dimensional problems

In this last numerical experiment of this chapter, we deal with high-dimensional stochastic problems, pushing the number of dimensions from 9 to 21, and from 51 up to 101, and comparing the performance of the reduced basis approximation and the stochastic collocation approximation. Note that in high dimensions $K = 101$, it is prohibitive to use the stochastic collocation method with tensor product grid (since we would need $3^{101} \approx 1.5 \times 10^{48}$ collocation points in total with 3 collocation points in each dimension), we use instead sparse grid of the anisotropic type to reduce the computational cost. The correlation length is $L = 1/128$, which enables us to consider an anisotropic problem but with the eigenvalues decaying very slowly ($\lambda_1 = 0.0138, \lambda_{50} = 0.0095$). The constant in (1.59) is chosen as $c = 20$ to guarantee that the stochastic problem is well posed with coercive elliptic operator. For the reduced basis approximation, we use 1000 samples randomly selected in $\Gamma = [-\sqrt{3}, \sqrt{3}]^K, K = 9, 21, 51, 101$ thanks to the rather smooth property of the solution in the parameter space, and for the stochastic collocation approximation, we construct adaptively an anisotropic sparse grid with $10^1, 10^2, 10^3, 10^4, 10^5, 10^6$ collocation nodes in an hierarchical way governed by the hierarchical surpluses [110]. To evaluate the error $||(||u||_X - ||u_{approx}||_X)||_{L^\infty(\Gamma)}$, we randomly select 100 samples in Γ. For the computation of the expectation as well as the error $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$, we apply the reduced basis - collocation approximation with 10^5 collocation nodes constructed from the anisotropic grid. The error $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$ is evaluated as a posteriori error by taking the best stochastic collocation approximation as the true or reference value.

The results for the high-dimensional stochastic problems are displayed in Figure 1.7, from which we can observe an exponential decay rate for both the $L^\infty(\Gamma)$ error and the expectation error by the reduced basis approximation, which is much larger than that of the stochastic collocation approximation. As the dimension increases from 9 to 101, the convergence rate decreases very fast for both the reduced basis approximation and the stochastic collocation approximation. As for the computational cost of the reduced basis method, it takes $86(K = 9), 424(K = 21), 2479(K = 51), 8986(K = 101)$ CPU seconds, respectively, for the offline construction with the mesh size $h = 1/8$, growing as $t_{RB} \propto O(K^2)$,

which verify the formula in Table 1.1 by Algorithm 2. In contrast, it would take $t_{SC} \propto O(K^w)$ where $w = q - K = 0, 1, 2, \ldots$ is the interpolation level of the isotropic Smolyak formula (1.15), which prevents large w for high-dimensional problems. We remark that although our numerical results are very promising for the reduced basis approximation, the size of the samples in the training set $\#\Xi_{train} = 1000$ and the testing set $\#\Xi_{test} = 100$ is rather small for the high-dimensional problems, which may bring insufficiency as for the approximation elsewhere. In order to increase the accuracy of the reduced basis approximation, we may construct the training set adaptively by replacing it with new set once the reduced basis approximation is good enough in the current one; see [212]. We also remark that the cost of the offline construction grows linearly with respect to the cardinality of the training set $t_{RB} \propto N_t$, as seen in Table 1.1.

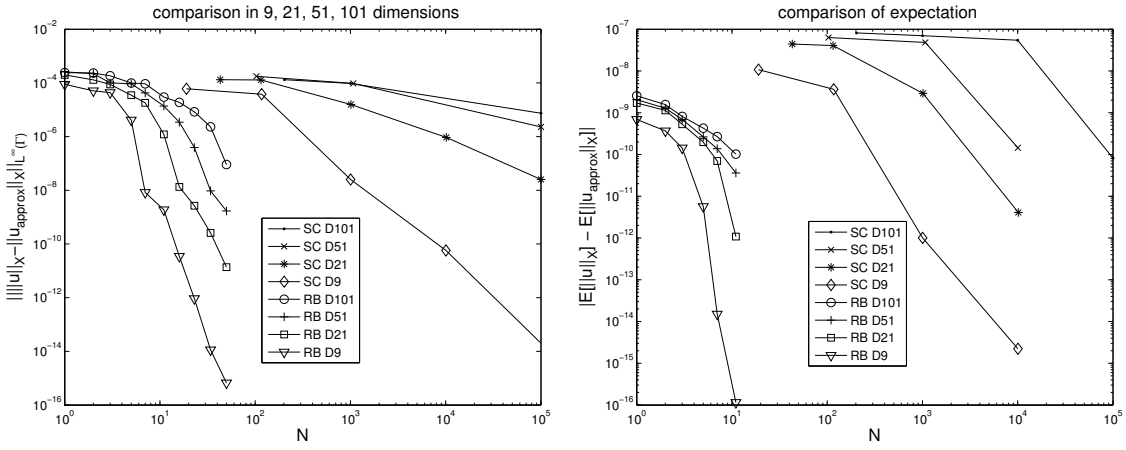

Figure 1.7: Comparison for the convergence rates of $||||u||_X - ||u_{approx}||_X||_{L^\infty(\Gamma)}$ (left) and $|\mathbb{E}[||u||_X] - \mathbb{E}[||u_{approx}||_X]|$ (right) between the anisotropic sparse grid stochastic collocation (SC) and the reduced basis (RB) methods in high dimensions 9D, 21D, 51D and 101D.

1.7 Summary

In this chapter, we carried out a detailed comparison of computational costs and accuracy between the reduced basis method and the stochastic collocation method for linear stochastic elliptic problems. The reduced basis method adopts Galerkin projection on the reduced basis space constructed from a greedy algorithm governed by a posteriori error estimate. It takes advantage of the affine structure of the stochastic problem to decompose the computation into an offline more expensive procedure and an online quite inexpensive procedure. The stochastic collocation method, on the other hand, follows essentially the Lagrange interpolation on the collocation nodes, which are taken as quadratures abscissas in order to achieve high order interpolation as well as integration for statistical computation.

The reduced basis method achieves an exponential convergence rate for smooth problems regardless of dimensions in our test case. The stochastic collocation method also exhibits an exponential convergence rate in the low-dimensional cases, though with a slower rate than the one featured by the reduced basis method; in contrast, in the multivariate case, especially for high-dimensional problems, it only achieves algebraic convergence rate. The computation of the stochastic collocation method takes less effort than the one needed by the reduced basis method in small-scale and low-dimensional problems, while it grows much faster than the reduced basis method in large-scale and high-dimensional problems, resulting in much heavier computational effort than the latter one. Note that the comparison depends essentially on the regularity of the stochastic solution, the dimension of the parameter space as well as the complexity of solving the underlying deterministic system, so that we presume that

similar comparison results hold reasonably beyond the case of linear stochastic elliptic problems considered here.

We succeeded in applying the reduced basis method and the anisotropic sparse grid stochastic collocation method in high-dimensional problems up to the order of (100). Nevertheless, the application is admittedly insufficient since the number of samples and collocation nodes is rather small. More advanced techniques such as sensitivity analysis and adaptive construction [88, 100] for both methods are being actively developed from the research community, more specifically to deal with high-dimensional stochastic systems. Moreover, the comparison has only been carried out for problems with solutions which depend smoothly on the parameters. As for non-smooth or low regularity stochastic problems, we expect that the reduced basis method, by taking advantage of solving a reduced problem (1.22) with the same mathematical structure as the original problem (46), can avoid Gibbs phenomenon as encountered by stochastic collocation method built upon dictionary basis (here Lagrange basis function), thus gaining further benefit on convergence; see also further examples in chapter 4 and in [41]. More research focusing on both theoretical and computational aspects is still needed when considering the reduced basis method and the stochastic collocation method, as well as their efficient combination, for solving more general problems, e.g., nonlinear, multiscale and multiphysics problems that feature low regularity and high dimensionality in the stochastic space.

2 A weighted reduced basis method for arbitrary probability measures

The results of the comparison between the reduced basis method and the stochastic collocation method in chapter 1 demonstrate that the former method is more efficient and cheaper than the latter for solving large-scale uncertainty quantification problems. However, our comparison was carried out only for uniformly distributed random variables. To our knowledge, the reduced basis method is currently used only for stochastic problems with uniformly distributed random inputs or parameter space with Lebesgue measure [24, 50]. In order to deal with more general uncertainties problems with arbitrary probability measures, we propose and analyze a extended version of the reduced basis method and name it the “weighted reduced basis method."

The basic idea of the weighted reduced basis method is to suitably assign a larger weight to those samples that are more important or have a higher probability to occur according to either the probability distribution function or some other available weight function depending on the specific application at hand. The benefit is to lighten the reduced space construction using a smaller number of bases without affecting the numerical accuracy. This idea is inspired by the generalization of polynomial chaos [208], where different polynomial bases representing the stochastic solution are chosen according to the probability density function of the input random variables, leading to fewer bases with the weights of the orthogonal polynomials exactly the same as the probability density function.

A priori convergence analysis for reduced basis method by greedy algorithm has been carried out in previous works [131, 30, 20, 115] under various assumptions. More specifically, the exponential convergence rate for a single-parameter elliptic PDE was obtained in [131] by exploring an eigenvalue problem; the algebraic or exponential convergence rate for greedy algorithm in multidimensional problem was achieved implicitly depending on the convergence rate of Kolmogorov N-width in [30] and improved in [20]; an exponential convergence rate was also recently obtained in [115] through direct expansion of the solution on a series of invertible elliptic operators. Different from these work, in this chapter we carry out a priori convergence analysis of the weighted reduced basis method based on Fourier analysis or constructive spectral approximation for analytic functions, as used in [8] for convergence analysis of the stochastic collocation method, which results in a direct a priori convergence rate for both single and multidimensional problems, and can be straightforwardly generalized to more general stochastic models once the stochastic regularity of the solution is obtained.

This chapter is organized as follows. Section 2.1 is devoted to the development of the weighted reduced

---

Reference for this chapter:

P. Chen, A. Quarteroni, and G. Rozza. A weighted reduced basis method for elliptic partial differential equation with random input data. SIAM Journal on Numerical Analysis, 51(6):3163–3185, 2013.

basis method, which is followed by regularity analysis and a priori convergence analysis in section 2.2. Numerical examples for both the one-dimensional problem and the multidimensional problem are presented as verification of the efficiency and convergence properties in section 2.3. Some brief remarks about further research perspectives are provided in the last section 2.4.

2.1 A weighted reduced basis method

We employ the linear and coercive stochastic elliptic PDE introduced in the preliminary chapter as the benchmark model to develop the weighted reduced basis method in this section. The basic idea of the weighted method is to assign different weights in the construction of reduced basis space at different values of parameter $y \in \Gamma$ according to a prescribed weight function $w(y)$. The motivation is that when the parameter y has distinctive weights $w(y)$ at different values $y \in \Gamma$, e.g., stochastic problems with random inputs obeying probability distribution far from uniform type, the weighted approach can considerably attenuate the computational effort for large scale computational problems. The weighted reduced basis method consists of the same elements, namely greedy algorithm, a posteriori error estimate and offline-online decomposition, as presented in chapter 1, section 1.3. In this chapter, we only highlight the new weighted scheme.

Let X be a high-fidelity approximation space of $H_0^1(D)$, equipped with the norm $||v||_X = \sqrt{A(v,v;\bar{y})}$ $\forall\, v(y) \in H_0^1(D)$ at some reference value $\bar{y} \in \Gamma$. Let X_w be a weighted approximation space with norm

$$||v(y)||_{X_w} = w(y)||v(y)||_X \quad \forall\, v \in X, \forall\, y \in \Gamma, \tag{2.1}$$

where $w : \Gamma \to \mathbb{R}_+$ is a weight function taking positive real values. Note that both X and X_w are equivalent to $H_0^1(D)$. The weighted greedy algorithm essentially deals with the $L^\infty(\Gamma; X_w)$ optimization problem in a greedy way [178], seeking a new parameter $y^N \in \Gamma$ such that

$$y^N = \arg\sup_{y\in\Gamma} ||u(y) - u_N(y)||_{X_w}, \tag{2.2}$$

where we recall that u_N is the reduced basis approximation of the solution u. By solving the infinite dimensional problem (2.2) we would locate the least matching point $y^N \in \Gamma$ in $||\cdot||_{X_w}$ norm. A computable (finite dimensional) greedy algorithm relies on (i) replacing the parameter domain Γ by a finite training set $\Xi_{train} \subset \Gamma$ with cardinality $|\Xi_{train}| = n_{train} < \infty$, and (ii) replacing the mismatching term $||u(y) - u_N(y)||_{X_w}$ by a cheap weighted a posteriori error bound $\triangle_N^w$ that should be as sharp as possible, i.e.,

$$c_N \triangle_N^w(y) \le ||u(y) - u_N(y)||_{X_w} \le C_N \triangle_N^w(y), \tag{2.3}$$

where C_N / c_N is expected to be close to 1. Let us look at an example of the weight function.

Example 2.1.1 *As shown in (1.27), the reduced basis approximation of the compliant output $s_N = F(u_N; y)$ has an error that scales quadratically with respect to the reduced basis approximation error of the solution [178], i.e.*

$$|s(y) - s_N(y)| \propto ||u(y) - u_N(y)||_X^2. \tag{2.4}$$

Therefore, in the evaluation of the expectation of s by formula (56), we may choose the weight as the following function of the probability density ρ:

$$w(y) = \sqrt{\rho(y)} \quad \forall\, y \in \Gamma, \tag{2.5}$$

so that the approximation error of the expectation can be bounded by

$$\mathbb{E}[s] - \mathbb{E}[s_N] = \int_\Gamma (s(y) - s_N(y))\rho(y)\,dy \le |\Gamma| \sup_{y\in\Gamma} |s(y) - s_N(y)|\rho(y) \le |\Gamma| \sup_{y\in\Gamma} (\triangle_N^w(y))^2, \tag{2.6}$$

where we assume that Γ *is bounded.*

Similar to (1.29), the reduced basis system can be written as

$$\sum_{m=1}^{N}\left(A_0(\zeta_m,\zeta_n)+\sum_{k=1}^{K}y_kA_k(\zeta_m,\zeta_n)\right)u_{Nm}(y)=(f_0,\zeta_n)+\sum_{k=1}^{K}(f_k,\zeta_n)y_k,\quad 1\le n\le N. \tag{2.7}$$

To solve (2.7) efficiently, we precompute and store $A_k(\zeta_m,\zeta_n), 0\le k\le K, 1\le m,n\le N_{max}$ and $(f_k,\zeta_n), 0\le k\le K, 1\le n\le N_{max}$ in the offline procedure. In the online procedure, we only need to assemble the stiffness matrix in (2.7) and solve the resulting $N\times N$ stiffness system with much less computational effort compared to solving a full $\mathcal{N}\times\mathcal{N}$ stiffness system. As for the computation of the error bound $\triangle_N^w(y)$, we need to evaluate $||\hat{e}(y)||_X$ at y chosen in the course of sampling procedure. The residual can be expanded as

$$R(v;y)=F(v;y)-A(u_N,v;y)=\sum_{k=0}^{K}(f_k,v)y_k-\sum_{n=1}^{N}u_{Nn}\left(\sum_{k=0}^{K}A_k(\zeta_n,v)y_k\right),\ \text{where } y_0=1. \tag{2.8}$$

Set $(\mathscr{C}_k,v)_X=(f_k,v)$ and $(\mathscr{L}_n^k,v)_X=-A_k(\zeta_n,v),\forall v\in X, 1\le n\le N, 0\le k\le K$, where $\mathscr{C}_k$ and $\mathscr{L}_n^k$ are the representatives of f_k and A_k^n (defined as $A_k^n(v)=-A_k(\zeta_n,v),\forall v\in X$) in X, respectively, whose existence is secured by the Riesz representation theorem. By recalling $(\hat{e}(y),v)_X=R(v;y)$, we obtain

$$\begin{aligned}||\hat{e}(y)||_X^2=&\sum_{k=0}^{K}y_k\left(\sum_{k'=0}^{K}y_{k'}(\mathscr{C}_k,\mathscr{C}_{k'})_X\right)\\&+\sum_{k=0}^{K}\sum_{n=1}^{N}y_ku_{Nn}(y)\left(\sum_{k'=0}^{K}y_{k'}2(\mathscr{C}_{k'},\mathscr{L}_n^k)_X+\sum_{k'=0}^{K}\sum_{n'=1}^{N}y_{k'}u_{Nn'}(y)(\mathscr{L}_n^k,\mathscr{L}_{n'}^{k'})_X\right).\end{aligned} \tag{2.9}$$

Therefore, we can compute and store $(\mathscr{C}_k,\mathscr{C}_{k'})_X,(\mathscr{C}_{k'},\mathscr{L}_n^k)_X,(\mathscr{L}_n^k,\mathscr{L}_{n'}^{k'})_X, 1\le n,n'\le N_{max}, 0\le k,k'\le K$, in the offline procedure and evaluate $||\hat{e}(y)||_X$ in the online procedure by assembling (2.9) with $O((K+1)^2N^2)$ scalar products, which is far more efficient provided that $O((K+1)^2N^2)\ll\mathcal{N}$.

2.2 Regularity and a priori error estimates

We work in the full space, which we still denote as X for ease of notation, rather than in the high fidelity discretization space (e.g., finite element space) for proving regularity and a priori error estimates for the weighted reduced basis method; the regularity with respect to random variables $y\in\Gamma$ and convergence results of the weighted reduced basis approximation hold the same in the discretization space.

2.2.1 Regularity results

Lemma 2.2.1 *Under Assumption 0.3, the solution to problem (46) satisfies* $u\in C^0(\Gamma;H_0^1(D))$. *Moreover, if* u *and* $\tilde{u}$ *are two weak solutions of problem (46) associated with data* a,f *and* $\tilde{a},\tilde{f}$*, respectively, we have the stability estimate*

$$||u-\tilde{u}||_{C^0(\Gamma;H_0^1(D))}\le\frac{C_P}{a_{min}}||f-\tilde{f}||_{C^0(\Gamma;L^2(D))}+\frac{C_P}{a_{min}^2}||\tilde{f}||_{C^0(\Gamma;L^2(D))}||a-\tilde{a}||_{C^0(\Gamma;L^\infty(D))} \tag{2.10}$$

Proof We rewrite (46) explicitly as

$$\forall y\in\Gamma,\int_D a(x,y)\nabla u(x,y)\cdot\nabla v(x)dx=\int_D f(x,y)v(x)dx\quad\forall v\in H_0^1(D). \tag{2.11}$$

A similar problem holds for $\tilde{f}$ and $\tilde{a}$. By subtraction we obtain the difference equation:

$$\int_D a\nabla(u-\tilde{u})\cdot\nabla v dx = \int_D (f-\tilde{f})v dx + \int_D (\tilde{a}-a)\nabla\tilde{u}\cdot\nabla v dx. \tag{2.12}$$

By taking $v = u-\tilde{u}$, applying the Cauchy–Schwarz and Poincaré inequalities, and using Assumption 0.2 we have

$$a_{min}||u-\tilde{u}||^2_{H_0^1(D)} \le C_P||f-\tilde{f}||_{L^2(D)}||u-\tilde{u}||_{H_0^1(D)} + ||\tilde{u}||_{H_0^1(D)}||u-\tilde{u}||_{H_0^1(D)}||a-\tilde{a}||_{L^\infty(D)}, \tag{2.13}$$

so that the following stability estimate holds for $\forall\, y\in\Gamma$ by the fact $||\tilde{u}||_{H_0^1(D)} \le (C_P/a_{min})||\tilde{f}||_{L^2(D)}$ (due to the Lax–Milgram theorem and Assumption 0.2 holding also for $\tilde{a}$):

$$||u(y)-\tilde{u}(y)||_{H_0^1(D)} \le \frac{C_P}{a_{min}}||f(y)-\tilde{f}(y)||_{L^2(D)} + \frac{C_P}{a_{min}^2}||\tilde{f}(y)||_{L^2(D)}||a(y)-\tilde{a}(y)||_{L^\infty(D)}. \tag{2.14}$$

Setting $\tilde{a}(y) = a(y+\delta y)$ and $\tilde{f}(y) = f(y+\delta y)$ such that $y+\delta y\in\Gamma$, we have by Assumption 0.3 that $\tilde{a}(y)\to a(y)$ in $L^\infty(D)$ and $\tilde{f}(y)\to f(y)$ in $L^2(D)$ so that $\tilde{u}(y) = u(y+\delta y)\to u(y)$ in $H_0^1(D)$ when $\delta y\to 0$. Therefore, the solution is continuous with respect to the parameter $y\in\Gamma$, i.e., $u\in C^0(\Gamma; H_0^1(D))$. □

A direct application of Lemma 2.2.1 leads to the following lemma for the existence of partial derivatives of the solution with respect to the parameter $y\in\Gamma$ as well as their bound in $H_0^1(D)$.

Lemma 2.2.2 *For any $y\in\Gamma$, there exists a unique $\partial_y^\nu u(y)$ in $H_0^1(D)$ provided that Assumption 0.3 are satisfied for any $y\in\Gamma$ and $\nu = (\nu_1,\dots,\nu_K)\in\Lambda$, where $\Lambda\subset\mathbb{N}^K$ is a multiple index set. Moreover, we have the following estimate:*

$$||\partial_y^\nu u(y)||_{H_0^1(D)} \le B(y)|\nu|!\eta^\nu + \frac{C_P}{a_{min}}|\nu|!\sum_{k:\nu_k\neq 0}\left(\eta^{\nu-e_k}||f_k||_{L^2(D)}\right), \tag{2.15}$$

where

$$B(y) = \frac{C_P}{a_{min}}||f(y)||_{L^2(D)}, \quad |\nu|! = (\nu_1+\cdots+\nu_K)!, \quad \eta^\nu = \prod_{k=1}^K \eta_k^{\nu_k}, \quad \eta_k = \frac{||a_k||_{L^\infty(D)}}{a_{min}}. \tag{2.16}$$

Proof We use an induction argument for the proof in the following few steps.

Step 1. First, when $|\nu| = 0$, there exists a unique solution $u\in H_0^1(D)$ of problem (46) for every $y\in\Gamma$ thanks to the Lax–Milgram theorem. Moreover, the estimate

$$||\partial_y^\nu u(y)||_{H_0^1(D)} = ||u(y)||_{H_0^1(D)} \le \frac{C_P}{a_{min}}||f(y)||_{L^2(D)} = B(y) \tag{2.17}$$

holds, which verifies (2.15) for $|\nu| = 0$.

Step 2. For $|\nu|\ge 1$, we are about to prove that there exists a unique function $\partial_y^\nu u(y)$ satisfying the following general recursive equation (write $a(y)$ in short for $a(x,y)$, etc.),

$$\int_D a(y)\nabla\partial_y^\nu u(y)\cdot\nabla v = -\sum_{k:\nu_k\neq 0}\nu_k\int_D a_k\nabla\partial_y^{\nu-e_k}u(y)\cdot\nabla v + \sum_{k:\nu=e_k}\int_D f_k v \quad \forall\, v\in H_0^1(D), \tag{2.18}$$

where e_k is a K dimensional vector with the k-th element as 1 and all the other elements as 0. To see this, let us first show that for $|\nu| = 1$, i.e., $\nu = e_k, 1\le k\le K$, there exists a unique solution $\partial_y^\nu u(y)$ to (2.18).

We take the perturbation $\tilde{a}(y) = a(y - he_k)$, $\tilde{f}(y) = f(y - he_k)$, and $\tilde{u}(y) = u(y - he_k)$ in (2.12) and set $D_h^k u = (u(y) - u(y - he_k))/h$; then (2.12) becomes

$$\int_D a(y)\nabla D_h^k u(y)\nabla v = \int_D f_k v - \int_D a_k \nabla u(y - he_k)\cdot\nabla v \quad \forall v \in H_0^1(D), \tag{2.19}$$

which results in a unique solution $D_h^k u(y) \in H_0^1(D)$ by the Lax–Milgram theorem. Taking the limit $h \to 0$, we have by the continuity result in Lemma 2.2.1 that $u(y - he_k) \to u(y)$ so that $D_h^k u(y) \to \partial_y^\nu u(y)$ exists. Therefore, $\partial_y^\nu u(y)$ is a unique solution of (2.18) for $\nu = e_k, 1 \le k \le K$. By induction we suppose that there exists a unique function $\partial_y^{\tilde{\nu}} u(y)$ satisfying (2.18) for $|\tilde{\nu}| = |\nu| - 1$, i.e., $\tilde{\nu} = \nu - e_j$ for some $j = 1, \ldots, K$; then we claim that there exists a unique function $\partial_y^\nu u(y)$ satisfying (2.18) for each ν such that $|\nu| > 1$. By the same argument of perturbation and continuity property, we are able to take the derivative of (2.18) with respect to y_j, where ν is replaced by $\tilde{\nu} = \nu - e_j$ in (2.18), yielding

$$\begin{aligned}\int_D a(y)\nabla\partial_y^\nu u(y)\cdot\nabla v + \int_D a_j \nabla\partial_y^{\nu-e_j} u(y)\cdot\nabla v = &-\sum_{k\neq j:\nu_k\neq 0} \nu_k \int_D a_k \nabla\partial_y^{\nu-e_k} u(y)\cdot\nabla v \\ &-(\nu_j - 1)\int_D a_j\nabla\partial_y^{\nu-e_j} u(y)\cdot\nabla v + \sum_{k:\nu=e_k}\int_D f_k v,\end{aligned} \tag{2.20}$$

which can be simplified by summing up the same terms to end up with (2.18). By the Lax–Milgram theorem, we have that there exists a unique solution $\partial_y^\nu u(y) \in H_0^1(D)$ to (2.18).

Step 3. We are going to show that the estimate (2.15) holds for $|\nu| \ge 1$ in this step. Upon replacing v by $\partial_y^\nu u(y)$ in (2.18), we have by Assumption 0.2 as well as the Cauchy–Schwarz and Poincaré inequalities the following estimate:

$$||\partial_y^\nu u(y)||_{H_0^1(D)} \le \sum_{k:\nu_k\neq 0} \nu_k \eta_k ||\partial_y^{\nu-e_k} u(y)||_{H_0^1(D)} + \frac{C_P}{a_{min}} \sum_{k:\nu=e_k} ||f_k||_{L^2(D)}. \tag{2.21}$$

Observe that when $|\nu| = 1$, i.e., $\nu = e_k, 1 \le k \le K$, estimate (2.21) becomes

$$||\partial_y^\nu u(y)||_{H_0^1(D)} = ||\partial_{y_k} u(y)||_{H_0^1(D)} \le B(y)\eta_k + \frac{C_P}{a_{\min}}||f_k||_{L^2(D)}, \tag{2.22}$$

which is the same as in (2.15). If $|\nu| > 1$, estimate (2.21) becomes

$$||\partial_y^\nu u(y)||_{H_0^1(D)} \le \sum_{k:\nu_k\neq 0} \nu_k\eta_k ||\partial_y^{\nu-e_k} u(y)||_{H_0^1(D)}. \tag{2.23}$$

Suppose estimate (2.15) holds for any $|\tilde{\nu}| < |\nu|$ with $|\nu| > 1$; then we have

$$\begin{aligned}||\partial_y^\nu u(y)||_{H_0^1(D)} &\le \sum_{j:\nu_j\neq 0} \nu_j\eta_j ||\partial_y^{\nu-e_j} u(y)||_{H_0^1(D)} \\ &\le \sum_{j:\nu_j\neq 0} \nu_j\eta_j \left(B(y)(|\nu|-1)!\eta^{\nu-e_j} + \frac{C_P}{a_{\min}}(|\nu|-1)! \sum_{k:\nu_k\neq 0} \left(\eta^{\nu-e_j-e_k}||f_k||_{L^2(D)}\right)\right) \\ &= B(y)\left(\sum_{j:\nu_j\neq 0}\nu_j\right)(|\nu|-1)!\eta^\nu + \frac{C_P}{a_{\min}}\left(\sum_{j:\nu_j\neq 0}\nu_j\right)(|\nu|-1)!\sum_{k:\nu_k\neq 0}\left(\eta^{\nu-e_k}||f_k||_{L^2(D)}\right) \\ &= B(y)|\nu|!\eta^\nu + \frac{C_P}{a_{min}}|\nu|!\sum_{k:\nu_k\neq 0}\left(\eta^{\nu-e_k}||f_k||_{L^2(D)}\right) \equiv C_{a,f}(y)|\nu|!\eta^\nu,\end{aligned} \tag{2.24}$$

where

$$C_{a,f}(y) = B(y) + C_P \sum_{k:\nu_k\neq 0, ||a_k||_{L^\infty(D)}\neq 0} \frac{||f_k||_{L^2(D)}}{||a_k||_{L^\infty(D)}}, \tag{2.25}$$

so that estimate (2.15) also holds for ν with $|\nu| > 1$. □

An analytic extension of the solution u in a certain region Σ such that $\Gamma \subset \Sigma$ is a consequence of the regularity result in Lemma 2.2.2 provided conditions are suitable, as stated in the following lemma.

Lemma 2.2.3 *Holding all the assumptions in Lemma 2.2.2, and defining*

$$\Sigma = \left\{ z \in \mathbb{C}^K : \exists\, y \in \Gamma \text{ s.t. } |(\eta \cdot |z-y|)| = \sum_{k=1}^{K} \eta_k |z_k - y_k| < 1 \right\}, \tag{2.26}$$

we can find an analytic extension of the stochastic solution u in the complex region Σ and we define $\Sigma(\Gamma;\tau) := \{z \in \Sigma : dist(z,\Gamma) \leq \tau\} \subset \Sigma$ for the largest possible vector $\tau = (\tau_1, \ldots, \tau_K)$.

Proof By the Taylor expansion of $u(z)$ about $y \in \Gamma$ in the complex domain we obtain

$$u(z) = \sum_{\nu} \frac{\partial_y^\nu u(y)}{\nu!} (z-y)^\nu \tag{2.27}$$

with $\nu! = \nu_1! \cdots \nu_K!$. Thanks to the regularity result in Lemma 2.2.2, we obtain

$$\begin{aligned} \left|\left| \sum_{\nu} \frac{\partial_y^\nu u(y)}{\nu!} (z-y)^\nu \right|\right|_{H_0^1(D)} &\leq \sum_{\nu} \frac{|z-y|^\nu}{\nu!} ||\partial_y^\nu u(y)||_{H_0^1(D)} \\ &\leq C_{a,f}(y) \sum_{n\geq 0:|\nu|=n} \frac{|\nu|!}{\nu!} \left(\eta \cdot |z-y|\right)^\nu \\ &= C_{a,f}(y) \sum_{n\geq 0} \left(\sum_{k=1}^{K} \eta_k |z_k - y_k| \right)^n \\ &= \frac{C_{a,f}(y)}{1 - \sum_{k=1}^{K} \eta_k |z_k - y_k|}, \end{aligned} \tag{2.28}$$

where the second inequality is due to Lemma 2.2.2 and the first equality follows from the generalized Newton binomial formula. In the complex region defined in (2.26), we obtain that the function $u(z)$ admits a Taylor expansion around $y \in \Gamma$ so that the solution u can be analytically extended to the complex region (2.26). □

2.2.2 A priori convergence analysis

To prove the exponential convergence of the weighted reduced basis method for problem (46) for the case of one random variable, i.e., $\Gamma \subset \mathbb{R}$, we bound the error by another type of constructive spectral approximation or more specifically, extension of the Chebyshev polynomial approximation for analytic functions (see [61, Chapter 7]). The idea has also been used in the proof of the exponential convergence property of the stochastic collocation method [8]. Based on this idea we also obtain the a priori error estimate of the reduced basis approximation for multidimensional problems, e.g., $\Gamma \subset \mathbb{R}^K, K > 1$.

We define the weighted space $C_w^0(\Gamma; X)$ equipped with the following norm

$$||v||_{C_w^0(\Gamma;X)} = \max_{y\in\Gamma}(w(y)||v(y)||_X) \tag{2.29}$$

for any positive continuous bounded weight function $w:\Gamma\to\mathbb{R}_+$. Because of Assumption 0.3, the linear coefficient a and forcing term f satisfy $a\in C^0(\Gamma;L^\infty(D))$ and $f\in C^0_w(\Gamma;L^2(D))$.

Theorem 2.2.4 *Under Assumption 0.3 with $\Gamma\subset\mathbb{R}$, the error between the reduced basis solution $u_N=P_Nu$ (recall that $P_N:u\to u_N$ represents the Galerkin projection operator) and the "truth" solution u of problem (46) enjoys the exponential convergence*

$$||u-P_Nu||_{C^0_w(\Gamma;X)}\le C^w e^{-rN}\max_{z\in\Sigma(\Gamma;\tau)}||u(z)||_X \tag{2.30}$$

where the constant C^w depends on the weight w and is independent of N, and the rate r is defined as

$$1<r=log\left(\frac{2\tau}{|\Gamma|}+\sqrt{1+\frac{4\tau^2}{|\Gamma|^2}}\right), \tag{2.31}$$

where τ is defined in Lemma 2.2.3.

Remark 2.2.1 *The convergence rate stated above does not depend on the specific problem (39). In fact, as long as $u=u(y)$ is an analytic function, the exponential convergence rate (2.30) holds for reduced basis approximation as demonstrated in the proof of this theorem later. The same a priori convergence property can therefore be established for problems other than the elliptic problem (39) under linear or affine assumptions (43) as studied in [131, 115].*

Proof First, we note that the results obtained in the above lemmas in $H^1_0(D)$ norm are still valid in the equivalent X norm. Given a bounded and continuous one-dimensional domain $\Gamma\subset\mathbb{R}$, we introduce the change of variables $y(t)=\bar{y}+\frac{|\Gamma|}{2}t$ with $t\in[-1,1]$ and $\bar{y}$ the center of domain Γ, so that $y:[-1,1]\to\Gamma$ is bijective. Let the solution of problem (46) be set as $\hat{u}(t)=u(y(t))$ for $t\in[-1,1]$; then we have that $\hat{u}:[-1,1]\to X$ can be analytically extended to $\Sigma([-1,1],2\tau/|\Gamma|)$ by Lemma 2.2.3. Consequently, there exists a spectral expansion of $\hat{u}$ on the standard Chebyshev polynomials $c_k:[-1,1]\to\mathbb{R}$ and $|c_n|\le 1, n=0,1,\ldots$, in the form

$$\hat{u}(t)=\frac{u_0}{2}+\sum_{n=1}^{\infty}\hat{u}_nc_n(t). \tag{2.32}$$

The n-th Chebyshev coefficient fulfils [61]

$$\hat{u}_n=\frac{1}{\pi}\int_{-\pi}^{\pi}\hat{u}(\cos(t))\cos(nt)dt,\quad ||\hat{u}_n||_X\le 2\varrho^{-n}\max_{z\in D_\varrho}||\hat{u}(z)||_X,\quad n=0,1,\ldots, \tag{2.33}$$

where the elliptic disc D_ϱ is bounded by the ellipse E_ϱ with *foci* ±1 and the sum of the half-axes $\varrho=2\tau/|\Gamma|+\sqrt{1+(4\tau^2/|\Gamma|^2)}$. Define the N-th order Chebyshev polynomial approximation of $\hat{u}$ as the truncation of (2.32) up to N terms, written as

$$\Pi_N\hat{u}=\frac{u_0}{2}+\sum_{n=1}^{N}\hat{u}_nc_n(t); \tag{2.34}$$

then the truncation error is bounded by using $|c_n|\le 1, n=N+1,\ldots$, and (2.33) as follows:

$$||\hat{u}-\Pi_N\hat{u}||_{C^0([-1,1];X)}\le\sum_{n\ge N+1}||\hat{u}_n||_X\le\frac{2}{\varrho-1}e^{-\log(\varrho)N}\max_{z\in D_\varrho}||\hat{u}(z)||_X, \tag{2.35}$$

Therefore, by the identity $\hat{u}(t)=u(y(t)), t\in[-1,1]$, we have

$$||u-\Pi_Nu||_{C^0(\Gamma;X)}\le\frac{2}{\varrho-1}e^{-rN}\max_{z\in D_\varrho}||\hat{u}(z)||_X\le\frac{2}{\varrho-1}e^{-rN}\max_{z\in\Sigma(\Gamma;\tau)}||u(z)||_X, \tag{2.36}$$

where we define $r := \log(\varrho)$, as given in (2.31). It is left to prove that the reduced basis approximation error can be bounded by the above truncation error. In fact, for any function $v \in \mathscr{P}_N(\Gamma) \otimes X$, a tensor product of polynomials with total degree at most N and X, we have that $\mathscr{I}_N v = v$ [36, 8], where $\mathscr{I}_N$ is the Lagrange interpolation operator based on the interpolation points $y^n, n = 1, \ldots, N+1$; see [8]. We have the following estimate with the help of the Lagrange interpolation operator

$$\begin{aligned} ||u - P_{N+1}u||_X &\leq C_0 \inf_{v \in X_{N+1}} ||u - v||_X \\ &\leq C_0 ||u - \mathscr{I}_N u||_X \\ &\leq C_0 \inf_{v \in \mathscr{P}_N(\Gamma)\otimes X} (||u - v||_X + ||v - \mathscr{I}_N u||_X) \\ &= C_0 \inf_{v \in \mathscr{P}_N(\Gamma)\otimes X} (||u - v||_X + ||\mathscr{I}_N v - \mathscr{I}_N u)||_X) \\ &\leq (C_0 + C_1) \inf_{v \in \mathscr{P}_N(\Gamma)\otimes X} ||u - v||_X, \end{aligned} \tag{2.37}$$

where the first inequality is due to Cea's lemma [165] with constant $C_0 < \infty$ and the second due to the fact $\inf_{v \in X_{N+1}} ||u - v||_X \leq ||u - \mathscr{I}_N u||_X$; as for the last inequality, we have used the property that the Lagrange interpolation operator $\mathscr{I}_N$ is linear and $||\mathscr{I}_N v||_X \leq C_1 ||v||_X \forall\, v \in C^0(\Gamma, X)$ for a constant $C_1 < \infty$, see [8]. Moreover, because the Chebyshev polynomials $c_k \in \mathscr{P}_N([-1,1]), k = 0, 1, \ldots, N$, we have

$$\inf_{v \in \mathscr{P}_N(\Gamma)\otimes X} ||u - v||_X = \inf_{\hat{v} \in \mathscr{P}_N([-1,1])\otimes X} ||\hat{u} - \hat{v}||_X \leq ||\hat{u} - \Pi_N \hat{u}||_X = ||u - \Pi_N u||_X. \tag{2.38}$$

A combination of (2.36), (2.37), and (2.38) leads to the following bound for the reduced basis approximation error with $C = 2(C_0 + C_1)e^r/(\varrho - 1)$:

$$||u - P_N u||_X \leq C e^{-rN} \max_{z \in \Sigma(\Gamma;\tau)} ||u(z)||_X. \tag{2.39}$$

Since the reduced basis approximation $P_N u$ satisfies the linear system (2.7), which can be written in the compact form as

$$A(P_N u, v; y) = F(v; y) \quad \forall\, v \in X_N, \tag{2.40}$$

we obtain the same regularity for $P_N u$ as for the solution u to system (46) with respect to the parameter y. In particular, $P_N u \in C^0_w(\Gamma; X)$, so that $u - P_N u \in C^0_w(\Gamma; X)$. Multiplying both sides of (2.39) by the weight function w and taking the maximum value over the parameter domain Γ, we obtain the exponential convergence result (2.30) with the constant $C^w = C \max_{y \in \Gamma} w(y)$.

□

Remark 2.2.2 *The exponential convergence result (2.30) holds for the case of a single parameter in a bounded parameter domain $|\Gamma| < \infty$. Extension to a single parameter in the unbounded domain, e.g., a normal distributed random variable, requires that the data a and f feature a fast decrease at the parameter far away from the origin, and the constructive approximation by spectral expansion on Chebyshev polynomials (2.32) is replaced by the one on Hermite polynomials [8]. The proof follows the same procedure as for Theorem 2.2.4 and we omit it for simplicity.*

As for the reduced basis approximation in the multidimensional case, we have the following a priori error estimate:

Theorem 2.2.5 *Under Assumption 0.3 with $\Gamma \subset \mathbb{R}^K, K > 1$, the approximation error of the reduced basis*

solution can be bounded by

$$||u - P_N u||_{C_w^0(\Gamma;X)} \leq \max_{z\in\Sigma(\Gamma;\tau)} ||u(z)||_X \sum_{k=1}^{K} C_k^w e^{-r_k N_k}, \tag{2.41}$$

where the constants $C_k^w, 1 \leq k \leq K$, depend on the weight w and dimension k but is independent of the number of nodes in the k-th dimension N_k, $N = \prod_{k=1}^{K} N_k$, and the rate r_k is defined as

$$1 < r_k = log\left(\frac{2\tau_k}{|\Gamma_k|} + \sqrt{1 + \frac{4\tau_k^2}{|\Gamma_k|^2}}\right), 1 \leq k \leq K. \tag{2.42}$$

Proof Let us choose the training set as all the nodes of a tensor product grid, i.e., $\Xi_{train} := \{(y_1^{n_1}, \ldots, y_K^{n_K}), 1 \leq n_k \leq N_k, 1 \leq k \leq K\}$, for instance the Gauss quadrature nodes corresponding to the probability density function of the random vector y. We define the reduced basis space $X_N^k, 1 \leq k \leq K$, as a linear combination of the snapshots $u(y)$ at $y = (y_k^1, y_k^*), \ldots, (y_k^N, y_k^*)$, where $y_k^n \in \Gamma_k, 1 \leq n \leq N$, and y_k^* is any point in the rest of the $K-1$ dimensional domain denoted as Γ_k^*. Correspondingly, we define the Galerkin projection operator $P_N^k : X \to X_N^k, 1 \leq k \leq K$, such that $P_N^k u$ is the solution of the reduced problem (2.40) in X_N^k whenever u is the solution of the original problem (46) in X at any $y = (y_k, y_k^*) \in \Gamma_k \times \Gamma_k^*$. Let X_N be the reduced basis space spanned by the snapshots at all the $N = \prod_{k=1}^{K} N_k$ samples and $P_N : X \to X_N$ be the associated Galerkin projection operator; then we have for the solution $u \in X$ of problem (46) at any $y = (y_1, y_1^*) \in \Gamma_1 \times \Gamma_1^*$,

$$P_N u = P_N^1 \circ \cdots \circ P_N^K u, \tag{2.43}$$

the symbol $\circ$ being the composition of the projection operators. By triangular inequality, we have

$$||u - P_N u||_X \leq ||u - P_N^1 u||_X + ||P_N^1(u - P_N^2 \circ \cdots \circ P_N^K u)||_X, \tag{2.44}$$

where we can bound the first term as in (2.39) by

$$||u - P_N^1 u||_X \leq C_1 e^{-r_1 N_1} \max_{(z_1, z_1^*)\in\Sigma(\Gamma_1\times\Gamma_1^*;\tau)} ||u(z)||_X \leq C_1 e^{-r_1 N_1} \max_{z\in\Sigma(\Gamma;\tau)} ||u(z)||_X, \tag{2.45}$$

where z_1, z_1^* are the complex elements associated with y_1 and y_1^*; the constant C_1 has similar definition as C in (2.39) and r_k is defined in (2.42). As for the second term, thanks to the fact that $||P_N^1 v||_X \leq ||v||_X$ we have

$$||P_N^1(u - P_N^2 \circ \cdots \circ P_N^K u)||_X \leq ||u - P_N^2 \circ \cdots \circ P_N^K u||_X. \tag{2.46}$$

By iteration, we obtain the error bound

$$||u - P_N u||_X \leq \max_{z\in\Sigma(\Gamma;\tau)} ||u(z)||_X \sum_{k=1}^{K} C_k e^{-r_k N_k}, \tag{2.47}$$

which leads to the a priori error estimate (2.41) by multiplying by the weight function w on both sides and noting that $P_N u \in C_w^0(\Gamma; X)$, where the constants $C_k^w := C_k \max_{y\in\Gamma} w(y), 1 \leq k \leq K$. □

Remark 2.2.3 *In practice, the training set Ξ_{train} can be chosen in a more general way, e.g., by sampling according to the probability density function, and the cardinality of the reduced basis space X_N is much lower than $\prod_{k=1}^{K} N_k$ given in the theorem. In fact the error estimate obtained in this theorem is rather crude. An improved convergence rate $e^{-r' N^{\beta/(\beta+1)}}$ was achieved in [20] provided that the Kolmogorov N-width by the optimal N dimensional approximation decays as $e^{-r N^{\beta}}$ in a more general setting, e.g., if*

Γ *is not bounded. However, the Kolmogorov N-width is not available in general.*

A direct consequence of Theorems 2.2.4 and 2.2.5 for the convergence of quantity of interest and its statistical moments is summarized in the following corollary:

Corollary 2.2.6 *Suppose that the assumptions in Theorem 2.2.4 hold, we have*

$$||s(u) - s(P_N u)||_{C^0_w(\Gamma)} \leq ||s||_{X'} ||u - P_N u||_{C^0_w(\Gamma;X)}, \tag{2.48}$$

and for the k-th order statistical moment, where $k = 1, 2, \ldots$, *we have*

$$\begin{aligned} |\mathbb{E}[s^k(u)] - \mathbb{E}[s^k(P_N u)]| &\approx \left| \sum_{m=1}^{M} w(y^m)(s(u;y^m) - s(P_N;y^m)) \left(\sum_{l=0}^{k-1} s^{k-1-l}(u;y^m) s^l(P_N u;y^m) \right) \right| \\ &\leq M ||s(u) - s(P_N u)||_{C^0_w(\Gamma)} C^k_s, \end{aligned} \tag{2.49}$$

where C^k_s *is a constant depending on the output s and the statistical moment k with* $C^1_s = 1$.

2.3 Numerical examples

In this section, we present several numerical examples to illustrate the efficiency of the weighted reduced basis method compared to the classical reduced basis method and the stochastic collocation method. The output of interest is defined as the integral of the solution over the physical domain D

$$s(y) = \int_D u(x,y)\, dx. \tag{2.50}$$

We define the following two errors as criteria of different numerical methods:

$$||s - s_N||_{C^0_w(\Gamma)} \quad \text{and} \quad |\mathbb{E}[s] - \mathbb{E}[s_N]|, \tag{2.51}$$

where s_N is the approximated value of s obtained using N bases for (weighted) reduced basis method or N collocation points for the stochastic collocation method. In particular, we use the weight function in one dimension as the probability density function of the random variable obeying Beta(α, β) distribution with shape parameter α and β providing distinctive property of the weight, defined as

$$w(y;\alpha,\beta) = \frac{1}{2\mathrm{B}(\alpha,\beta)} (1+y)^{\alpha-1} (1-y)^{\beta-1} \quad y \in [-1,1], \tag{2.52}$$

where the beta function B is a normalization constant such that the total probability integrates to 1. In our numerical experiments, we use the Gauss-Jacobi quadrature formula to compute the expectation (2.51) with the solution at the abscissas evaluated by the reduced basis methods. As for the stochastic collocation method, we use the Gauss-Jacobi abscissas as the collocation points, which is more accurate than other choices, especially when the weight function is more concentrated. We specify the detailed setting of the weighted reduced basis method in the following subsections. The physical domain is a square $D = (-1,1)^2$ and homogeneous Dirichlet boundary conditions are prescribed on the entire boundary ∂D.

2.3.1 One-dimensional problem

We set the stochastic coefficient $a(x,\omega), x = (x_1, x_2) \in D$, in problem (39) (depending only on x_1) as

$$a(x,\omega) = \frac{1}{10}(1.1 + \sin(2\pi x_1)Y(\omega)) \tag{2.53}$$

with random variable $Y \sim \text{Beta}(\alpha,\beta)$ with $(\alpha,\beta) = (1,1)$, $(10,10)$ and $(100,100)$, respectively. We remark that when $(\alpha,\beta) = (1,1)$ the weighted reduced basis method becomes a reduced basis method with uniformly distributed random variable, which has been examined in [50]. The left of Figure 2.1 depicts the shape of weight at different locations. For simplicity, the forcing term is taken with a deterministic value $f = 1$. We use the same tolerance $\varepsilon = 1 \times 10^{-15}$ for three different weight functions to stop the greedy algorithm; $n_{train} = 1000$ samples are uniformly selected to construct the reduced basis space. Another 1000 samples are used to test the accuracy of different methods. The exponential convergence of the error $||s - s_N||_{C^0_w(\Gamma)}$ and its error bound in logarithmic scale for three different weight functions are displayed on the right side of Figure 2.1 for the weighted reduced basis method. The maximum number of bases $N_{max} = 16, 11, 6$ built at the training samples are visualized with selection order identified by the marker size on the left side of Figure 2.1; they are quite different for different weight functions. From the location and selecting order of the samples on the left of Figure 2.1, we can conclude that the weight function plays an important role in choosing the most representative bases.

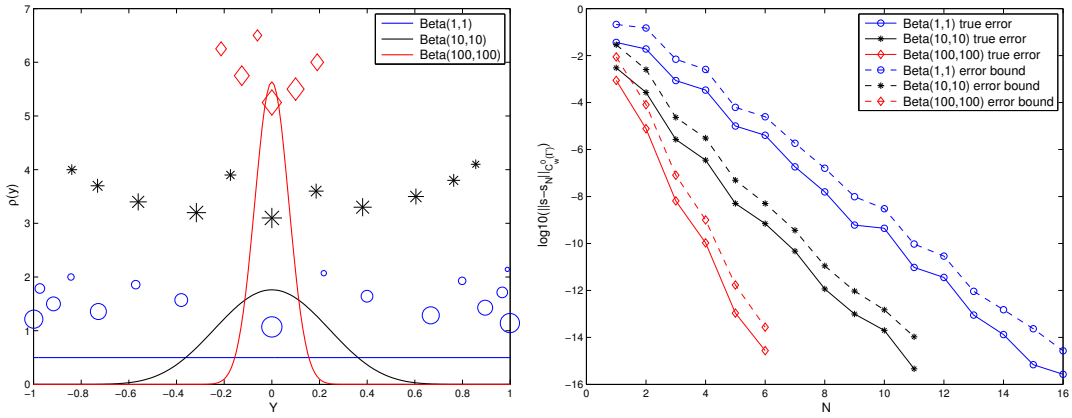

Figure 2.1: Left: probability density function of Beta(α,β) distribution with different α,β and samples selected by weighted reduced basis approximation; the bigger the size the earlier it has been selected. Right: convergence of the error $\log_{10}\left(||s - s_N||_{C^0_w(\Gamma)}\right)$ by the weighted reduced basis method.

In the comparison of the convergence property of the reduced basis method, the weighted reduced basis method as well as the stochastic collocation method, we select the weight function of Beta$(10,10)$ and compute the two errors defined in (2.51) with the results shown in Figure 2.2. It is evident that the weighted reduced basis method outperforms the reduced basis method in both norms, and these two methods are more accurate than the stochastic collocation method in the $||\cdot||_{C^0_w(\Gamma)}$ norm. As for the expectation, the weighted reduced basis method is the best and the reduced basis method does not beat the stochastic collocation method because it does not take the weight into account.

However, as demonstrated in [50], the computation of both reduced basis methods for the one-dimensional stochastic problem is more expensive than that of the stochastic collocation method because of the offline construction with a large number of training samples, especially for the problem requiring low computational effort in one deterministic solving. Similar numerical examples for some other weight functions are presented in the appendix for expository convenience.

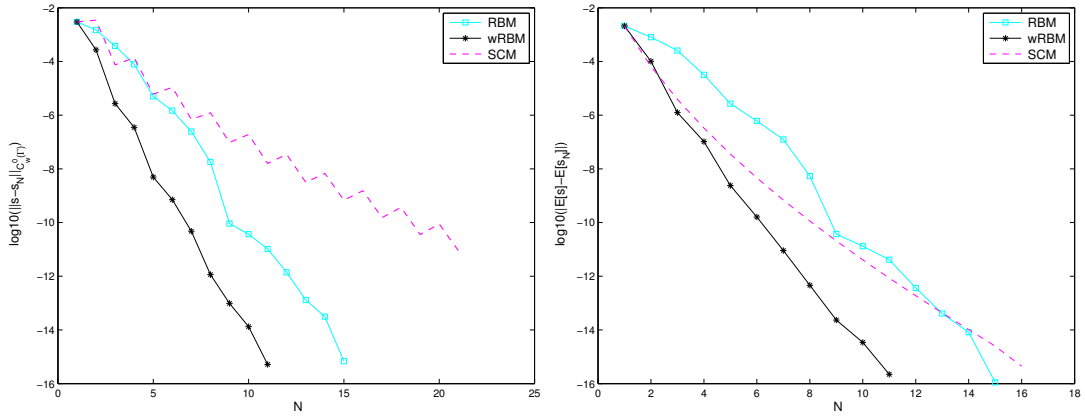

Figure 2.2: Left: convergence of the error $\log_{10}\left(||s-s_N||_{C^0_w(\Gamma)}\right)$ by the reduced basis method (RBM), the weighted reduced basis method (wRBM) and the stochastic collocation method (SCM). Right: convergence of the error $\log_{10}(|E[s]-E[s_N]|)$ by RBM, wRBM, and SCM, both with $K=1$, Beta$(10,10)$.

2.3.2 Multidimensional problem

For the test of a multidimensional problem, we consider the following coefficient

$$a(x,\omega)=\frac{1}{10}\left(4+\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2}y_1(\omega)+\sum_{n=1}^{2}\sqrt{\lambda_n}\left(\sin(n\pi x_1)y_{2n}(\omega)+\cos(n\pi x_1)y_{2n+1}(\omega)\right)\right), \tag{2.54}$$

where $y_k, 1\le k\le 5$, obeying Beta$(100,100)$, $L=1/4$ and $\lambda_1=0.3798, \lambda_2=0.2391$. A sufficient number of $n_{train}=10000$ samples (in fact $n_{train}=1000$ provides almost the same result in this example) obeying independent and identically distributed $y_k\sim$ Beta$(100,100), 1\le k\le 5$, are taken within the parameter domain $\Gamma=[-1,1]^5$ to construct the reduced basis space and another 1000 samples following the same distribution are taken independently to test different methods. We compare the performance of the weighted reduced basis method, the reduced basis method, and a sparse grid collocation method, with results displayed in Figure 2.3. The two reduced basis methods are obviously more efficient in both norms (2.51) with the weighted type providing faster convergence: the number of bases constructed for the weighted reduced basis method ($N_{max}=15$) is half that necessary for the reduced basis method ($N_{max}=30$).

As for the computational effort, the stochastic collocation method with sparse grid depends critically on the dimension [149], while the reduced basis methods are near the best approximation in the sense that they considerably alleviate the “curse of dimensionality" for analytic problems and save the computational effort significantly for high-dimensional problems, especially those with big cost for one deterministic solving. The weighted reduced basis method uses fewer bases than the conventional reduced basis method in both offline construction and online evaluation and thus costs less computational effort, particularly for high concentrated weight function, as shown in the above examples. For a detailed comparison of computational cost for the reduced basis method and the stochastic collocation method in various conditions, notably for large-scale and high-dimensional problems, see [50] and results anticipated in chapter 1.

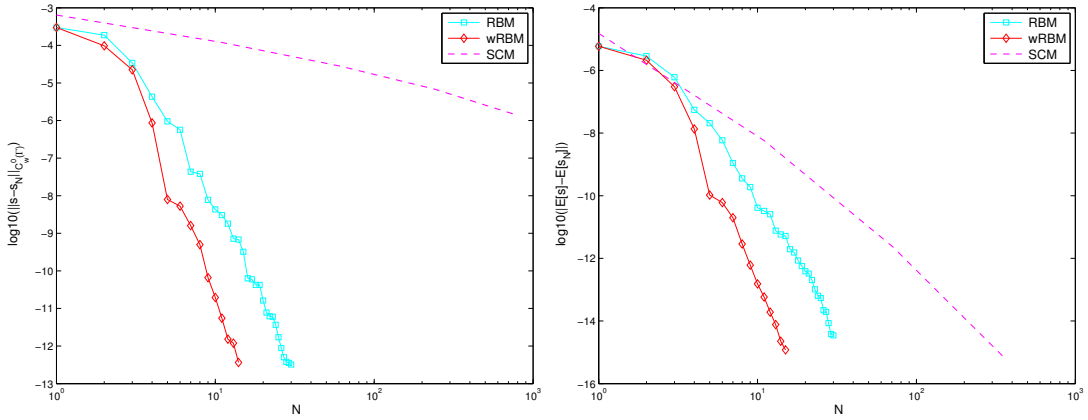

Figure 2.3: Left: convergence of the error $\log_{10}\left(||s-s_N||_{C^0_w(\Gamma)}\right)$. Right: convergence of the error $\log_{10}\left(|E[s]-E[s_N]|\right)$, computed by RBM, wRBM, and SCM, both with $K=5$, Beta$(100,100)$.

2.4 Summary

We proposed a weighted reduced basis method to deal with parametric elliptic problems with distinctive weight or importance at different values of the parameters. This method is particularly suited in solving stochastic problems with random variables obeying various probability distributions. Analytic regularity of the stochastic solution with respect to random variables was obtained under certain assumptions for the random input data, based on which an exponential convergence property of this method was studied by constructive approximation of general functions with analytic dependence on the parameters. The computational efficiency of the proposed method was compared to the ones of the classical reduced basis method and the (sparse grid) stochastic collocation method and was demonstrated numerically for both univariate and multivariate stochastic elliptic problems.

There are a few potential limitations we would like to warn the reader about: first, the performance of the weighted reduced basis method for low regularity problems is to be investigated, possibly improved by combination with the hp-adaptive reduced basis method [65]. Second, the efficient empirical interpolation method [11, 48] needs to be applied in order to use the weighted reduced basis method to solve nonlinear stochastic problems or linear stochastic problems with nonaffine random inputs exhibiting various probability structure. Finally, we would like to mention that application of the weighted reduced basis method to more general problems, e.g., parabolic problems [87], fluid dynamics [162], multiscale and multiphysics problems [117], stochastic optimization problems [47], and inverse problems [133], as well as more general stochastic problems with different probability structures is seen as upcoming perspectives.

Appendix A

To further illustrate the efficiency of the weighted reduced basis method, we present the following numerical examples with some widely used weight functions other than those introduced in section 2.3.

1. Weight function as truncated probability density function of normal distributed random variable:

$$a(x,\omega)=\frac{1}{10}(3.1+\sin(2\pi x_1)\,Y(\omega)\mathbb{I}(|Y|\le 3)),\ Y\sim \text{Normal}(\mu,\sigma);\ w(y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right);$$

2. Weight function as truncated probability density function of gamma distributed random variable:

$$a(x,\omega) = \frac{1}{10}(10.1 + \sin(2\pi x_1)\, Y(\omega)\mathbb{I}(Y \le 10)),\ Y \sim \text{Gamma}(k,\gamma);\ w(y) = \frac{1}{\gamma^k \Gamma(k)} y^{k-1} \exp(-\frac{y}{\gamma});$$

3. Weight function as truncated probability density function of Poisson distributed random variable:

$$a(x,\omega) = \frac{1}{10}(100.1 + \sin(2\pi x_1)\, Y(\omega)\mathbb{I}(Y \le 100)),\ Y \sim \text{Poisson}(\lambda);\ w(y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

The selected samples for different weight functions and error of $\log_{10}(||s - s_N||_{C^0_w(\Gamma)})$ are displayed in Figures 2.4, 2.5, and 2.6, respectively, from which we can observe that the samples are effectively chosen according to the weight functions. Consequently, both the offline construction and the online evaluation become more efficient by the weighted reduced basis method than the conventional one.

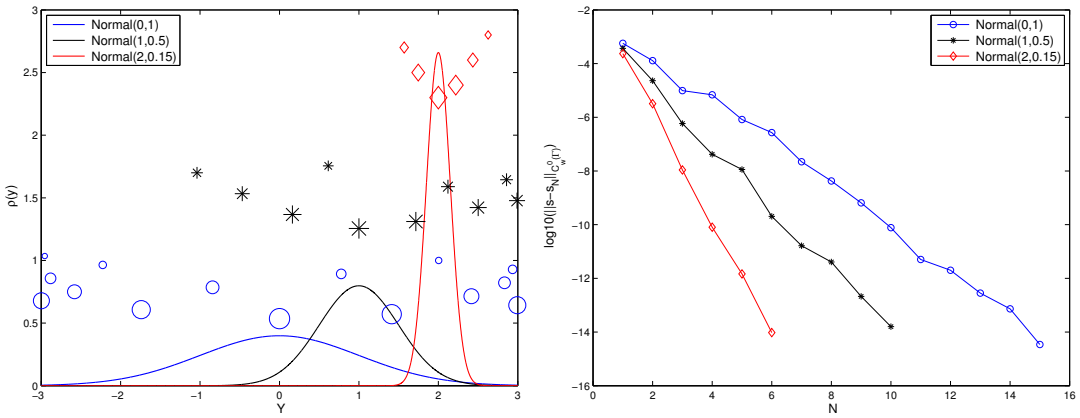

Figure 2.4: Left: probability density function of $Y \sim \text{Normal}(\mu,\sigma)$ with different μ,σ and samples selected by weighted reduced basis approximation; the bigger the size the earlier it has been selected. Right: convergence of the error $\log_{10}\left(||s - s_N||_{C^0_w(\Gamma)}\right)$ by the weighted reduced basis method.

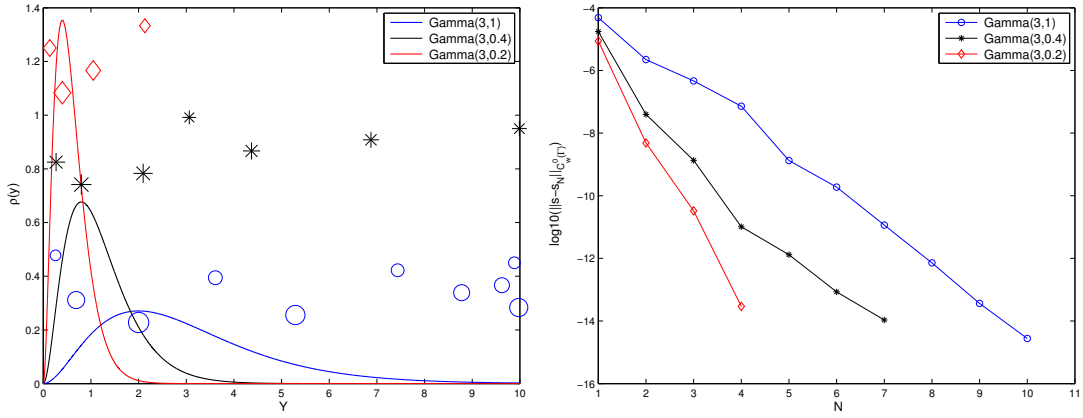

Figure 2.5: Left: probability density function of $Y \sim \text{Gamma}(k,\gamma)$ with different γ and samples selected by weighted reduced basis approximation; the bigger the size the earlier it has been selected. Right: convergence of the error $\log_{10}\left(||s-s_N||_{C^0_w(\Gamma)}\right)$ by the weighted reduced basis method.

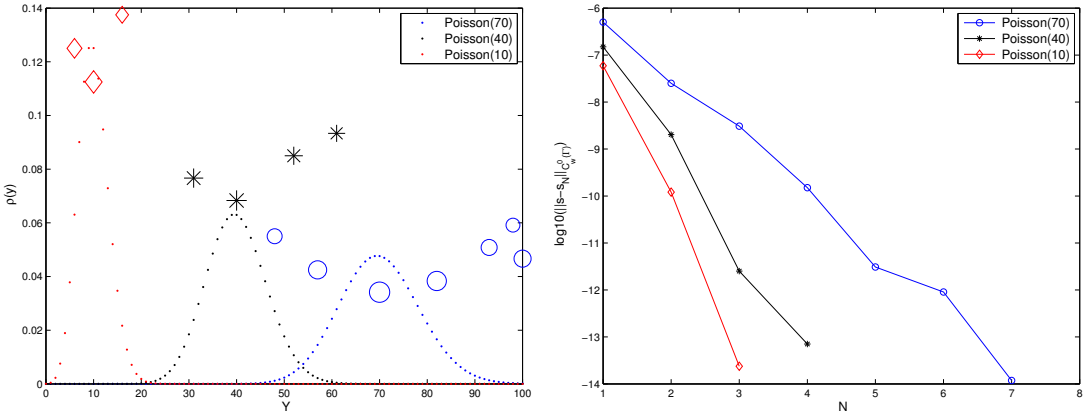

Figure 2.6: Left: probability density function of $Y \sim \text{Poisson}(\lambda)$ with different λ and samples selected by weighted reduced basis approximation; the bigger the size the earlier it has been selected. Right: convergence of the error $\log_{10}\left(||s-s_N||_{C^0_w(\Gamma)}\right)$ by the weighted reduced basis method.

3 Decomposition of nonaffine fields – a weighted empirical interpolation

A critical assumption made in both chapter 1 and chapter 2 to achieve the efficiency of the reduced basis method is that the random fields have an affine structure, which makes the offline-online decomposition possible. However, in practice the random fields are not necessarily given in the form of affine expansion; for instance, a lognormal random field

$$g(x,y) = \exp\left(\sum_{k=1}^{K} g_k(x)\, y_k\right), \tag{3.1}$$

is most often assumed for a diffusion coefficient that is strictly positive. This is of course nonaffine in the random variables $y = (y_1, \ldots, y_K)$. In order to retain the advantage of the offline-online decomposition, a common strategy consists in approximating the nonaffine structure by an affine function involving a separation of physical variables and random variables.

The empirical interpolation method [11] was originally developed to approximate the nonaffine coefficients of a partial differential equation in order to effectively decompose the reduced basis method into the offline construction and the online evaluation. Since its development, many applications of this method have been considered and several extentions proposed [86, 129, 201, 63, 176, 118, 40, 163]. In particular, we mention its application and analysis in the context of the reduced basis approximation for nonlinear elliptic and parabolic equations [86], in the geometrical parametrization of domains and shapes [132], and its extension to a general, multipurpose interpolation procedure [129], in which a priori error estimate compared to Kolmogorov N-width was obtained.

The basic idea behind empirical interpolation for parametric function $g(x,y)$ is to choose the parameter samples $y^1, y^2, \ldots$ and the interpolation nodes $x^1, x^2, \ldots$ recursively in a greedy approach according to the criteria that the values y^m and x^m selected at each step $m = 1,2,\ldots$ are the most representative ones in L^∞ norm, that is the ones where the function is worst approximated by the interpolation formula constructed from the previous steps [11]. This is essentially different from the conventional interpolation construction which requires the interpolation nodes to be chosen a priori according to a specific rule, e.g., roots of orthogonal polynomials [164]. The so called “magic points" [129] (y^m, x^m), $m = 1,2,\ldots$ obtained by the goal-oriented or the function-specified empirical interpolation procedure are supposed to identify an interpolation formula by capturing some specific features (e.g., regularity, extreme values) of the given function, thus providing higher interpolation accuracy. Another

---

Reference for this chapter:

P. Chen, A. Quarteroni, and G. Rozza. A weighted empirical interpolation method: A priori convergence analysis and applications. ESAIM: Mathematical Modelling and Numerical Analysis, in press, online doi: 10.1051/m2an/2013128, 2013.

superiority of the empirical interpolation construction is attributed to the affine expansion of the function given in whatever form, leading to the separation of the physical variables and the random variables in the following expression:

$$g(x,y) \approx \mathscr{I}_M[g] = \sum_{j=1}^{M} \Theta_j(y) q_j(x), \tag{3.2}$$

which can be efficiently exploited in conducting mathematical manipulation, e.g., numerical integration or reduced basis approximation [86]. Conventionally, one supposes that the parameter y, if viewed as a random variable, is uniformly distributed in a bounded space Γ.

However, in many applications, e.g., stochastic problems with parametrized random variables that obey normal distribution, the request of the boundedness of the parameter space Γ and that of the uniform distribution of the parameter y is quite difficult to be fulfilled. In these situations, the approximation to some quantities of interest (e.g., statistics of the function) based on the parameter samples and interpolation nodes, selected by the empirical interpolation procedure, would not lead to results that are as accurate or efficient as those expected when taking distinct weights of the parameter at different values into account. In this chapter we propose a weighted empirical interpolation method (wEIM) by considering a weighted optimization problem and analyzing its convergence properties by improving the a priori error estimate obtained in [129]. To demonstrate numerically its effectiveness and efficiency, we apply the wEIM to approximating nonlinear parametric functions, geometric Brownian motion in one dimension, exponential Karhunen–Loève expansion in multiple dimensions as well as reduced basis approximation to nonaffine stochastic elliptic problems, and we compare it with the conventional empirical interpolation method (EIM) and sparse grid stochastic collocation method. It is worth mentioning that constructing a goal-oriented numerical method is a quite common procedure in adaptive finite element methods [14, 82] and has also been applied to construct adaptive reduced basis method [41].

This chapter is organized as follows. We present the weighted empirical interpolation method in section 3.1. A priori convergence analysis is carried out in section 3.2, followed by section 3.3 where different applications of this method are addressed. Some limitations and perspectives of the weighted empirical interpolation are provided in section 3.4.

3.1 Weighted empirical interpolation method (wEIM)

For notational convenience, we introduce the spaces $L^\infty(D)$ defined in a bounded physical domain $D \subset \mathbb{R}^d, d \in \mathbb{N}_+$ and $C^0_w(\Gamma)$ defined in a parameter space (not necessarily bounded) $\Gamma \subset \mathbb{R}^K, K \in \mathbb{N}_+$, which are equipped with the following norms:

$$||g||_{L^\infty(D)} = \operatorname*{ess\,sup}_{x\in D} |g(x)| \text{ and } ||g||_{C^0_w(\Gamma)} = \max_{y\in\Gamma} w(y)|g(y)| \tag{3.3}$$

for a given positive weight function $w : \Gamma \to \mathbb{R}_+$. We also define the Bochner space $L^\infty(D; C^0_w(\Gamma))$ for a parameter dependent function equipped with the norm

$$||g||_{L^\infty(D;C^0_w(\Gamma))} = \operatorname*{ess\,sup}_{x\in D} \left(\max_{y\in\Gamma} w(y)|g(x,y)| \right) \equiv \max_{y\in\Gamma} w(y) \left(\operatorname*{ess\,sup}_{x\in D} |g(x,y)| \right). \tag{3.4}$$

We note that $L^\infty(D)$, as used in [11, 86, 129], is usually replaced with $C^0(D)$ for conventional interpolation of continuous functions [164].

At the discrete level, the physical domain D is replaced by a set of vertices $x \in V_x$ with finite cardinality $n_x = |V_x| < \infty$, for instance finite element nodes, and the parameter space Γ is represented by a sample

set Ξ_y of finite cardinality $n_y = |\Xi_y| < \infty$. We present the weighted empirical interpolation method in Algorithm 3. We emphasize that the initial sample y^1 is chosen such that the weighted function is maximized in $L^\infty(V_x; C^0_w(\Xi_y))$ norm:

$$y^1 = \arg\max_{y\in\Xi_y} \left[w(y) \left(\operatorname*{ess\,sup}_{x\in V_x} |g(x,y)| \right) \right]. \tag{3.5}$$

In the course of the construction procedure, the quasi-optimal samples $y^{M+1}, M \geq 1$, can be chosen by a greedy algorithm to minimize the weighted optimal approximation error (3.6) in the subspace $W_M := \text{span}\{g(\cdot, y^i), 1 \leq i \leq M\}$, i.e., find $y^{M+1} \in \Xi_y$ such that

$$y^{M+1} = \arg\max_{y\in\Xi_y} \left[w(y) \left(\inf_{h\in W_M} ||g(y) - h||_{L^\infty(V_x)} \right) \right]. \tag{3.6}$$

However, the weighted L^∞ optimization problem (3.6) is expensive to solve by linear programming if $|V_x|$ and $|\Xi_y|$ are large. In practice, it can be efficiently replaced by a weighted L^2 optimization problem [86] or by a surrogate weighted L^∞ optimization problem (3.9) [129].

We state several properties of the weighted empirical interpolation method in the following lemmas, whose proof is straightforward by noting the fact that the weight function $w : \Gamma \to \mathbb{R}_+$ is positive, and therefore omitted here; see, for instance, [11, 86, 129] for details.

Lemma 3.1.1 *For any $M < M_{max}$, the subspace $Q_M = span\{q_m, 1 \leq m \leq M\}$ is of dimension M. Moreover, the matrix B^M formed in (3.12) is lower triangular with unity diagonal and thus invertible.*

Lemma 3.1.2 *For any function $h \in Q_M$, the empirical interpolation formula given by (3.7) is exact, i.e., $r_{M+1}(x,y) = 0 \; \forall x \in V_x$ and $y \in \Xi_y$. In general, for any function $g \in L^\infty(D; C^0_w(\Gamma))$, we have*

$$||g - \mathcal{I}_M[g]||_{L^\infty(D)} \leq (1 + \Lambda_M) \inf_{h\in Q_M} ||g - h||_{L^\infty(D)} \text{ with } \Lambda_M \leq 2^M - 1. \tag{3.13}$$

Remark 3.1.1 *In practice, the empirical interpolation method is always carried out on a discrete finite vertex set V_x instead of the domain D. In order to make this idea more explicit, a variant version of the empirical interpolation method, under the name of* discrete empirical interpolation method (DEIM)*, is proposed in [40] to solve nonlinear problems. In particular, nonlinear systems of ordinary differential equations (which have similar structure as nonlinear time-dependent partial differential equations after spatial discretization) were efficiently treated by the DEIM in [40].*

3.2 A priori convergence analysis

The interpolation error obtained in (3.13) with the Lebesgue constant $\Lambda_M \leq 2^M - 1$ (see proof in [86]) by the empirical interpolation procedure is too pessimistic, far from the result for conventional interpolation error based on certain prescribed interpolation nodes (e.g., Chebyshev nodes with $\Lambda_M \sim \log(M)$ [164]). An explicit a priori convergence rate of the weighted empirical interpolation error is not available for generic functions. In order to measure the accuracy of the approximation by the weighted empirical interpolation method, in the following theorem we compare it with the Kolmogorov N-width [160], which quantifies the optimal approximation error of a subset $\mathscr{F}$ in a Banach space $\mathscr{H}$ by any possible N dimensional subspace F_N, defined as

$$d_N(\mathscr{F}, \mathscr{H}) := \inf_{F_N \subset \mathscr{H}} \sup_{g\in\mathscr{F}} \inf_{f\in F_N} ||g - f||_{\mathscr{H}}. \tag{3.14}$$

Algorithm 3 The weighted empirical interpolation method

procedure INITIALIZATION:
  Given finite vertex set $V_x \subset D$, sample set $\Xi_y \subset \Gamma$, weight w and function $g \in L^\infty(V_x; C^0_w(\Xi_y))$;
  find $y^1 \in \Xi_y$ such that $y^1 = \arg\max_{y\in\Xi_y} w(y)(\operatorname{ess\,sup}_{x\in V_x} |g(x,y)|)$; set $W_1 = \operatorname{span}\{g(x,y^1)\}$;
  find $x^1 \in V_x$ such that $x^1 = \arg\operatorname{ess\,sup}_{x\in V_x} |g(x,y^1)|$;
  define $r_1 = wg$, $q_1(x) = r_1(x,y^1)/r_1(x^1,y^1)$, $B^1_{11} = 1$, set $M = 1$, specify tolerance ε_{tol};
end procedure
procedure CONSTRUCTION:
  while $M < M_{max}$ & $r_M(x^M, y^M) > \varepsilon_{tol}$ **do**
    find $\Theta^M(y) = (\Theta^M_1(y), \ldots, \Theta^M_M(y))^T$ by solving

$$\sum_{j=1}^{M} \Theta^M_j(y) q_j(x^i) = g(x^i, y) \quad 1 \le i \le M; \tag{3.7}$$

    define $r_{M+1} : D \times \Gamma \to \mathbb{R}$ as

$$r_{M+1}(x,y) = g(x,y) - \sum_{j=1}^{M} \Theta^M_j(y) q_j(x); \tag{3.8}$$

    find $y^{M+1} \in \Xi_y$ such that

$$y^{M+1} = \arg\max_{y\in\Xi_y} \left[w(y) \left(\operatorname*{ess\,sup}_{x\in V_x} |r_{M+1}(x,y)| \right) \right], \tag{3.9}$$

    find $x^{M+1} \in V_x$ such that

$$x^{M+1} = \arg\operatorname*{ess\,sup}_{x\in V_x} |r_{M+1}(x, y^{M+1})|; \tag{3.10}$$

    define $q_{M+1} : D \to \mathbb{R}$ as

$$q_{M+1}(x) = \frac{r_{M+1}(x, y^{M+1})}{r_{M+1}(x^{M+1}, y^{M+1})}; \tag{3.11}$$

    update matrix $B^{M+1} \in \mathbb{R}^{(M+1)\times(M+1)}$ as

$$B^{M+1}_{ij} = q_j(x^i) \quad 1 \le i, j \le M+1; \tag{3.12}$$

    set M = M+1;
  end while
end procedure
procedure EVALUATION:
  For $\forall y \in \Xi_y$, construct approximation (3.2) by solving (3.7), then evaluate (3.2) at $\forall x \in V_x$.
end procedure

In the context of empirical interpolation, we consider $\mathscr{H} \equiv L^\infty(D)$ and $\mathscr{F} \equiv F_g(D)$ as the image of the function g in Γ, i.e. $g : \Gamma \to F_g(D)$.

Theorem 3.2.1 *The error of the weighted empirical interpolation method can be bounded as follows:*

$$||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \le C_w(M+1)2^M d_M(F_g(V_x), L^\infty(V_x)), \tag{3.15}$$

where the constant C_w depends on the weight function w but is independent of M.

Remark 3.2.1 *In fact, the result (3.15) is obtained in the subspace $L^\infty(V_x)$ for the constructive weighted empirical interpolation method and can be straightforwardly extended to $L^\infty(D)$ when the vertex set V_x tends to D such that the points outside the vertex set V_x can be sufficiently well represented by the points inside. To be rigorous, we take the vertex set V_x such that for almost every $x \in D$, there exists $y \in V_x$ satisfying*

$$|g(x) - g(y)| \leq ||g - \mathscr{I}_M[g]||_{L^\infty(V_x)}. \tag{3.16}$$

Consequently, we have the error bound

$$\begin{aligned} ||g - \mathscr{I}_M[g]||_{L^\infty(D)} &\leq ||g - \mathscr{I}_M[g]||_{L^\infty(D\setminus V_x)} + ||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \\ &\leq 2||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \\ &\leq C_w(M+1)2^{M+1} d_M(F_g(V_x), L^\infty(V_x)) \\ &\leq C_w(M+1)2^{M+1} d_M(F_g(V_x), L^\infty(D)). \end{aligned} \tag{3.17}$$

The proof of (3.15) adopts a constructive approach inspired from that for the greedy algorithm for the reduced basis method [20]. Some preliminary results are provided in the next two lemmas.

For simplicity, we use the shorthand notation $r_m(x) = r_m(x, y^m), 1 \leq m \leq M+1$, obtained in Algorithm 3 and define the functions $t_j(x^i) = r_i(x^j), 1 \leq i, j \leq M+1$, and $t_j(x^i) = 0, 1 \leq j \leq M+1, i > M+1$.

Lemma 3.2.2 *The matrix T^{M+1} defined by $T^{M+1}_{ij} = t_j(x^i), 1 \leq i, j \leq M+1$, is upper triangular with dominating diagonal elements, i.e., $t_j(x^i) = 0, i > j$, and $|t_j(x^i)| \leq |t_j(x^j)|, i \leq j$.*

Proof From the result of Lemma (3.1.1), we know that the matrix B^{M+1} is lower triangular with unity diagonal. By the definition of $q_i, 1 \leq i \leq M+1$, in (3.11) and the definition of $t_j, 1 \leq j \leq M+1$, we have $t_j(x^i) = q_i(x^j) r_i(x^i)$, so that $t_j(x^i) = q_i(x^j) = 0, i > j$, and $|t_j(x^i)| \leq |t_j(x^j)| = |r_j(x^j)|, i \leq j$, due to (3.10). □

Lemma 3.2.3 *For any $1 \leq m \leq M+1$, there exists a unique $b = (b_1, \ldots, b_m)^T \in \mathbb{R}^m$ such that*

$$r_m(x) e_m(x) = \sum_{j=1}^{m} b_j t_j(x) \quad \forall x \in V_x, \tag{3.18}$$

where $e_m, 1 \leq m \leq M+1$, are unit vectors, i.e., $e_m(x^m) = 1$ and $e_m(x^n) = 0$ if $n \neq m$. In addition, we have $b_m = 1$ and the bound $|b_i| \leq 2^{m-i-1}, 1 \leq i < m$, so that $|b_1| + \cdots + |b_m| \leq 2^{m-1}$.

Proof For any $x = x^i, i > M+1$, we have $e_m(x) = 0$ and $t_j(x) = 0$, so that both sides of the equation vanish and we only need to verify the statement for $x = x^i, 1 \leq i \leq M+1$, in which case the system (3.18) becomes

$$s = Tb \quad \text{with } s = (0, \ldots, 0, r_m(x^m))^T. \tag{3.19}$$

Thanks to Lemma 3.2.2, we have that T is invertible and thus there exists a unique solution b. Moreover, the last row of the system (3.19) $r_m(x^m) = t_m(x^m) b_m$ leads to the solution $b_m = 1$ since $r_m(x^m) = t_m(x^m)$. For any other row $i, 1 \leq i < m$, we have by the fact that T is an upper triangular matrix

$$0 = \sum_{j=i}^{m} b_j t_j(x^i). \tag{3.20}$$

Recall that $|t_j(x^i)| \leq |t_i(x^i)|$, $j > i$, which yields the following bound for $b_i, 1 \leq i < m$,

$$|b_i| = \left| - \sum_{j=i+1}^{m} b_j \frac{t_j(x^i)}{t_i(x^i)} \right| \leq \sum_{j=i+1}^{m} |b_j|, \tag{3.21}$$

so that $|b_i| \leq 2^{m-i-1}, 1 \leq i < m$, and $|b_1| + \cdots + |b_m| \leq 2^{m-1}$ being $b_m = 1$ and using a recursive argument. □

We are now ready to prove Theorem 3.2.1 using the representation of the residual in Lemma 3.2.3.

Proof Suppose there exists a subspace $H_M \subset L^\infty(V_x)$ of dimension M achieving the Kolmogorov M-width as defined in (3.14), then we have a sequence of elements $h_j \in H_M, 1 \leq j \leq M+1$, such that

$$||t_j - h_j||_{L^\infty(V_x)} \leq d_M(F_g(V_x), L^\infty(V_x)), 1 \leq j \leq M+1. \tag{3.22}$$

We define the functions

$$s_m(x) = \sum_{j=1}^{m} b_j h_j(x), 1 \leq m \leq M+1. \tag{3.23}$$

Since all the elements $h_j, 1 \leq j \leq M+1$, belong to the M dimensional subspace H_M and s_m is a linear combination of these elements for any $m = 1, \ldots, M$, there exists a vector $\alpha = (\alpha_1, \ldots, \alpha_{M+1})^T$ with $|\alpha_1| + \cdots + |\alpha_{M+1}| = 1$ such that

$$\sum_{m=1}^{M+1} \alpha_m s_m = 0. \tag{3.24}$$

Thanks to the result in Lemma 3.2.3, together with bound (3.22) and representation (3.23) and (3.24), we obtain the following bound for every $x \in V_x$:

$$\begin{aligned} \left| \sum_{m=1}^{M+1} \alpha_m r_m(x) e_m(x) \right| &= \left| \sum_{m=1}^{M+1} \alpha_m \left(r_m(x) e_m(x) - s_m(x) \right) \right| \\ &\leq \left(\sum_{m=1}^{M+1} |\alpha_m| \right) \max_{m=1,\ldots,M+1} ||r_m e_m - s_m||_{L^\infty(V_x)} \\ &\leq \max_{m=1,\ldots,M+1} \left(\sum_{j=1}^{m} |b_j| \right) \max_{j=1,\ldots,m} ||t_j - h_j||_{L^\infty(V_x)} \\ &\leq 2^M d_M(F_g(V_x), L^\infty(V_x)). \end{aligned} \tag{3.25}$$

Since $|\alpha_1| + \cdots + |\alpha_{M+1}| = 1$, there must exist α_m such that $|\alpha_m| \geq 1/(M+1)$. Setting $x = x^m$ in (3.25), we have $|\alpha_m r_m(x^m)| \leq 2^M d_M(F_g(V_x), L^\infty(V_x))$, and thus

$$|r_m(x^m)| \leq (M+1) 2^M d_M(F_g(V_x), L^\infty(V_x)). \tag{3.26}$$

By the construction of the weighted empirical interpolation approximation in Algorithm 3, we have

$$\operatorname*{ess\,sup}_{x \in V_x} |r_{M+1}(x)| \leq |r_{M+1}(x^{M+1})| \leq |r_M(x^M)| \leq \cdots \leq |r_m(x^m)|. \tag{3.27}$$

A combination of (3.26) and (3.27) leads to the following error bound:

$$||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \leq \operatorname*{ess\,sup}_{x \in V_x} |r_{M+1}(x)| \leq (M+1) 2^M d_M(F_g(V_x), L^\infty(V_x)). \tag{3.28}$$

□

Corollary 3.2.4 *Under the assumption $d_M(F_g(V_x), L^\infty(V_x)) \le ce^{-rM}$ with $r > \log(2)$, we have the following a priori error estimate of the wEIM: $\forall g \in L^\infty(V_x; C^0_w(\Xi_y))$*

$$||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \le c(M+1)e^{-(r-\log(2))M}. \tag{3.29}$$

Remark 3.2.2 *The result (3.29) is an improvement of that recently obtained in [129], in which r is required to satisfy $r > 2\log(2)$ and the exponential convergence rate becomes $r - 2\log(2)$. In fact, when the function g is analytic with respect to the parameter $y \in \mathbb{R}$, the Kolmogorov width is bounded by the exponentially decaying error from the truncation of Fourier expansion of order M of g; see [61].*

Remark 3.2.3 *The result obtained in Theorem 3.2.1 can not be improved in the exponential growth 2^M for a priori convergence analysis of general parametric functions. In fact, it can be proved that $||g - \mathscr{I}_M[g]||_{L^\infty(V_x)} \ge (1-\varepsilon)2^M d_M(F_g(V_x), L^\infty(V_x))$ for arbitrary small $\varepsilon > 0$ under certain assumptions; see [20].*

3.3 Numerical experiments

In this section, we study the accuracy and the efficiency of the weighted empirical interpolation method (wEIM) compared to the conventional empirical interpolation method (EIM) as well as the stochastic collocation method (SCM) for one-dimensional problems and the sparse grid stochastic collocation method (SG-SCM) [149] for multidimensional problems. Given a function g, we denote by g_M its approximation using M "elements" (either basis functions for wEIM and EIM, or interpolation nodes for SCM and SG-SCM) and we define the approximation error in the following two norms

$$||g - g_M||_{L^\infty(D;C^0(\Gamma))} \quad \text{and} \quad ||\mathbb{E}[g] - \mathbb{E}[g_M]||_{L^\infty(D)}, \tag{3.30}$$

where the expectation $\mathbb{E}[g]$ is computed by Gauss quadrature formula specified when in need.

3.3.1 Parametric function in one dimension – geometric Brownian motion

We consider a geometric Brownian motion S_t satisfying a stochastic ordinary differential equation $dS_t = kS_t dt + \sigma S_t dB_t$ (This is, e.g., the most widely used model of stock price S_t at time t with drift k, volatility σ and standard Brown motion B_t [157]). The solution is given by $S_t = \exp(\sigma B_t + (k - \sigma^2/2)t)$. For simplicity, we set $S_0 = 1$, $\sigma = 1$ and $k = 1/2$ so that S_t can be written as $S_t = \exp(\sqrt{t}B_1)$, where B_1 is a standard Gauss random variable $B_1 \sim \mathcal{N}(0,1)$. By denoting $x \equiv t$, $y \equiv B_1 \in \mathbb{R}^K, K = 1$, and $g = S_t$, we seek the following affine expansion by wEIM given in Algorithm 3

$$g(x,y) = \exp(\sqrt{x}y) \approx g_M(x,y) = \sum_{j=1}^{M} \Theta_j(y) q_j(x) \text{ where } y \sim \mathcal{N}(0,1). \tag{3.31}$$

Moreover, we are interested in the expectation of g at time x, which can be approximated by Gauss-Hermite quadrature with abscissas and weights $(y_n, w_n), 1 \le n \le N$,

$$\mathbb{E}_y[g](x) \approx \sum_{j=1}^{M} \left(\int_{-\infty}^{\infty} \Theta_j(y)\rho(y)dy \right) q_j(x) \approx \sum_{j=1}^{M} \left(\sum_{n=1}^{N} \Theta_j(y_n) w_n \right) q_j(x), \tag{3.32}$$

where ρ is standard normal density function. The advantage of (3.32) is that we do not need to compute the function g for $y_n, 1 \le n \le N$, at every x but only at the empirical interpolation nodes $x^m, 1 \le m \le M$,

which is attributed to solving a small linear system (3.7) for $\Theta_j(y_n), 1 \le j \le M, 1 \le n \le N$. When the evaluation of the function itself at (x, y) is expensive and we have a large number of points x, the wEIM can be employed for efficient computation of the statistics. We set the tolerance as $\varepsilon_{tol} = 1 \times 10^{-12}$, and take 1000 equidistant points in the vertex set V_x and 1000 normal distributed samples in the sample set Ξ_y; we also take an independent 1000 normal distributed samples to test different interpolation methods. The weight in Algorithm 3 is taken as the normalized Gauss density function $w(y) = \rho(y)/\rho(0)$. As for the evaluation of the expectation of $\mathbb{E}[g_M]$, we use 12 quadrature abscissas in (3.32), which is sufficiently accurate for this example. We examine the convergence of "EIM bound" and "wEIM bound" ($r_M(x^M)$), error by "EIM test" and "wEIM test" (error computed from test samples) and test error by stochastic collocation method "SCM test".

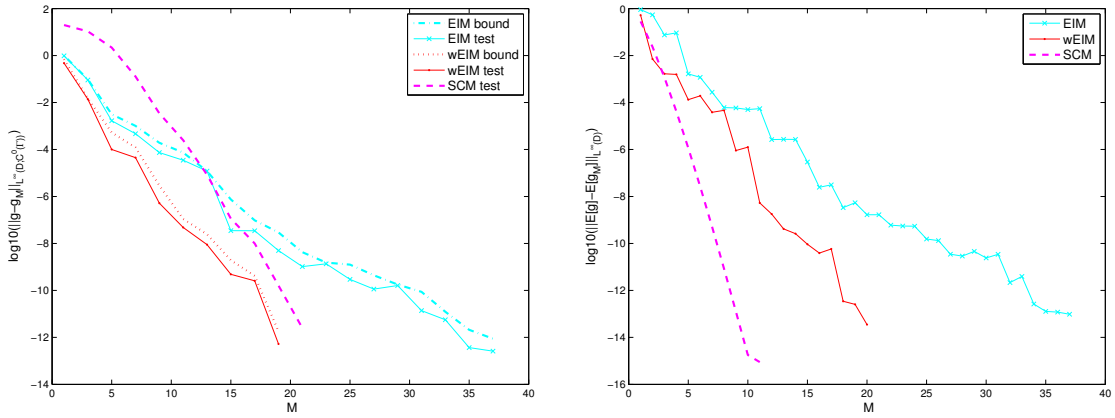

Figure 3.1: Comparison of the convergence rates of EIM, wEIM and SCM in different norms. Left: decreasing of the error $||g - g_M||_{L^\infty(D;C^0(\Gamma))}$; right: decreasing of the error $||\mathbb{E}[g] - \mathbb{E}[g_M]||_{L^\infty(D)}$.

The convergence property of different methods is displayed in Figure 3.1, from which we can see that all the methods achieve an exponential convergence rate and wEIM converges faster than both SCM and EIM in $L^\infty(D; C^0(\Gamma))$ norm. However, as for the expectation in $L^\infty(D)$ norm, SCM is the best and wEIM is evidently better than EIM, which does not take the weight into consideration. The reason for these results is that wEIM and EIM select the samples by $L^\infty_w(\Xi_y)$ and $L^\infty(\Xi_y)$ optimization, leading to small error in $L^\infty(D; C^0(\Gamma))$ norm and relatively large error for the evaluation of the expectation.

3.3.2 Parametric function in multiple dimensions – Karhunen–Loève expansion

For the case of multidimensional parameters, we consider the function g truncated from Karhunen–Loève expansion of a Gaussian random field with correlation length L and eigenvalues $\lambda_n, 1 \le n \le N_t$, written as [149]

$$g(x,y) - g_0(x) = C \exp\left(\left(\frac{\sqrt{\pi}L}{2}\right)^{\frac{1}{2}} y_1(\omega) + \sum_{n=1}^{N_t} \sqrt{\lambda_n}\left(\sin(n\pi x) y_{2n}(\omega) + \cos(n\pi x) y_{2n+1}(\omega)\right)\right), \tag{3.33}$$

where $y_i \sim \mathcal{N}(0,1), 1 \le i \le 2N_t + 1$, are standard Gauss random variables defined in the sample space $\Omega \ni \omega$. This function is widely used, e.g., in modelling the random property of porous medium in material science, geophysics, etc.. To compare the convergence properties of different methods, we take $g_0 = 0$, $C = \exp(5)$, $N_t = 2$, $L = 1/8$, and $\lambda_1 = 0.213, \lambda_2 = 0.190$; $x \in [0,1]$ is discretized by 1000 equidistant vertices. We set tolerance $\varepsilon_{tol} = 1 \times 10^{-12}$, and use 1000 five dimensional independent normal distributed samples and another 1000 test samples. For the computation of $\mathbb{E}[g]$, we apply SG-SCM based on Gauss-Hermite quadrature with the deepest interpolation level 4 in each dimension [149].

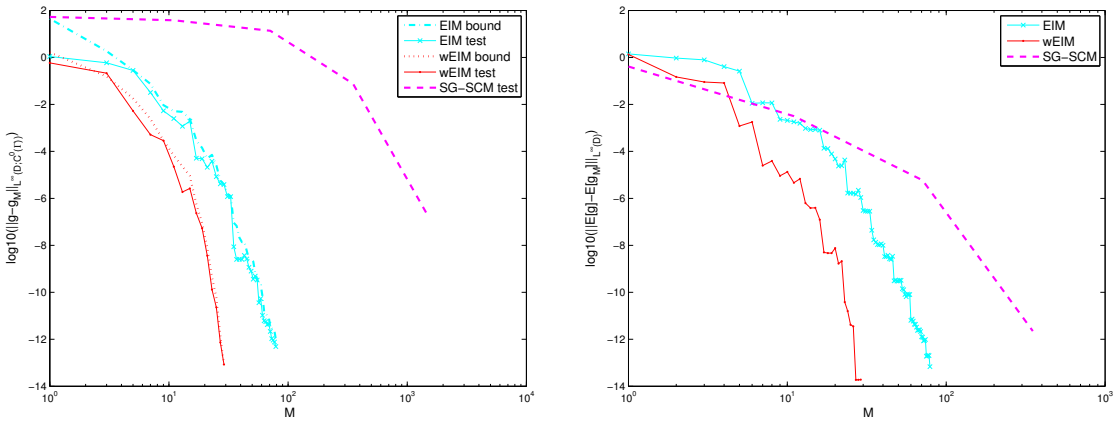

Figure 3.2: Comparison of the convergence rates of EIM, wEIM and SG-SCM in different norms. Left: decreasing of the error $||g - g_M||_{L^\infty(D;C^0(\Gamma))}$; right: decreasing of the error $||\mathbb{E}[g] - \mathbb{E}[g_M]||_{L^\infty(D)}$.

Figure 3.2 depicts the convergence rate of different methods, from which we can observe that in multi-dimensional problems wEIM and EIM perform much better than SG-SCM in both $||g - g_M||_{L^\infty(D;C^0(\Gamma))}$ error and $||\mathbb{E}[g] - \mathbb{E}[g_M]||_{L^\infty(D)}$ error. Both wEIM and EIM achieve fast exponential convergence rate and considerably alleviate the "curse of dimensionality" suffered by SG-SCM. wEIM uses only 29 samples while EIM needs 80 samples and thus 80 expansion terms, which is far less efficient than the weighted version in practical applications, e.g., in approximating the nonaffine terms of reduced basis method.

3.3.3 Parametric PDEs – application to the reduced basis method

As mentioned before, EIM was originally developed to deal with nonaffine terms in reduced basis discretization of partial differential equations (PDEs) in [11]. The efficiency of the reduced basis method depends critically on the number of affine terms for both offline construction and online evaluation [86, 50, 116, 134]. Therefore, wEIM is more suitable for reduced basis approximation of nonaffine parametric equation with weighted parameters or random variables with arbitrary probability measures.

We consider the following elliptic equation with a random coefficient and homogeneous Dirichlet boundary condition: find $u : D \times \Omega \to \mathbb{R}$ such that

$$-\nabla(g(x,\omega)\nabla u(x,\omega)) = f(x) \quad (x,\omega) \in D \times \Omega, \tag{3.34}$$

where the random coefficient $g(x,\omega)$ is a Gauss random field represented by a truncated Karhunen–Loève expansion as in (3.33). We set $D = (0,1)^2$, $f = 1$, $g_0 = 0.1$, $C = \exp(5)$, $L = 1/16$, $N_t = 5$, $\lambda_1 = 0.110, \lambda_2 = 0.107, \lambda_3 = 0.101, \lambda_4 = 0.095, \lambda_5 = 0.087$, and identify the eigenfunctions in (3.33) as $\sin(n\pi x_1)$ and $\cos(n\pi x_2)$ with $x_1, x_2 \in [0,1]$. The tolerance for weighted empirical interpolation method is taken as $\varepsilon_{tol} = 1 \times 10^{-12}$. Note that the problem has 11 independent and normal distributed random variables $y_K \sim \mathcal{N}(0,1), 1 \le K \le 11$, and all the random variables have relatively equivalent importance due to very close eigenvalues. Therefore, we employ isotropic sparse grid stochastic collocation method based on Gauss-Hermite quadrature [149] for the computation of statistics.

We first run wEIM and EIM with finite element vertices $|V_x| = 185$ and normal distributed samples $|\Xi_y| = 10000$ to build an affine expansion 3.2 for the coefficient g of problem (3.34). Another independent 1000 normal distributed samples are used to test the accuracy of the two expansions. The results are shown on the left of Figure 3.3, from which we can observe that wEIM is much more efficient with only 31

affine terms than EIM requiring 94 terms to achieve the same approximation accuracy in $L^\infty(D;C^0(\Gamma))$ norm.

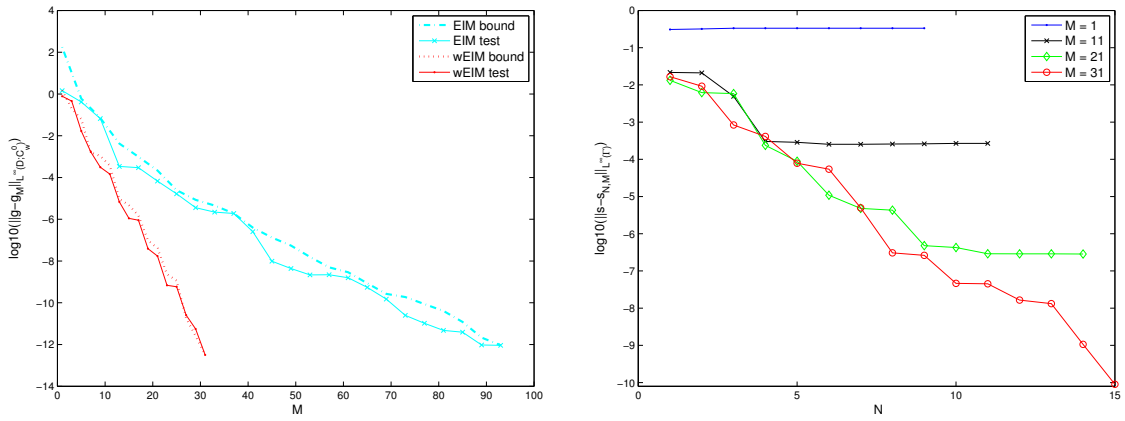

Figure 3.3: The convergence rate of wEIM in reduced basis approximation. Left: decreasing of the error $||g-g_M||_{L^\infty(D;C^0(\Gamma))}$ for EIM and wEIM; right: decreasing of the error $||s-s_{N,M}||_{L^\infty(\Gamma)}$.

We use the affine expansion constructed by wEIM to build a weighted reduced basis approximation, as introduced in chapter 2, with finite element discretization in physical domain D to the stochastic elliptic problem (3.34). The quantity of interest is the integral of the solution over the physical domain D, $s=\int_D u dx$, which is computed from the finite element solution. We denote $s_{N,M}$ the approximation of s based on using N reduced bases and M affine terms. The convergence of $||s-s_{N,M}||_{L^\infty(\Gamma)}$ is displayed on the right side of Figure 3.3, which demonstrates that wEIM is efficient in the application of the reduced basis method resulting in only a few elements in the reduced basis space. Moreover, we can see that the accuracy of wEIM, represented by different number of affine terms $M=1,11,21,31$, is clearly quite influential to the accuracy of the reduced basis approximation.

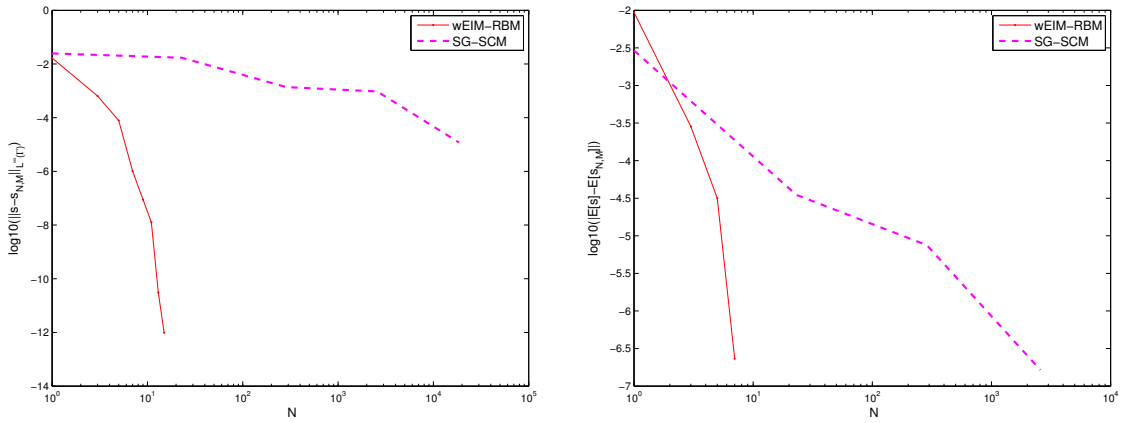

Figure 3.4: Comparison of the convergence rates between methods wEIM-RBM and SG-SCM. Left: decreasing of the error $||s-s_{N,M}||_{L^\infty(\Gamma)}$; right: decreasing of the error $|\mathbb{E}[s]-\mathbb{E}[s_{N,M}]|$.

Finally, we compare the proposed approach, a combination of weighted empirical interpolation with reduced basis approximation (wEIM-RBM), to one of the most efficient stochastic computational methods - SG-SCM [149] for their accuracy and efficiency. The result of this comparison, for the $||s-s_{N,M}||_{L^\infty(\Gamma)}$ norm, is depicted on the left of Figure 3.4, from which the curse of dimensionality of

SG-SCM can be obviously observed. In contrast, wEIM-RBM effectively alleviates this computational burden, using merely 15 bases to accurately approximate the stochastic solution depending on 11 independent normal distributed random variables.

As for the approximation of expectation $\mathbb{E}[s]$, we only need to compute the quantity $s_{N,M}$ with $N = 15$ and $M = 31$ by online evaluation of reduced basis method at the sparse Gauss quadrature abscissas and then $\mathbb{E}[s_{N,M}]$ by sparse grid Gauss quadrature formula.

The comparison of wEIM-RBM with SG-SCM on the right side of Figure 3.4 shows that in order to achieve the same accuracy, it takes only 7 bases by reduced basis approximation while 2575 collocation nodes for stochastic collocation approximation. It is worth to mention that the online evaluation of the reduced basis method is independent of the degree of freedom ($|V_x|$) of the deterministic system. Therefore, when solving the underlying deterministic system is computational demanding (with large $|V_x|$) and the dimension of the stochastic space becomes high (with more random variables), wEIM-RBM is much more efficient than SG-SCM for nonaffine stochastic problems; see [50] for detailed comparison of computational cost.

3.4 Summary

In order to approximate parametric functions with weighted parameters, e.g., random variables with various probability distributions, we extended the empirical interpolation method by taking the weight into account for the construction of the interpolation formula. A priori convergence analysis of the weighted empirical interpolation method has been provided. We obtained a direct comparison of the interpolation error to the Kolmogorov width, which improved the result obtained recently in [129].

By the applications in approximating geometric Brownian motion in one dimension and exponential Karhunen–Loève expansion in multiple dimensions, we demonstrated numerically the exponential convergence rate of the weighted empirical interpolation method and its advantage in accuracy and efficiency over the empirical interpolation method as well as over the sparse grid stochastic collocation method. We also applied the proposed method to the weighted reduced basis approximation [49] for a nonaffine stochastic elliptic equation and illustrated its efficiency and especially its effectiveness in alleviating the curse of dimensionality in comparison with the sparse grid stochastic collocation method.

The weighted empirical interpolation method can be straightforwardly applied to nonlinear stochastic partial differential equations with reduced basis approximation and can also be employed effectively in several fields embracing weighted parameters or random variables, e.g., image science, geophysics, mathematical finance, material science, bioengineering, to cite a few without being exhaustive.

4 Hybrid and goal-oriented adaptive reduced basis methods for risk analysis

In chapter 1, the reduced basis method was demonstrated to be very efficient and sufficiently accurate in solving uncertainty quantification problems, then further improved in chapter 2 thanks to the weighted algorithm for evaluation of statistical moments. Besides statistical moments and their related quantities of interest, such as variance-based sensitivity analysis, other quantities of interest depend on pointwise evaluation of the stochastic/parametric outputs. In this framework, risk analysis provides the most representative example. This chapter is devoted to the development of suitable algorithms in order to accurately apply the reduced basis method for uncertainty quantification problems requiring pointwise evaluation, in particular, but not limited to, evaluation of failure probability.

Several computational methods, such as the Monte Carlo method [72], the first or second order reliability method [166, 187], the response surface method [69, 29], etc., have been developed for the evaluation of failure probability for risk prediction or reliability analysis of a given system featuring various uncertainties or random inputs. However, efficient and accurate evaluation of failure probability is difficult to achieve, especially for a given system modeled by partial differential equations (PDEs) with high-dimensional random inputs. As a matter of fact, evaluation of the output at each realization requires a complete solution of the underlying PDE with expensive computational cost, making the direct approach of solving PDEs and evaluating outputs for a large number of realizations sampled from the high-dimensional probability space prohibitive [187, 148]. Secondly, the topological and geometrical properties of the limit state surface defined by a critical failure value play a crucial role in the design of appropriate computational methods, which bring a significant difficulty for accurate approximation when the surface lacks smoothness and/or features possible discontinuity, disconnectivity and singularity [69, 166]. At third, it is a common challenge to perform effective and efficient sampling in the probability space in order to evaluate an extreme failure probability of some rare event with high consequence [29, 120]. In this chapter, we are mainly treating the first two difficulties of computational complexity and accurate approximation. The third one will be tackled in a forthcoming research work by combining the computational strategies developed in this work with suitable sampling techniques, such as importance sampling with efficient adaptive procedure guided by sensitivity analysis [38].

To avoid solving the full PDE many times, efficient computational methods have been designed for constructing accurate and inexpensive surrogate models of the original PDEs. However, it has been noticed [121] that no matter how accurate the surrogate model is, the resulting failure probability evaluated via the surrogate model can be incorrect due to the nonsmoothness of the limit state surface.

Reference for this chapter:

P. Chen and A. Quarteroni. Accurate and efficient evaluation of failure probability for partial differential equations with random input data. Computer Methods in Applied Mechanics and Engineering, 267(0):233–260, 2013

For instance, when approximating a function by either projective or interpolative methods based on prescribed dictionary bases, the approximation error of the surrogate function can converge to zero when the number of basis functions increases. Nevertheless, the surrogate function may oscillate about the original function due to jump discontinuity because of the Gibbs phenomenon, producing therefore erroneous failure probability estimates if the discontinuity lies in the limit surface space. To deal with this problem, a hybrid approach consisting in combining the outputs computed from both the surrogate and the original models was proposed in [121]. The idea is that whenever the surrogate output is close enough to the critical value controlled by a threshold parameter, one uses the original output computed by solving the full PDE. However, the threshold parameter of the proposed direct algorithm as well as the step size and the stopping criterion of the iterative algorithm are exposed to arbitrariness, potentially leading to a biased failure probability estimate or less efficient surrogate model. When it comes to high-dimensional problems, most of the surrogate models constructed by projective and interpolative approximation based on prescribed dictionary bases may become poorly accurate. In real-world engineering problems, most of the high-dimensional stochastic problems reside in a relatively low-dimensional stochastic manifold named universality phenomenon [196], which provides rationality for the application of model order reduction techniques to reconstruct the low-dimensional manifold of the stochastic solution based on a series of snapshots, i.e., solutions at some representative samples.

In this chapter, we develop a hybrid and goal-oriented adaptive computational strategy based on the certified reduced basis method introduced in chapter 1 and 2 in order to efficiently and accurately evaluate the failure probability of a PDE with random inputs. In dealing with high-dimensional random input problems, we introduce a reduced basis approximation space constructed by a goal-oriented greedy algorithm. An accurate and sharp a posteriori error bound for the approximate output is employed for the construction, which results in a limited number of reduced bases when the output lives in a low-dimensional manifold. For an accurate evaluation of the failure probability when the limit state surface is nonsmooth, we design a hybrid computational approach. The idea is to use the surrogate model constructed by the reduced basis method to evaluate a surrogate output. If the latter can be determined to be a failure or a success by a certification indicator, we use this certified output and do not need to solve the full PDE. Otherwise, we solve the full PDE and evaluate the truth output in order to judge if it is a failure. Since the sample of the uncertified output is very near to or lives in the limit state surface, we enrich the reduced basis space by the solution at this sample to build a more accurate surrogate model, especially for samples near the limit state surface.

For efficient application of the computational strategy to more general PDE models, we present some generalizations of our technique, including the adapted primal-dual approach, POD-greedy sampling algorithm, and empirical interpolation algorithm for efficient decomposition of nonaffine functions.

This chapter is organized as follows. In section 4.1 we state the problem of failure probability evaluation based on a benchmark model, followed by section 4.2 for the development of the hybrid and goal-oriented adaptive algorithm based on the reduced basis method. We extend the proposed methods to more general PDE models in section 4.3 and carry out a series of experiments to compare and illustrate the advantages of our methods in section 4.4. Summary on the advantages of this algorithm and further development is provided in the last section 4.5.

4.1 Problem statement

We first present the generic formulation of failure probability of some quantities of interest depending on the stochastic solution of a given elliptic partial differential equation (PDE) with random inputs as introduced in the preliminary chapter, and we will later extend our proposed methods to more general PDE models in section 4.3.

Recall that $u : \bar{D} \times \Omega \to \mathbb{R}$ stands for the solution of the stochastic elliptic problem (39). Suppose that it depends only on a given finite random vector $y(\omega) = (y_1(\omega), \ldots, y_K(\omega)) : \Omega \to \Gamma = \prod_{k=1}^{K} \Gamma_k \subset \mathbb{R}^K$ with probability density function $\rho : \Gamma \to \mathbb{R}$. In the context of risk prediction or reliability analysis, without loss of generality we are interested in computing the following failure probability [121]:

$$P_0 := P(\omega \in \Omega : s(u(y(\omega))) < s_0) = \int_\Gamma \mathscr{X}_{\Gamma_0}(y)\rho(y)dy, \tag{4.1}$$

where s is a functional of the stochastic solution u, conventionally called limit state function or performance function in reliability problem; s_0 is a critical value defining the failure domain $\Gamma_0 := \{y \in \Gamma : s(u(y)) < s_0\}$ and the characteristic function $\mathscr{X}_{\Gamma_0}$ is defined as

$$\mathscr{X}_{\Gamma_0}(y) = \begin{cases} 1 & \text{if } y \in \Gamma_0, \\ 0 & \text{if } y \notin \Gamma_0. \end{cases} \tag{4.2}$$

The Monte Carlo method can be straightforwardly applied to solve the stochastic system as well as evaluate the failure probability [72]. The idea is to generate a series of samples $y^m \in \Gamma, m = 1, 2, \ldots, M$ according to the probability density function $\rho(y)$, solve the underlying PDE problem at each sample to get the stochastic solution $u(y^m)$, compute the output of interest $s(u(y^m))$ and evaluate the failure probability (Monte Carlo failure probability, denoted as P_0^m) by taking the average as

$$P_0^m = \frac{1}{M} \sum_{m=1}^{M} \mathscr{X}_{\Gamma_0}(y^m). \tag{4.3}$$

This method requires no additional effort for modification of the deterministic solver of the PDE. However, in practical application it is too expensive because one PDE has to be fully solved for each of a large number of samples, leading in general to a prohibitive computational cost. Several accelerated variations of Monte Carlo method have been developed and used in evaluation of failure probability, such as quasi Monte Carlo, Latin hypercube sampling, multi-level techniques, to name a few [171, 81, 62].

4.2 Reduced basis methods for evaluation of failure probability

We first present the formula for the reduced basis method in the evaluation of failure probability, then we propose a hybrid approach for evaluation of the failure probability guided by a posteriori error bound. Finally, we present a goal-oriented adaptive reduced basis method for efficient evaluation of the failure probability.

4.2.1 The reduced basis method

We apply the reduced basis method introduced in chapter 1 to first compute the surrogate output s_N with Algorithm 2 and then evaluate the surrogate failure probability by

$$P_0^s = \frac{1}{M} \sum_{m=1}^{M} \mathscr{X}_{\Gamma_0^s}(y^m), \tag{4.4}$$

where the surrogate approximate failure domain is defined as $\Gamma_0^s := \{y \in \Gamma : s_N(y) < s_0\}$. Unfortunately, the surrogate output s_N may lead to an inaccurate failure probability due to the reduced basis approximation error. In fact, this drawback is commonly present for most of the surrogate models.

4.2.2 A hybrid reduced basis method

As already noticed, the Monte Carlo method is an accurate and straightforward approach for evaluation of the failure probability by (4.3), however it is prohibitively expensive as it requires the solution of a large number of PDEs. In contrast, surrogate models built on other methods may improve computational efficiency at the expense of producing incorrect output and thus wrong failure probability estimate. In order to balance the trade-off of computational efficiency and numerical accuracy, a hybrid approach with either direct or iterative algorithms has been developed in [121]. The direct hybrid algorithm predefines a neighborhood region of the critical value by a threshold parameter, then it uses a surrogate model to compute the (surrogate) outputs at samples outside that region and directly solves the PDEs to evaluate the (direct) outputs at samples inside the region. However, the choice of the threshold value depends crucially on the accuracy of the surrogate model, which is not provided in general. On the other hand, the iterative hybrid algorithm replaces some surrogate output closest to the critical value by direct outputs and conduct the replacement iteratively until meeting a posteriori error tolerance. This algorithm does not need to choose the value of a threshold parameter but the accuracy of the failure probability estimate is again affected by the unknown error of the surrogate model. To improve on this, we propose a hybrid reduced basis method certified by a posteriori error bound, achieving both the computational efficiency and the numerical accuracy.

Since the approximation error of the output at sample y can be bounded by (1.27), we can define the certified surrogate failure domain

$$\Gamma_0^c := \{y \in \Gamma : s_N(y) < s_0, \triangle_N^s(y) < s_0 - s_N(y)\}, \tag{4.5}$$

and the uncertified surrogate failure domain

$$\Gamma_0^u := \{y \in \Gamma : \triangle_N^s(y) \geq |s_0 - s_N(y)|\}. \tag{4.6}$$

Whenever the sample y falls in the certified surrogate failure domain Γ_0^c, we have

$$s(y) = (s(y) - s_N(y)) + s_N(y) \leq \triangle_N^s(y) + s_N(y) < s_0 - s_N(y) + s_N(y) = s_0, \tag{4.7}$$

so that any sample $y \in \Gamma_0^c$ also falls in the original failure domain Γ_0. As for the sample in uncertified failure domain $y \in \Gamma_0^u$, we compute a real output $s(y) = s(u(y))$ from the solution $u(y)$ by fully solving the PDE (46). Thus, the hybrid failure domain is defined as

$$\Gamma_0^h := \Gamma_0^c \cup \left(\Gamma_0^u \cap \{y \in \Gamma : s(y) < s_0\}\right), \tag{4.8}$$

and the hybrid failure probability is evaluated by

$$P_0^h = \frac{1}{M} \sum_{m=1}^{M} \mathscr{X}_{\Gamma_0^h}(y^m). \tag{4.9}$$

By construction, we have that the evaluation of the hybrid failure probability is cheap thanks to the use of the surrogate model and accurate, as it is equal to the Monte Carlo failure probability, $P_0^h = P_0^m$.

In dealing with high-dimensional problems, we usually apply an iterative algorithm for Monte Carlo sampling with an increasing number of samples to enhance computational efficiency on the one hand and provide a posteriori error estimate for the Monte Carlo evaluation on the other. The following Algorithm 4 describes the hybrid reduced basis method.

Algorithm 4 Iterative algorithm for the hybrid reduced basis method

procedure OFFLINE CONSTRUCTION:
    Construct a reduced basis space X_N by Algorithm 2.
end procedure

Initialize tolerance ϵ_{tol} and a posteriori error $e_1^p = 2\epsilon_{tol}$, choose the number of initial samples M, adaptive size parameter β as well as a maximum iteration number I_{max};
procedure ITERATIVE EVALUATION:
    for $i = 1, \ldots, I_{max}$ **do**
        sample Ξ_M with $|\Xi_M| = M$, pre-compute and store $\alpha_{LB}(y)$, $y \in \Xi_M$ by SCM [102];
        compute surrogate output $s_N(y)$ and the error bound $\triangle_N^s(y)$ by (1.27) for $\forall\, y \in \Xi_M$;
        evaluation the failure probability $P_0^{h,i}$ by formula (4.9);
        if $i > 1$ **then**
            compute the a posteriori error for failure probability $e_i^p = |P_0^{h,i} - P_0^{h,i-1}|$;
            if $e_i^p < \epsilon_{tol}$ **then**
                $I_{max} = i$;
                return ;
            end if
        end if
        increase the number of sample size by setting $M = \beta^{i+1} M$;
    end for
end procedure

4.2.3 A goal-oriented adaptive reduced basis method

In order to avoid too many direct solves of the full underlying PDE, we need to increase the portion of the samples in the certified surrogate failure domain, which in turn requires using a more accurate surrogate model constructed with more reduced basis functions. However, the computational cost of both the offline construction and the online evaluation of the reduced basis method critically depends on the number of reduced basis functions, suggesting therefore the use of a low number of reduced basis functions, especially for high-dimensional problems. In addition, when the surrogate output is far from the critical value, a rather crude surrogate approximation with a small number of reduced basis functions would be sufficient as long as the a posteriori error bound for the approximation error of the output is smaller than the distance between the surrogate output and the critical value. To take full advantage of the reduced basis approximation and a posteriori error bound, we develop a goal-oriented adaptive strategy to construct a surrogate model with fine approximation of the output manifold close to the limit state surface $\{y \in \Gamma : s(y) = s_0\}$ and coarse approximation of the output manifold far away from it.

Goal-oriented adaptive strategies have been developed in many contexts (e.g. [155, 156]). For their application in the construction of surrogate models, we first run the Algorithm 2 for the reduced basis method with a relatively small training set Ξ_{train} and large tolerance ε_{tol} as stopping criteria. Given any new sample set Ξ_M with M samples, we compute the surrogate outputs s_N and the associated error bounds $\triangle_N^s$, from which we define the following adaptive criteria

$$\triangle_N^a(y) = \frac{\triangle_N^s(y)}{|s_N(y) - s_0|} \quad \forall\, y \in \Xi_M. \tag{4.10}$$

We remark that in some extreme case, there may exist too many samples for which s_N is too close to s_0. In this case, we can pick the sample such that the condition $|s_N - s_0| \geq \epsilon_s$ with $\epsilon_s > 0$ very small is also satisfied.

We apply again the greedy algorithm to select the most mismatching sample

$$y^{N+1} = \arg\max_{y\in\Xi_M} \triangle_N^a(y) \text{ such that } \triangle_N^a(y) \geq 1, \tag{4.11}$$

and enrich the reduced basis space by $X_{N+1} = X_N \oplus \text{span}\{\zeta^{N+1}\}$ where ζ^{N+1} is the orthonormalized version of the solution $u(y^{N+1})$. We carry out the sample procedure of reduced basis construction with $N = N+1$ until $\triangle_N^a(y^{N+1}) < 1$. Then we compute the failure probability by formula (4.4), which is accurate (the same as Monte-Carlo evaluation) since $\triangle_N^s(y) < |s_N(y) - s_0| \; \forall\, y \in \Xi_M$.

Algorithm 5 combines the goal-oriented adaptive strategy with the iterative scheme for Monte Carlo evaluation of failure probability. As a byproduct of the adaptive construction of the reduced basis space, the failure probability is computed asymptotically based on a posteriori error in Algorithm 4 and 5. In order to further quantify the precision of the failure probability, one may also provide the binomial confidence interval for the failure probability [187], e.g. the normal approximation interval $P_0 \pm z_{1-\alpha_e/2}\sqrt{P_0(1-P_0)/M}$, being P_0 the failure probability computed by either the hybrid or the adaptive algorithm with M Monte Carlo samples, $z_{1-\alpha_e/2}$ a percentile of a standard normal distribution associated with a prescribed tail probability α_e.

4.2.4 Remarks on approximation error and computational cost

The approximation error of the failure probability by the three different approaches described above can be generally split into the one arising from the surrogate models and the other from Monte Carlo method. In the first approach (described in section 4.2.1), the approximation error of the surrogate model may lead to a large error or even wrong evaluation of the failure probability due to the discontinuous or singular properties of the limit state surface, while in the last two approaches (described in section 4.2.2 and 4.2.3), the contribution of the approximation error from surrogate models is null and the Monte Carlo approximation error takes full responsibility with a slow algebraic decaying rate $M^{-1/2}$.

As for computational cost, the first approach is the cheapest one as it does not necessitate to solve a full PDE in the evaluation procedure once the offline construction is finished. In contrast, the hybrid approach is relatively expensive, as it requires to solve the full PDE whenever the a posteriori error bound is larger than the distance between the surrogate output and the critical value. The goal-oriented adaptive approach is much cheaper than the hybrid one since it starts from a rather crude reduced basis construction and replaces many direct outputs in the hybrid approach by surrogate outputs based on adaptively enriched reduced basis space. Moreover, it might be even cheaper than the first approach if its total offline construction is less expensive than that of the first approach.

4.3 Extension to more general PDE models

The development of both hybrid and goal-oriented adaptive reduced basis methods is based on the benchmark linear elliptic coercive affine PDE with random inputs (39), which is assumed to be compliant in the output, time independent, affine in the random inputs and coercive. In this section, we remove these limitations and extend the proposed methods to more general PDE models. The key elements in the extension are to accurately compute cheap, reliable and sharp a posteriori error bound for the approximation error of the output and efficiently decompose the approximation procedure into the offline construction stage and the online evaluation stage. We remark that most of the techniques we are using have been well studied for the development and application of the reduced basis method [163], and we briefly summarize them with specific application in the context of failure probability computation. In addition, most of the proposed algorithms can be extended to the case of multiple functional outputs of the solution field. In fact, for a small number of independent functional outputs,

Algorithm 5 Iterative algorithm for goal-oriented adaptive reduced basis method

procedure OFFLINE CONSTRUCTION:
  Construct a crude reduced basis space X_N by Algorithm 2.
end procedure

Initialize tolerance ϵ_{tol} and a posteriori error $e_1^p = 2\epsilon_{tol}$, choose the number of initial samples M, adaptive size parameter β as well as a maximum iteration number I_{max};
procedure ADAPTIVE CONSTRUCTION:
  for $i = 1, \ldots, I_{max}$ **do**
    sample Ξ_M with $|\Xi_M| = M$, pre-compute and store $\alpha_{LB}(y), y \in \Xi_M$ by SCM [102];
    compute surrogate outputs $s_N(y)$ and adaptive criteria $\triangle_N^a(y)$ by (4.10) for $\forall y \in \Xi_M$;
    choose adaptive sample $y^{N+1} = \arg\max_{y\in\Xi_M} \triangle_N^a(y)$;
    while $\triangle_N^a(y^{N+1}) \geq 1$ **do**
      augment the sample space $S_{N+1} = S_N \cup \{y^{N+1}\}$;
      solve problem (46) at y^{N+1} to obtain $u(y^{N+1})$;
      orthonormalize the solution $u(y^{N+1})$ by Gram-Schmidt process to get ζ_{N+1};
      augment the reduced basis space $X_{N+1} = X_N \oplus \text{span}\{\zeta_{N+1}\}$;
      compute and store $A_q(\zeta_{N+1}, \zeta_n), A_q(\zeta_n, \zeta_{N+1}), 1 \leq q \leq Q_a, 1 \leq n \leq N+1$ and $F_q(\zeta_{N+1})$, $1 \leq q \leq Q_f$;
      compute and store $(\mathscr{C}_q, \mathscr{C}_{q'})_X, (\mathscr{C}_q, \mathscr{L}_p^{N+1})_X, (\mathscr{L}_p^{N+1}, \mathscr{L}_{p'}^n)_X, (\mathscr{L}_p^n, \mathscr{L}_{p'}^{N+1})_X, 1 \leq q, q' \leq Q_f$, $1 \leq p, p' \leq Q_a, 1 \leq n \leq N+1$;
      set N = N + 1;
      compute $s_N(y)$ and $\triangle_N^a(y)$ by (4.10) $\forall y \in \Xi_M$;
      choose adaptive sample $y^{N+1} = \arg\max_{y\in\Xi_M} \triangle_N^a(y)$;
    end while
    evaluation the failure probability $P_0^{s,i}$ by formula (4.4);
    if $i > 1$ **then**
      compute the a posteriori error for failure probability $e_i^p = |P_0^{s,i} - P_0^{s,i-1}|$;
      if $e_i^p < \epsilon_{tol}$ **then**
        $I_{max} = i$;
        return ;
      end if
    end if
    increase the number of sample size by setting $M = \beta^{i+1} M$;
  end for
end procedure

we may treat each of them separately as in the case of a single functional output. When there are many coupled functional outputs, a common coarse reduced basis space is more convenient to be used in combination with suitable refinement with respect to each functional output in order to keep the total computational cost under control.

4.3.1 Noncompliant problems

When the output is compliant, i.e., $s(y) \equiv s(u(y); y) = F(u(y); y), y \in \Gamma$, we obtain a posteriori error bound $\triangle_N^s(y)$ being quadratic with respect to the residual norm $||\hat{e}(y)||_X$. However, when the output is noncompliant in more general conditions, i.e.

$$s(y) \equiv s(u(y); y) = L(u(y); y), \tag{4.12}$$

where $L : X \to \mathbb{R}$ is a bounded and affine functional, $L \neq F$, we have the following upper bound

$$|s(y) - s_N(y)| \leq ||L(y)||_{X'} ||u(y) - u_N(y)||_X \leq \frac{1}{\alpha(y)} ||L(y)||_{X'} ||\hat{e}(y)||_X, \tag{4.13}$$

which depends only linearly on the residual norm $||\hat{e}(y)||_X$. Moreover, evaluation of the dual norm of the functional $||L(y)||_{X'}$ is expensive and might not be uniformly bounded in the probability domain Γ. In order to seek an effective and efficient a posteriori error bound for the output approximation error, we apply the primal-dual computational strategy [178, 158, 163] by solving an additional problem, known as the dual problem associated to the functional L: $\forall\, y \in \Gamma$ find the dual variable $\psi(y) \in X$ such that

$$A(v, \psi(y); y) = -L(v; y) \quad \forall\, v \in X. \tag{4.14}$$

By the same reduced basis approximation procedure as in section 4.2.1, we construct the reduced basis space for the approximation of the dual variable ψ as $X^{du}_{N_{du}} := \text{span}\{\zeta^{du}_1, \ldots, \zeta^{du}_{N_{du}}\}$ where $\zeta^{du}_n, 1 \leq n, \leq N_{du}$ are determined via orthonormalization from the solution $\{\psi(y^n), 1 \leq n \leq N_{du}\}$ (at suitable values of $y^n, 1 \leq n \leq N_{du}$), then the reduced basis solution $\psi_{N_{du}}(y)$ at sample $y \in \Gamma$ is obtained by solving the reduced system

$$A(v, \psi_{N_{du}}(y); y) = -L(v; y) \quad \forall\, v \in X^{du}_{N_{du}}. \tag{4.15}$$

Let us denote the primal reduced basis space as $X^{pr}_{N_{pr}} := \text{span}\{\zeta^{pr}_1, \ldots, \zeta^{pr}_{N_{pr}}\}$ and rewrite the reduced system for the primal reduced basis solution $u_{N_{pr}}$ as

$$A(u_{N_{pr}}(y), v; y) = F(v; y) \quad \forall\, v \in X^{pr}_{N_{pr}}. \tag{4.16}$$

Furthermore, let us define the primal residual and dual residual respectively as

$$R^{pr}(v; y) = F(v; y) - A(u_{N_{pr}}(y), v; y) \quad \text{and} \quad R^{du}(v; y) = -L(v; y) - A(v, \psi_{N_{du}}(y); y). \tag{4.17}$$

By solving the primal and dual reduced system, we can evaluate the noncompliant output by

$$s_N(y) = L(u_{N_{pr}}(y)) - R^{pr}(\psi_{N_{du}}(y); y). \tag{4.18}$$

The following lemma provides an efficient a posteriori error bound for the output [178, 158, 163].

Lemma 4.3.1 *The approximation error on the output $|s(y) - s_N(y)|$ is bounded from above by the following a posteriori error bound $\Delta^s_N(y)$*

$$|s(y) - s_N(y)| \leq \Delta^s_N(y) := \frac{||R^{pr}(\cdot; y)||_{X'} ||R^{du}(\cdot; y)||_{X'}}{\alpha_{LB}(y)} \quad \forall\, y \in \Gamma, \tag{4.19}$$

where $||R^{pr}(\cdot; y)||_{X'}$ and $||R^{du}(\cdot; y)||_{X'}$ are the dual norms of the primal and dual residuals, respectively.

Remark 4.3.1 *Besides converging faster, the primal-dual computational strategy does not require the computation of the dual norm $||L(y)||_{X'}, \forall\, y \in \Gamma$. On their turn, the dual norms $||R^{pr}(\cdot; y)||_{X'}$ and $||R^{du}(\cdot; y)||_{X'}$ can be efficiently evaluated by the offline-online computational decomposition.*

As for the evaluation of failure probability in noncompliant problems, the reduced basis method in Algorithm 2 remains the same as in the compliant case, and the hybrid reduced basis method in Algorithm 4 is essentially the same as in compliant problems except for the replacement of a posteriori error bound (4.19). In the goal-oriented adaptive Algorithm 5, we enrich simultaneously both the primal and the dual reduced basis spaces governed by the a posteriori error bound (4.19) in order to gain more computational efficiency for the evaluation of failure probability in noncompliant problems. In this way, we recover the “square effect" in the convergence of the error.

4.3.2 Unsteady problems

If the state variable depends not only on the spatial variable $x \in D$ but also on the temporal variable $t \in I \equiv [0, T]$, we have to face an unsteady PDE; a suitable time discretization needs to be operated on both the offline construction of reduced basis space and the online evaluation of the output. For the sake of simplicity [87, 163], we consider the following parabolic problem in semi-weak formulation: find $u(y) \in L^2(I; X) \cap C^0(I; L^2(D))$ such that it holds, almost surely

$$M\left(\frac{\partial u}{\partial t}(t; y), v; y\right) + A(u(t; y), v; y) = g(t)F(v; y) \quad \forall v \in X, \tag{4.20}$$

subject to initial condition $u(0; y) = u_0 \in L^2(D)$. Here, $g \in L^2(I)$ is a time dependent control function; X is a spatial approximation space as defined in section 4.2.1, e.g., a finite element space; the bilinear form A and linear form F are defined as in the elliptic problem, and the bilinear form M is assumed to be uniformly continuous and coercive and featuring the following affine expansion

$$M(w, v; y) = \sum_{q=1}^{Q_m} \Theta_q^m(y) M_q(w, v) \quad \forall w, v \in X. \tag{4.21}$$

Using (without loss of generality) the backward Euler scheme for time discretization, we find at every time step

$$M(u^i(y), v; y) + \triangle t A(u^i(y), v; y) = \triangle t g(t^i) F(v; y) + M\left(u^{i-1}(y), v; y\right) \quad \forall v \in X, \tag{4.22}$$

subject to the initial condition $u(t^0; y) = u_0$, where $\triangle t$ is the time step size, $u^i(y) \simeq u(t^i; y), 0 \le i \le I_T \equiv T/\triangle t$. We remark that we don't take into account the time discretization error for the sake of simplicity. We consider a compliant output $s(t^i; y) = F(u^i(y); y), 1 \le i \le I_T, y \in \Gamma$. For noncompliant output, we apply the primal-dual computational strategy presented in section 4.3.1; see unsteady problems with more general outputs in [87, 173, 145]. A reduced problem associated to (4.22) can be formulated as: find $u_N^i(y) \in X_N, 1 \le i \le I_T$ such that

$$M(u_N^i(y), v; y) + \triangle t A(u_N^i(y), v; y) = \triangle t g(t^i) F(v; y) + M\left(u_N^{i-1}(y), v; y\right) \quad \forall v \in X_N, \tag{4.23}$$

where the reduced basis space X_N can be constructed by a POD-greedy sampling algorithm governed by cheap a posteriori error bound as well as an efficient offline-online computational decomposition procedure, which are presented in the following subsections respectively.

A POD-greedy algorithm

In unsteady problems, the samples for the construction of the reduced basis space involve not only the random samples $y \in \Xi_{train} \subset \mathbb{R}^K$ in multiple dimensions but also the temporal samples $t^i \in I \subset \mathbb{R}, 1 \le i \le I_T$ in one dimension. A pure greedy sampling algorithm in both probability space and temporal space has been demonstrated inefficient and resulting in occasional infinite loop [93]. A POD-greedy algorithm, based on POD selection in temporal space and greedy selection in probability space, has been effectively used in [93, 146] for tackling these difficulties. A general formulation for POD is stated as follows: given a training set X_{train} with n_{train} elements, the function $X_M = \mathrm{POD}(X_{train}, M)$ leads to an optimal subset $X_M \subset \mathrm{span}\{X_{train}\}$ with M bases such that

$$X_M = \arg \inf_{Y_M \subset \mathrm{span}\{X_{train}\}} \left(\frac{1}{n_{train}} \sum_{v \in X_{train}} \inf_{w \in Y_M} ||v - w||_X^2 \right)^{1/2}. \tag{4.24}$$

In practice, we solve the eigenvalue problem $C\zeta = \lambda\zeta$, where the correlation matrix C is assembled by the weighted correlation of the elements $v_n \in X_{train}, 1 \le n \le n_{train}$ as

$$C_{mn} = \frac{1}{n_{train}}(v_m, v_n)_X, 1 \le m, n \le n_{train}, \tag{4.25}$$

and the subset $X_M = \text{span}\{\zeta_m, 1 \le m \le M\}$ where $\zeta_m, 1 \le m \le M$ are the orthonormal eigenfunctions corresponding to the M largest eigenvalues. Provided a tolerance ε_{pod} is given, we can also redefine the function $X_M = \text{POD}(X_{train}, \varepsilon_{pod})$ such that the sum of $n_{train} - M$ smallest eigenvalues is smaller than ε_{tol}. The POD-greedy algorithm for the construction of reduced basis space in unsteady problems is recalled in Algorithm 6 [163, 93, 146].

Algorithm 6 A POD-greedy algorithm

Initialize a random sample $y^* \in \Xi_{train}$, the tolerances ε_{tol} and ε_{pod}, an empty reduced basis space Y as well as a maximum number of reduced basis functions N_{max}, set $N = 0$;
procedure ITERATIVE CONSTRUCTION:
  while $N \le N_{max}$ **do**
    solve the parabolic problem (4.22) at sample y^* and time $t^i, 1 \le i \le I_t$;
    compute $X_{M_1} = \text{POD}(\{u^i(y^*), 1 \le i \le I_t\}, \varepsilon_{pod})$;
    enrich the reduced basis space $Y = Y \cup X_{M_1}$;
    update the number of reduced basis functions $N = N + M_2$, where $M_2 \le M_1$;
    construct the reduced basis space $X_N = \text{POD}(Y, N)$;
    choose sample $y^* = \arg\max_{y \in \Xi_{train}} \triangle_N^s(T; y)$ by greedy algorithm;
    if $\triangle_N^s(T; y^*) \le \varepsilon_{tol}$ **then**
      $N_{max} = N$;
      return ;
    end if
  end while
end procedure

We underline that in Algorithm 6 the step integer M_1 is controlled by the tolerance of the internal POD algorithm, offering flexibility in choosing the number of reduced basis functions from the elements $u^i(y^*), 1 \le i \le I_t$, and M_2 is chosen to be smaller than M_1 in order to minimize duplication of the reduced basis functions (several options/combinations are possible). The random sample y^*, which might be the same in different iteration steps, is chosen by greedy algorithm governed by cheap and sharp a posteriori error bound $\triangle_N^s(T; y), y \in \Gamma$ constructed in the following sections.

Construction of a posteriori error bound

We follow the procedure in section 1.3.3 of chapter 1 to derive an a posteriori error bound for the parabolic problem (4.22). Firstly, we define the reduced residual for $1 \le i \le I_t$,

$$R^i(v; y) = g(t^i)F(v; y) - \frac{1}{\triangle t}M(u_N^i(y) - u_N^{i-1}(y), v; y) - A(u_N^i(y), v; y) \quad \forall v \in X_N. \tag{4.26}$$

By Riesz representation theorem [68], we have a unique function $\hat{e}^i(y) \in X, 1 \le i \le I_t$ such that $(\hat{e}^i(y), v)_X = R^i(v; y)$ and $||\hat{e}^i(y)||_X = ||R^i(\cdot; y)||_{X'}, 1 \le i \le I_t$. Furthermore, it can be proven that the reduced basis approximation error for the output is bounded by [87]

$$|s(t^i; y) - s_N(t^i; y)| \le \triangle_N^s(t^i; y) := \frac{1}{\alpha_{LB}(y)} \sum_{i'=1}^{i} ||\hat{e}^{i'}(y)||_X^2, 1 \le i \le I_t. \tag{4.27}$$

Offline-online computational decomposition

By expansion of the reduced basis solution at time $t^i, 1 \le i \le I_t$ in the reduced basis functions

$$u_N^i(y) = \sum_{m=1}^{N} u_{Nm}^i(y)\zeta_m, \tag{4.28}$$

we have the reduced basis problem by Galerkin projection in (4.22) as: find $u_{Nm}^i(y), 1 \le m \le N, 1 \le i \le I_t$ such that

$$\begin{aligned} &\sum_{m=1}^{N}\sum_{q=1}^{Q_m} \Theta_q^m(y) M_q(\zeta_m,\zeta_n) u_{Nm}^i(y) + \triangle t \sum_{m=1}^{N}\sum_{q=1}^{Q_a} \Theta_q^a(y) A_q(\zeta_m,\zeta_n) u_{Nm}^i(y) \\ &= \triangle t g(t^i) \sum_{q=1}^{Q_f} \Theta_q^f(y) F_q(\zeta_n) + \sum_{m=1}^{N}\sum_{q=1}^{Q_m} \Theta_q^m(y) M_q(\zeta_m,\zeta_n) u_{Nm}^{i-1}(y) \quad 1 \le n \le N, \end{aligned} \tag{4.29}$$

where the matrices $M_q(\zeta_m,\zeta_n), 1 \le q \le Q_m, 1 \le m,n \le N$, $A_q(\zeta_m,\zeta_n), 1 \le q \le Q_a, 1 \le m,n \le N$ and the vectors $F_q(\zeta_n), 1 \le q \le Q_f, 1 \le n \le N$ can be pre-computed and stored in the offline construction stage. In the online evaluation stage, we only need to assemble and solve a $N \times N$ system (4.29) to get the solution $u_N^i(y)$ and evaluate the output by NQ_f operations

$$s_N(t^i;y) = F(u_N^i(y);y) = \sum_{n=1}^{N} \left(\sum_{q=1}^{Q_f} \Theta_q^f(y) F_q(\zeta_n) \right) u_{Nn}^i(y). \tag{4.30}$$

As for the evaluation of the error bound (4.27), we substitute the reduced basis solution (4.28) in the residual (4.26) and compute the residual norm $||\hat{e}^i(y)||_X$ by assembling

$$\begin{aligned} ||\hat{e}^i(y)||_X^2 = {} & g^2(t^i) \sum_{q}^{Q_f}\sum_{q'}^{Q_f} \Theta_q^f(y)\Theta_{q'}^f(y)(\mathscr{C}_q,\mathscr{C}_{q'})_X \\ & + 2\frac{g(t^i)}{\triangle t} \sum_{n=1}^{N}\sum_{q=1}^{Q_f}\sum_{q'=1}^{Q_m} \Theta_q^f(y)\Theta_{q'}^m(y)(\mathscr{C}_q,\mathscr{M}_{q'}^n)_X \varphi_{Nn}^i(y) \\ & + 2g(t^i) \sum_{n=1}^{N}\sum_{q=1}^{Q_f}\sum_{q'=1}^{Q_a} \Theta_q^f(y)\Theta_{q'}^a(y)(\mathscr{C}_q,\mathscr{L}_{q'}^n)_X u_{Nn}^i(y) \\ & + \frac{1}{\triangle t^2} \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q_m}\sum_{q'=1}^{Q_m} \Theta_q^m(y)\Theta_{q'}^m(y)\varphi_{Nn}^i(y)(\mathscr{M}_q^n,\mathscr{M}_{q'}^{n'})_X \varphi_{Nn'}^i(y) \\ & + 2\frac{1}{\triangle t} \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q_m}\sum_{q'=1}^{Q_a} \Theta_q^m(y)\Theta_{q'}^a(y)\varphi_{Nn}^i(y)(\mathscr{M}_q^n,\mathscr{L}_{q'}^{n'})_X u_{Nn'}^i(y) \\ & + \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q_a}\sum_{q'=1}^{Q_a} \Theta_q^a(y)\Theta_{q'}^a(y) u_{Nn}^i(y)(\mathscr{L}_q^n,\mathscr{L}_{q'}^{n'})_X u_{Nn'}^i(y), \end{aligned} \tag{4.31}$$

where $\varphi_{Nn}^i(y) = u_{Nn}^i(y) - u_{Nn}^{i-1}(y)$, $\mathscr{C}_q, 1 \le q \le Q_f$ and $\mathscr{L}_q^n, 1 \le q \le Q_a, 1 \le n \le N$ are defined as in the elliptic case, and $\mathscr{M}_q^n, 1 \le q \le Q_m, 1 \le n \le N$ are defined such that $(\mathscr{M}_q^n, v)_X = -M_q(\zeta_n, v), \forall v \in X$, which are pre-computed and stored in the offline stage. In the online stage, we only need to assemble (4.31) by $O((Q_f + NQ_m + NQ_a)^2)$ operations, which is very efficient because the values $Q_f, Q_m, Q_a, N \ll \mathscr{N}$ are small.

Methods for the evaluation of failure probability in unsteady problems is not different than those used in the elliptic problems. In particular, we can use the same goal-oriented adaptive and iterative procedure of Algorithm 5 with the POD-greedy sampling Algorithm 6 governed by the a posterior error

bound (4.27).

4.3.3 Nonaffine problems

The affine assumption is crucial for an effective offline-online computational decomposition. In the case of a more general nonaffine random field denoted by $g(x,y)$, we apply the empirical interpolation method as introduced in chapter 3 to approximate the random field $g(x,y)$ by finite affine terms, given by (3.2). Note that for pointwise evaluation, we assume a constant weight function $w = 1$ in Algorithm 3.

Global a posteriori error estimate

Let us now extend the affine Assumption 0.3 to more general nonaffine random fields for both the diffusion coefficient a and the force term f in (39). By the empirical interpolation introduced in chapter 3, we obtain the following affine decomposition

$$a \approx a_{Q_a} \equiv \mathscr{I}_{Q_a}[a] = \sum_{q=1}^{Q_a} \Theta_q^a(y) a_q(x) \text{ and } f \approx f_{Q_f} \equiv \mathscr{I}_{Q_f}[f] = \sum_{q=1}^{Q_f} \Theta_q^f(y) f_q(x). \tag{4.32}$$

For the reduced basis approximation with affine decomposition of the nonaffine random inputs, we state the following two lemmas for global a posteriori reduced basis approximation error estimate of the solution and the output.

Lemma 4.3.2 *Suppose the approximation by affine decomposition (4.32) results in a high-fidelity approximate solution u_Q and a reduced basis solution $u_{Q,N}$. The following a posteriori error bound for the reduced basis approximation error holds*

$$||u(y) - u_{Q,N}(y)||_X \leq \mathscr{E}_Q^u(y) + \Delta_N^u(y), \tag{4.33}$$

where Δ_N^u is the a posteriori error bound for the reduced basis approximation as defined in (1.26) of chapter 1, $\mathscr{E}_Q$ the error due to the affine approximation of the data a and f, defined as

$$\mathscr{E}_Q^u(y) := \frac{C_1}{\alpha_{LB}(y)} ||f(y) - f_{Q_f}(y)||_{L^\infty(D)} + \frac{C_1 C_2}{\alpha_{LB}^2(y)} ||a(y) - a_{Q_a}(y)||_{L^\infty(D)} ||f_{Q_f}(y)||_{L^\infty(D)}, \tag{4.34}$$

C_1, C_2 two constants bounded by (4.52) and $\alpha_{LB}(y), y \in \Gamma$ a lower bound of the coercivity constant of the bilinear form (4.38) with respect to the norm $||\cdot||_X$.

Remark 4.3.2 *We remark that instead of using the $L^\infty(D)$-norm, we may use $L^p(D)$-norm for $1 \leq p < \infty$ under the condition that $a, f \in L^p(D)$, which leads to a smaller error estimate.*

Proof The total approximation error can be bounded by the sum of two terms

$$||u(y) - u_{Q,N}(y)||_X \leq ||u(y) - u_Q(y)||_X + ||u_Q(y) - u_{Q,N}(y)||_X, \tag{4.35}$$

the former due to the affine approximation error of the random fields a and f, the latter arising from the reduced basis approximation error, respectively. Using (1.26), we have

$$||u_Q(y) - u_{Q,N}(y)||_X \leq \Delta_N^u. \tag{4.36}$$

Thus, we only need to control the first part with an error bound denoted as

$$||u(y) - u_Q(y)||_X \leq \mathscr{E}_Q^u(y). \tag{4.37}$$

To bound the first term, we consider the weak formulation of the problem (39) with the original random fields a and f as well as the approximate a_{Q_a} and f_{Q_a},

$$(a\nabla u, \nabla v) = (f, v) \quad \forall v \in H_0^1(D) \tag{4.38}$$

and

$$(a_{Q_a}\nabla u_Q, \nabla v) = (f_{Q_f}, v) \quad \forall v \in H_0^1(D), \tag{4.39}$$

respectively. Subtracting (4.39) from (4.38), we have

$$(a\nabla u - a_{Q_a}\nabla u_Q, \nabla v) = (f - f_{Q_f}, v) \quad \forall v \in H_0^1(D), \tag{4.40}$$

which can be transformed by adding and subtracting $a\nabla u_Q$ as

$$(a\nabla(u - u_Q), \nabla v) = (f - f_{Q_f}, v) + ((a_{Q_a} - a)\nabla u_Q, \nabla v) \quad \forall v \in H_0^1(D). \tag{4.41}$$

By taking $v = u - u_Q$ in (4.41) and applying the coercive property, we have

$$l.h.s. \geq \alpha(y)||u(y) - u_Q(y)||_X^2 \geq \alpha_{LB}(y)||u(y) - u_Q(y)||_X^2. \tag{4.42}$$

As for the right hand side of (4.41), we have the following bound by Hölder's inequality,

$$\begin{aligned} r.h.s \leq\ & ||f(y) - f_{Q_f}(y)||_{L^\infty(D)}||u(y) - u_Q(y)||_{L^1(D)} \\ & + ||a(y) - a_{Q_a}(y)||_{L^\infty(D)}||\nabla u_Q(y)||_{L^2(D)}||\nabla(u(y) - u_Q(y))||_{L^2(D)}. \end{aligned} \tag{4.43}$$

By applying Poincaré inequality in L^1-norm [2], we have that

$$||u(y) - u_Q(y)||_{L^1(D)} \leq C_P||\nabla(u(y) - u_Q(y))||_{L^1(D)} \tag{4.44}$$

where $C_P \leq d_D/2$ with d_D standing for the diameter of the domain D. Moreover, we have again by Cauchy-Schwarz inequality the following relation

$$||\nabla(u(y) - u_Q(y))||_{L^1(D)} \leq C_D||\nabla(u(y) - u_Q(y))||_{L^2(D)}, \tag{4.45}$$

where $C_D = \sqrt{|D|}$ with $|D|$ representing the Lebesgue measure of the domain D. By the definition of the norm $||v||_X = \sqrt{(a(\bar{y})\nabla v, \nabla v)}$ at a reference value $\bar{y} \in \Gamma$, we have

$$||\nabla v||_{L^2(D)} \leq C_X||v||_X \quad \forall v \in H_0^1(D), \tag{4.46}$$

where $C_X \leq \sqrt{||1/a(\bar{y})||_{L^\infty(D)}}$. Using the inequalities (4.44), (4.45) and (4.46), we have the following bound for the right hand side (4.43)

$$\begin{aligned} r.h.s \leq\ & C_D C_P C_X||f(y) - f_{Q_f}(y)||_{L^\infty(D)}||u(y) - u_Q(y)||_X \\ & + C_X^2||a(y) - a_{Q_a}(y)||_{L^\infty(D)}||u_Q(y)||_X||u(y) - u_Q(y)||_X. \end{aligned} \tag{4.47}$$

Furthermore, by setting $v = u_Q$ in the weak formulation (4.39), we obtain

$$||u_Q(y)||_X \leq \frac{C_D C_P C_X}{\alpha_{LB}(y)}||f_{Q_f}(y)||_{L^\infty(D)}, \tag{4.48}$$

for which we have used the following coercive property with lower bound $\alpha_{LB}(y) \le \alpha_{Q_a}(y)$

$$(a_{Q_a}\nabla u_Q, \nabla u_Q) \ge \alpha_{Q_a}(y)||u_Q(y)||_X^2 \ge \alpha_{LB}(y)||u_Q(y)||_X^2 \tag{4.49}$$

as well as the following bound by the inequalities (4.44), (4.45) and (4.46)

$$(f_{Q_f}, u_Q) \le ||f_{Q_f}(y)||_{L^\infty(D)}||u_Q(y)||_{L^1(D)} \le C_D C_P C_X||f_{Q_f}(y)||_{L^\infty(D)}||u_Q(y)||_X. \tag{4.50}$$

A combination of (4.47) and (4.48) leads to the following bound for the right hand side of (4.41)

$$\begin{aligned} r.h.s \le\ & C_D C_P C_X||f(y) - f_{Q_f}(y)||_{L^\infty(D)}||u(y) - u_Q(y)||_X \\ & + \frac{C_D C_P C_X^3}{\alpha_{LB}(y)}||a(y) - a_{Q_a}(y)||_{L^\infty(D)}||f_{Q_f}(y)||_{L^\infty(D)}||u(y) - u_Q(y)||_X. \end{aligned} \tag{4.51}$$

By comparing the left hand side (4.42) and the right hand side (4.51), we obtain the error bound (4.34) depending only the data a, f and their empirical interpolation errors, where C_1 and C_2 are defined as

$$C_1 := C_D C_P C_X \le \sqrt{|D|}\frac{d_D}{2}\sqrt{\left|\left|\frac{1}{a(\bar{y})}\right|\right|_{L^\infty(D)}} \text{ and } C_2 := C_X^2 \le \left|\left|\frac{1}{a(\bar{y})}\right|\right|_{L^\infty(D)}. \tag{4.52}$$

□

Lemma 4.3.3 *As for the approximation error between the compliant output $s(y) = (f(y), u(y))$ and the approximate compliant output $s_{Q,N}(y) = (f_{Q_f}(y), u_{Q,N}(y))$, we have*

$$|s(y) - s_{Q,N}(y)| \le \mathscr{E}_Q^s(y) + \Delta_N^s(y), \tag{4.53}$$

where Δ_N^s is the a posteriori error bound for the reduced basis approximation corresponding to (1.27), $\mathscr{E}_Q^s$ is the error due to the affine approximation of data a and f, defined as

$$\mathscr{E}_Q^s(y) := \frac{C_1^2}{\alpha_{LB}(y)}||f(y) - f_{Q_f}(y)||_{L^\infty(D)}||f_{Q_f}(y)||_{L^\infty(D)} + C_1||f(y)||_{L^\infty(D)}\mathscr{E}_Q^u(y), \tag{4.54}$$

where the constant C_1, the lower bound $\alpha_{LB}(y)$ and $\mathscr{E}_Q^u(y)$, $y \in \Gamma$, are defined in Lemma 4.3.2.

Proof Similar to the proof of Lemma 4.3.2, we split the output approximation error into

$$|s(y) - s_{Q,N}(y)| \le |s(y) - s_Q(y)| + |s_Q(y) - s_{Q,N}(y)|, \tag{4.55}$$

where the first part corresponds to the affine approximation error of the random fields a and f and the second part arises from the reduced basis approximation error bounded by

$$|s_Q(y) - s_{Q,N}(y)| \le \Delta_N^s(y), \tag{4.56}$$

which can be evaluated from (1.27). As for the first part, we seek a bound denoted as

$$|s(y) - s_Q(y)| \le \mathscr{E}_Q^s(y). \tag{4.57}$$

By definition of the output $s = (f, u)$ and the approximate output $s_Q = (f_{Q_f}, u_Q)$, we have

$$\begin{aligned}
|s(y) - s_Q(y)| &= |(f(y), u(y)) + (f_{Q_f}(y), u_Q(y))| \\
&\leq |(f(y) - f_{Q_f}(y), u_Q(y))| + |(f(y), u(y) - u_Q(y))| \\
&\leq ||f(y) - f_{Q_f}(y)||_{L^\infty(D)} ||u_Q(y)||_{L^1(D)} + ||f(y)||_{L^\infty(D)} ||u(y) - u_Q(y)||_{L^1(D)} \\
&\leq C_1 ||f(y) - f_{Q_f}(y)||_{L^\infty(D)} ||u_Q(y)||_X + C_1 ||f(y)||_{L^\infty(D)} ||u(y) - u_Q(y)||_X \\
&\leq \frac{C_1^2}{\alpha_{LB}(y)} ||f(y) - f_{Q_f}(y)||_{L^\infty(D)} ||f_{Q_f}(y)||_{L^\infty(D)} + C_1 ||f(y)||_{L^\infty(D)} \mathscr{E}^u_{a,f}(y),
\end{aligned} \tag{4.58}$$

where the first inequality is due to the triangular inequality, the second one to the Hölder's inequality, the third one follows from combining (4.44), (4.45) and (4.46), and the fourth inequality follows from using (4.48) and the error bound (4.37), respectively. □

Remark 4.3.3 *As a result of Lemma 4.3.2 and Lemma 4.3.3, the approximation error for both the solution and the output can be split into two components: one arising from the empirical interpolation error of the random fields and another one from the reduced basis approximation error. Unfortunately, the evaluation of the empirical interpolation error for each sample $y \in \Gamma$ in (4.34) and (4.54) involves computing $||\cdot||_{L^\infty(D)}$ norm with at least $O(\mathcal{N})$ operations, being $\mathcal{N} = |V_x|$ the number of the finite element nodes. This would spoil the cheap online evaluation cost for a large number of samples required in the computation of failure probability, especially when $\mathcal{N}$ becomes very large.*

Inexpensive a posteriori error bound

To order to overcome the drawback of computational inefficiency pointed out in Remark 4.3.3, we seek the upper bounds $\mathscr{E}^{u,b}_Q$ and $\mathscr{E}^{s,b}_Q$ for the affine approximation error of the solution $\mathscr{E}^u_Q \leq \mathscr{E}^{u,b}_Q$ and the output $\mathscr{E}^s_Q \leq \mathscr{E}^{s,b}_Q$, whose computational cost is small and independent of $\mathcal{N}$.

By the empirical interpolation Algorithm 3, we obtain from (3.9) and (3.10) the error bound

$$||a(y) - a_{Q_a}(y)||_{L^\infty(D)} \leq r^a_{Q_a+1}(x^{Q_a+1}, y^{Q_a+1}) \quad \forall y \in \Xi^a_y \tag{4.59}$$

and

$$||f(y) - f_{Q_f}(y)||_{L^\infty(D)} \leq r^f_{Q_f+1}(x^{Q_f+1}, y^{Q_f+1}) \quad \forall y \in \Xi^f_y, \tag{4.60}$$

where $r^a_{Q_a+1}$ and $r^f_{Q_f+1}$ are the the empirical interpolation errors defined in (3.8) corresponding to the nonaffine random fields a and f, respectively. Although the relation (4.59) and (4.60) hold true only in the sample sets Ξ^a_y and Ξ^f_y, we remark that in practice they also often hold in the whole probability domain Γ, especially when the cardinality of sample sets is big or the random fields are rather smooth with respect to the random vector y.

Since computing $||f_{Q_f}(y)||_{L^\infty(D)}$ in (4.34) and (4.54) for $y \in \Gamma$ is expensive, we bound the quantity $||u_Q(y)||_X$ in (4.47) directly by

$$||u_Q(y)||_X \leq ||u_{Q,N}(y)||_X + \triangle^u_N(y), \tag{4.61}$$

(instead than by (4.48)), which can be cheaply evaluated in the online stage. Now we can compute the following error bound for the affine approximation error of the solution by using (4.59), (4.60) and

(4.61),

$$\begin{aligned}\mathscr{E}_Q^{u,b}(y) := &\frac{C_1}{\alpha_{LB}(y)} r^f_{Q_f+1}(x^{Q_f+1}, y^{Q_f+1}) \\ &+ \frac{C_2}{\alpha_{LB}(y)} r^a_{Q_a+1}(x^{Q_a+1}, y^{Q_a+1})(||u_{Q,N}(y)||_X + \Delta_N^u(y)).\end{aligned} \tag{4.62}$$

As for the error bound $\mathscr{E}_Q^{s,b}(y)$, we also need to compute $||f(y)||_{L^\infty(D)}$ for $y \in \Gamma$, which is rather expensive. Alternatively, we can bound the second term $|(f(y), u(y) - u_Q(y))|$ in (4.58) by

$$\begin{aligned}|(f(y), u(y) - u_Q(y))| &= |(a(y)\nabla u(y), \nabla(u(y) - u_Q(y)))| \\ &\le a_{max} C_2 ||u(y)||_X ||u(y) - u_Q(y)||_X \\ &\le a_{max} C_2 (\mathscr{E}_Q^{u,b}(y) + \Delta_N^u(y) + ||u_{Q,N}||_X) \mathscr{E}_Q^{u,b}(y),\end{aligned} \tag{4.63}$$

where the first inequality follows from the definition of the constants a_{max} in (41) and C_2 in (4.46), while the second inequality holds because of the triangular inequality with the associated error bounds

$$||u(y)||_X \le ||u(y) - u_Q(y)||_X + ||u_Q(y) - u_{Q,N}(y)||_X + ||u_{Q,N}(y)||_X. \tag{4.64}$$

In conclusion, a cheaper error bound for the output $\mathscr{E}_Q^{s,b}(y)$ reads

$$\begin{aligned}\mathscr{E}_Q^{s,b}(y) := &C_1 r^f_{Q_f+1}(x^{Q_f+1}, y^{Q_f+1}) \left(||u_{Q,N}(y)||_X + \Delta_N^u(y)\right) \\ &+ a_{max} C_2 \left(||u_{Q,N}||_X + \Delta_N^u(y) + \mathscr{E}_Q^{u,b}(y)\right) \mathscr{E}_Q^{u,b}(y).\end{aligned} \tag{4.65}$$

The error bound $\mathscr{E}_Q^{s,b}(y)$ is cheap to evaluate since the solution $u_{Q,N}(y)$ and the error bounds $\Delta_N^u(y)$ and $\mathscr{E}_Q^{u,b}(y)$ can be computed online with at most $O((Q_f + NQ_a)^2)$ operations, and the other constants in (4.65), i.e., $C_1, r^f_{Q_f+1}(x^{Q_f+1}, y^{Q_f+1}), a_{max}, C_2$, are evaluated only once for all the samples.

On the evaluation of failure probability

In the evaluation of failure probability, the reduced basis method stays the same as presented in Algorithm 2 , while the a posteriori error bound used in the hybrid reduced basis method in Algorithm 4 is modified as the global a posteriori error bound $\mathscr{E}_Q^{s,b} + \Delta_N^s$. In both methods, we prefer to construct a more accurate empirical interpolation for the nonaffine random fields and a richer reduced basis space with small approximation error in order to improve the computational accuracy and efficiency, especially for $\mathscr{N}$ large entailing a costly solution of the full PDE. As for the goal-oriented reduced basis method, we adopt different computational strategies for different properties of the nonaffine random fields.

When the random fields are rather regular (smooth manifold) with respect to the random vector y, the decay of the optimal approximation error or Kolmogorov width d_Q is very fast, so the empirical interpolation error also converges rapidly to zero thanks to Theorem 3.2.1. In this case, the affine approximation error could be very small and dominated by the reduced basis approximation error. Therefore, goal-oriented adaptive reduced basis construction is still effectively governed by the a posteriori reduced basis approximation error bound. Whenever the distance between the approximate output and the critical value is smaller than the affine approximation error bound at sample $y \in \Gamma$, i.e., $|s_{Q,N}(y) - s_0| \le \mathscr{E}_Q^{s,b}(y)$, which is extremely rare, we solve the full PDE to evaluate an accurate output.

On the other hand, if the nonaffine random fields are far from smooth in the probability space, the affine approximation error bound $\mathscr{E}_Q^{s,b}(y)$ could be relatively large for small Q. In order to guarantee

that the affine approximation error bound is dominated by the reduced basis approximation error bound, the number of the affine terms Q_a, Q_f might be very large, resulting in relatively more expensive online evaluation with $O((Q_f + NQ_a)^2)$ operations. In this circumstance, we choose to start from a crude approximation with small Q_a, Q_f, N for sake of computational efficiency and adaptively enrich the bases in the reduced basis space as well as refine the empirical interpolation with more affine terms governed by the error bounds $\triangle_N^s$ and $\mathscr{E}_Q^{s,b}$.

4.4 Numerical experiments

In this section, we carry out several numerical experiments to illustrate the computational difficulties encountered by conventional methods and demonstrate the accuracy and efficiency of our proposed methods for the evaluation of failure probability. Moreover, we apply our methods to more general PDE models including noncompliant, unsteady and nonaffine problems.

4.4.1 Benchmark models

One-dimensional problems

First of all, we study the benchmark model of the elliptic coercive affine scalar problem (39) with different one-dimensional random inputs. The physical domain is specified as a square $D = (0,1)^2$. We take a deterministic force term $f = 1$ for simplicity and consider the random diffusion coefficient a in different cases. The solution of the PDE model in the physical domain is approximated by piecewise linear finite element functions. In the probability domain Γ, we approximate the solution by the stochastic collocation method introduced in chapter 1 and the reduced basis method. For the latter, we use a uniform lower bound $\alpha_{LB} \leq \alpha(y), \forall\, y \in \Gamma$, for the sake of computational efficiency, see [174, 178].

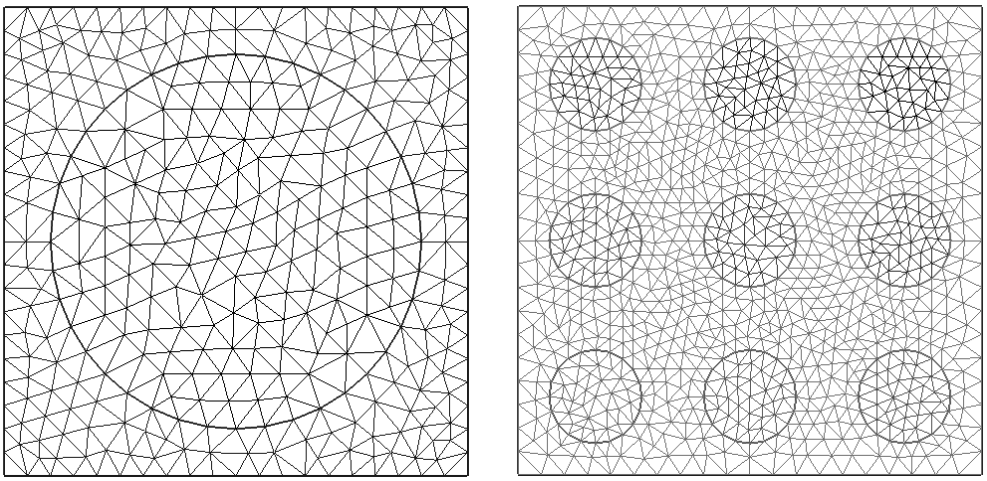

Figure 4.1: Finite element mesh for the physical domain D with 1 disk (left) and 9 disks (right).

In the first test, we take a random field $a(x,y) = (1.1 + y\mathscr{X}_1(x))/10, x \in D, y \in \Gamma$ where y is a random variable uniformly distributed in $\Gamma = [-1,1]$ and $\mathscr{X}_1$ is a characteristic function supported on a disk with radius 0.4 and center $(0.5,0.5)$, i.e., $\mathscr{X}_1(x) = 1$ if $(x_1 - 0.5)^2 + (x_2 - 0.5)^2 \leq 0.4^2$; see the left of Figure 4.1. Note that the random field a is a first order polynomial of y, thus smooth in the probability domain Γ. In the second test, we take $a(x,y) = (1.1 + (1 - 2\mathscr{X}_{0.5}(y))\mathscr{X}_1(x))/10, x \in D, y \in \Gamma$, where the characteristic function is $\mathscr{X}_{0.5}(y) = 1$ if $|y| \leq 0.5$. The random field a is now discontinuous in the probability domain

Γ, in fact taking only two different values. The critical value of the output is taken as $s_0 = 0.2845$ in the first test and $s_0 = 0.2726$ in the second. For the approximation of the output s in probability domain, we first approximate the solution u by employing the stochastic collocation method with hierarchical Clenshaw–Curtis rule [149], where the number of collocation nodes is $N = 2^n + 1, 1 \le n \le 5$, then evaluate the output $s_N = s(u_N)$ at the approximate solution u_N.

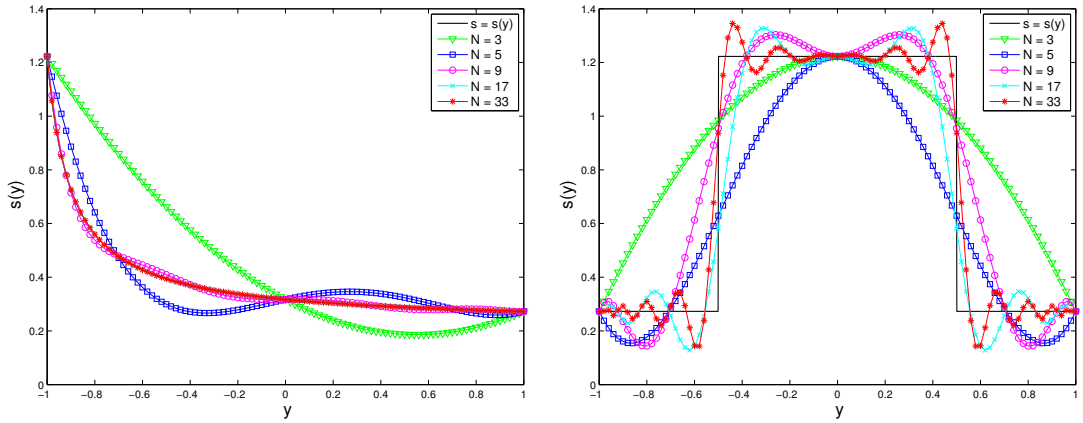

Figure 4.2: The stochastic collocation approximation of the output with different collocation nodes. Left: the random coefficient a is smooth; right, the random coefficient a is discontinuous in Γ.

Figure 4.2 displays the output $s(y), y \in \Gamma$ and the stochastic collocation approximation of the output for both the smooth and the discontinuous random fields. From the left of Figure 4.2, we can observe that the output approximated by the stochastic collocation method converges to the accurate output when increasing the number of collocation nodes. The worst approximation error $\max_{y\in\Gamma} |s(y) - s_N(y)|$ computed in the sample set Ξ_{new} with $|\Xi_{new}| = 1000$ is shown in Table 4.1, which decreases to zero very fast and the failure probability $P(\omega \in \Omega : s(y(\omega)) < s_0)$ converges to the true value 0.20. As for the discontinuous test, we can see from the right of Figure 4.2 that the approximate output oscillates around and does not converge to the accurate output, because of the Gibbs phenomenon (see also [61]). Due to the Gibbs phenomenon, the worst approximation error does not converge to zero but increases and the failure probability evaluated by the stochastic collocation method is far from the true value 0, as can be seen in Table 4.1 for Test 2. In order to compute an accurate failure probability, the threshold value in the hybrid approach must be so large that too many outputs (at samples in half of the probability domain in this example) have to be evaluated by fully solving the underlying PDE, which severely deteriorates the advantage of hybrid scheme. In the extreme case, the hybrid scheme may not gain any computational efficiency due to the fact that the outputs at most of the samples have to be evaluated by solving a full PDE.

Test \ Number of collocation nodes	$N = 3$	$N = 5$	$N = 9$	$N = 17$	$N = 33$
Test 1, $\max_{y\in\Gamma} \|s(y) - s_N(y)\|$	0.41	0.16	0.026	7.7e-4	6.3e-7
Test 1, $P(\omega \in \Omega : s(y(\omega)) < s_0)$	0.46	0.31	0.27	0.21	0.20
Test 2, $\max_{y\in\Gamma} \|s(y) - s_N(y)\|$	0.95	0.95	1.03	1.06	1.07
Test 2, $P(\omega \in \Omega : s(y(\omega)) < s_0)$	0.00	0.28	0.22	0.24	0.20

Table 4.1: Worst approximation error and failure probability of Test 1 (smooth) and Test 2 (discontinuous) evaluated by the stochastic collocation method with different number of nodes.

In comparison, the worst approximation error for the output by the reduced basis method (where we have set $\varepsilon_{tol} = 1.0 \times 10^{-14}$) decreases extremely fast, reaching 2.4×10^{-14} with only four bases in the

first test of smooth random field, and it completely vanishes with only two bases in the second test of discontinuous random field. The failure probability evaluated by the reduced basis method is exact in both tests. This remarkable computational accuracy and efficiency can be attributed to the fact that the reduced basis method takes the solution (only two different solutions in the discontinuous case) as the approximation basis and solves a reduced PDE that inherits the same structure of the full PDE. Consequently, when the random input function is discontinuous as shown in this example, the reduced basis method overcomes the challenge of Gibbs phenomenon by avoiding the usage of dictionary basis. This conclusion holds for more general nonsmooth random input functions as long as the model outputs depend smoothly on the inputs.

Multidimensional problems

To further investigate the computational accuracy and efficiency of different methods for the evaluation of failure probability, we consider a multidimensional problem with many random inputs. The physical domain D and force term f are specified as in previous case. We suppose that there are nine disks in the domain (see the right of Figure 4.1) and define the background coefficient as $a_0(x,y) = 1, x \in D, y \in \Gamma$ and coefficients in the disks as $a_k(x,y) = 10^{y_k}\mathcal{X}_k(x), 1 \le k \le 9, x \in D, y \in \Gamma$, where $y_k, 1 \le k \le 9$ are independent and obeying uniform distribution in $\Gamma_k = [-2,2]$, the characteristic functions are defined as $\mathcal{X}_k(x) = 1, (x_1 - x_1^k)^2 + (x_2 - x_2^k)^2 \le 0.1^2$, with the centers at the points $((2i-1)/6, (2j-1)/6), 1 \le i, j \le 3$ where $3(i-1)+j = k$. The random coefficient a is defined as $a = (a_0 + a_1 + \cdots + a_9)/10$.

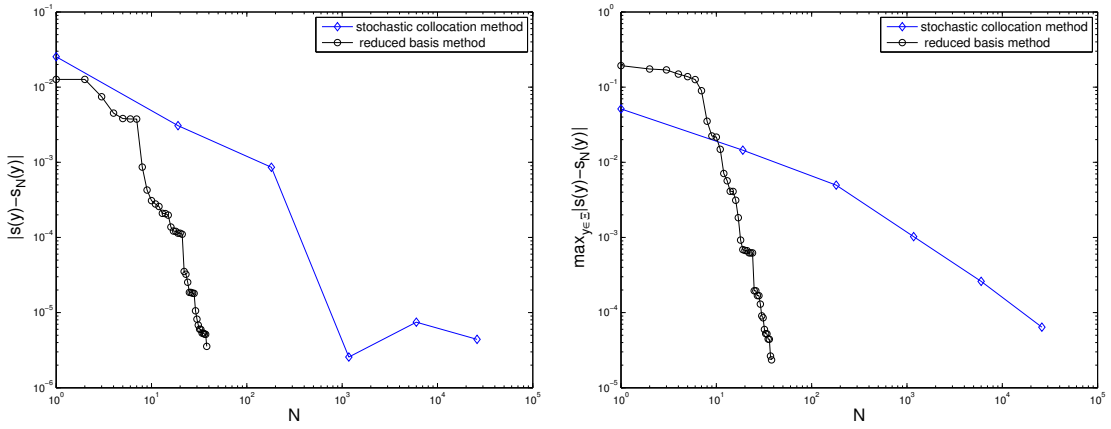

Figure 4.3: Comparison of output error between stochastic collocation approximation and reduced basis approximation. Left: error at one sample; right: worst approximation error.

In this numerical test, we employ the sparse grid stochastic collocation method introduced in chapter 1 to approximate the output s directly; 100 realizations of the random input $y \in \Gamma$ are sampled according to its probability distribution to specify the training set for the construction of the reduced basis space and another 100 realizations are sampled to test the two approximation methods. Figure 4.3 reports the comparison of the output error $|s - s_N|$ between stochastic collocation approximation and reduced basis approximation. On the left, the comparison is performed at one sample randomly chosen from the probability domain Γ, from which we can observe that the reduced basis approximation error decreases monotonically and much faster than the stochastic collocation approximation error, which starts to oscillate when the number of collocation nodes gets large due to over fitting problem (Gibbs phenomenon). On the right, the comparison is carried out for the worst approximation error (the largest approximation error among 100 test samples randomly chosen in the probability domain), which shows that the reduced basis approximation is much more efficient than the stochastic collocation approximation in that only a small number (≤ 38) of the full PDEs need to be solved in order to gain the

same worst approximation error compared to a significant large number (≥ 26017) of samples for the sparse grid collocation approach. The method becomes especially efficient when the solution resides in a low-dimensional manifold while the random inputs are in high dimensions.

Figure 4.4 displays the effectivity of the employment of a posteriori error bound. On the left, we report the decay of the error bound $\triangle_N^s$ and the real output error $|s - s_N|$ with respect to the number of reduced basis functions at one sample randomly chosen from the probability domain. On the right, the effectivity defined as $\triangle_N^s/|s - s_N|$ at 100 test samples is shown. It proves that $\triangle_N^s \geq |s - s_N|$ for all the samples and the error bound $\triangle_N^s$ is not far from the real error $|s - s_N|$ at most of the samples, so that it is reasonable to use the a posteriori error bound for both certification of the approximation output and construction of the reduced basis space.

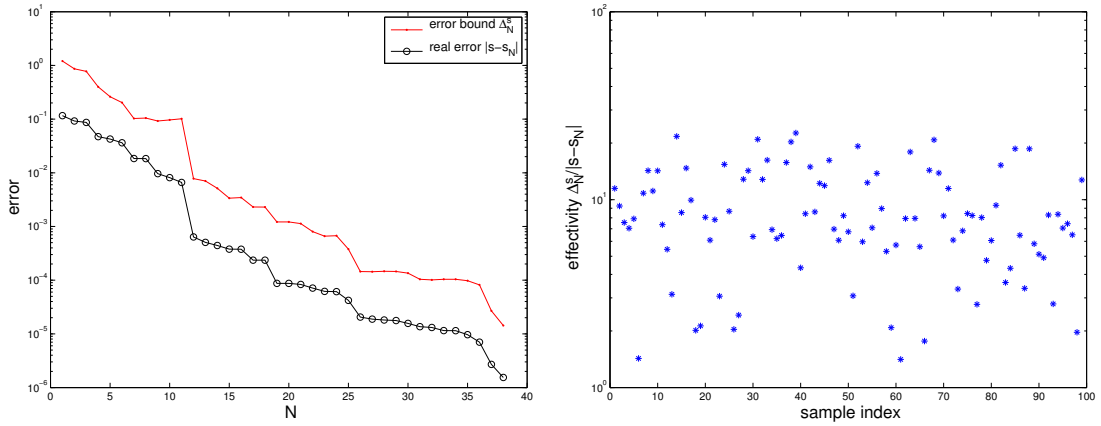

Figure 4.4: Left: comparison of error bound $\triangle_N^s$ and real error $|s - s_N|$ with respect to the number of reduced basis functions N at one sample; right: effectivity $\triangle_N^s/|s - s_N|$ at 100 test samples.

For the evaluation of the failure probability, we test both the hybrid reduced basis method and the goal-oriented reduced basis method. From the same training set with 100 samples, we construct a fine reduced basis space with tolerance $\varepsilon_{tol} = 1 \times 10^{-4}$ for the former method, resulting in 38 bases, and a coarse reduced basis space with tolerance $\varepsilon_{tol} = 1 \times 10^{-2}$ for the latter method, leading to 18 bases. We compute the failure probability by hybrid algorithm 4 and goal-oriented adaptive Algorithm 5 by setting $M_0 = 1000$ initial samples, the scaling parameter $\beta = 4$ and a posterior error tolerance $\epsilon_{tol} = 1 \times 10^{-3}$. We remark that a small value of the adaptive parameter β leads to a relatively small number of samples in each adaptation step, which is favorable for computational efficiency for the offline construction of the reduced basis space since there are less samples that need to be searched over. Large β potentially produces large difference of the a posteriori error e_i^p in Algorithm 5, which drives I_{max} big and results in relatively more accurate failure probability. Here and in the following numerical experiments, we set $\beta = 4$ as a trade-off between computational efficiency and numerical accuracy. The comparison results are recorded in Table 4.2, from which we can see that the reduced basis space for the hybrid method is fine enough and we only need to solve 329 full PDEs in total in order to evaluate the outputs at 341000 samples. By the goal-oriented adaptive approach, the total number of full PDEs that should be solved is 132. Nevertheless, only 36 PDEs need a full solving thanks to the adaptation of the reduced basis space at each iteration, which achieves further computational efficiency. Moreover, owing to an effective and cheap a posteriori error bound, both the hybrid approach and the goal-oriented adaptive approach result in the same failure probability (0.027 for a critical value $s_0 = 0.25$) as being solved directly by Monte Carlo method. In summary, both the hybrid and the goal-oriented adaptive reduced basis methods have been successfully applied to efficiently and accurately compute the failure probability, with the goal-oriented adaptive approach gaining remarkable computational efficiency thus more suitable to solve complex PDEs with time-consuming solver.

Number of Monte Carlo samples	$1M_0$	$4M_0$	$16M_0$	$64M_0$	$256M_0$	$341M_0$
Hybrid RBM, # ($\|s-s_N\|<\Delta_N^s$)	0	3	22	59	245	329
Adaptive RBM, # ($\|s-s_N\|<\Delta_N^s$)	41	41	20	8	22	132
Adaptive RBM, # adapted bases	8	9	6	5	8	36
Failure probability $P_0^m=P_0^h=P_0^g$	0.043	0.033	0.030	0.027	0.027	0.027

Table 4.2: Comparison between hybrid RBM and goal-oriented adaptive RBM in terms of the number of samples for which the full PDE have to be solved; $M_0 = 1000$.

4.4.2 Noncompliant problems

We take $D=(0,1)^2$ and suppose that the covariance fields of the random inputs are available and both the diffusion coefficient a and the force term f are obtained from truncation of the Karhunen-Loève expansion of covariance fields [189], expressed as

$$a(x,y(\omega))=\mathbb{E}[a]+\sum_{q=1}^{Q_a}\sqrt{\lambda_q^a}a_q(x)y_q(\omega) \text{ and } f(x,y(\omega))=\mathbb{E}[f]+\sum_{q=1}^{Q_f}\sqrt{\lambda_q^f}f_q(x)y_q^f(\omega), \tag{4.66}$$

where $(\lambda_q^a,a_q)_{q=1}^{Q_a}$ and $(\lambda_q^f,f_q)_{q=1}^{Q_f}$ are the eigenvalues and orthonormal eigenfunctions associated to their corresponding covariance fields, $y_q^a, 1\le q\le Q_a$ and $y_q^f, 1\le q\le Q_f$ are mutually uncorrelated with mean zero and unit variance. For the $i-$th coordinate, $i=1,\dots,d$, the general formula of a Gaussian random field $g(x_i,y)$ is written as [149]

$$g(x_i,y)=\mathbb{E}[g]+\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2}y_1^g(\omega)+\sum_{k=1}^{K}\sqrt{\lambda_n}\left(\sin(k\pi x_i)y_{2k}^g(\omega)+\cos(k\pi x_i)y_{2k+1}^g(\omega)\right), \tag{4.67}$$

where the random variables $y_k^g, 1\le k\le 2K+1$ are assumed to be uniformly distributed in $[-\sqrt{3},\sqrt{3}]$. For simplicity, we assume that the covariance fields for a and f are Gaussian fields depending on x_1 coordinate and x_2 coordinate, respectively, with the same correlation length $L=1/4$ and eigenvalues $\lambda_1=0.3798, \lambda_2=0.2391, \lambda_3=0.1106, \lambda_4=0.0376, \lambda_5=0.0094, \lambda_6=0.0017$, etc.. We take $Q_a=Q_f=13$ with $K=6$ in (4.67) leading to a 26 dimensional problem, which accounts for around 99% uncertainties of the random field. The expectation of the random force f given by (4.67) is taken as $\mathbb{E}[f]=6$; the expectation of a random field $\tilde{a}$ given by (4.67) is specified as $\mathbb{E}[\tilde{a}]=5$ and we take $a=\tilde{a}/10$. The output $s(y)=s(u(y))=\int_D 10u(x,y)dx$ is different from the force term.

We adopt the primal-dual computational strategy for noncompliant problems presented in section 4.3.1. We set the tolerance $\varepsilon_{tol}=1\times10^{-4}$ for $||R^{pr}||_{X'}^2/\alpha_{LB}$ and $||R^{du}||_{X'}^2/\alpha_{LB}$ (see the definition of residual in (4.17)) in the hybrid reduced basis method and $\varepsilon_{tol}=1\times10^{-2}$ in the goal-oriented adaptive reduced basis method. The constructed hybrid reduced basis space for the primal problem contains 27 bases and 14 bases for the dual problem, while for the construction of the goal-oriented adaptive reduced basis method, there are 9 and 7 bases respectively. We test the reduced basis approximation for both the primal and the dual problems with 100 test samples and present the worst approximation errors in Figure 4.5, which illustrates the exponentially fast convergence of the reduced basis approximation in high-dimensional random inputs. Figure 4.6 depicts the dependence of the worst approximation error for the output $\max_{y\in\Xi_{test}}|s(y)-s_N(y)|$ (Ξ_{test} denotes the test sample set with 100 samples) with respect to the number of bases in the primal and dual reduced basis space (left) as well as the effectivity of the a posterior error bound defined as $\Delta_N^s(y)/|s(y)-s_N(y)|$ (right), from which we can observe that simultaneous increase of the bases in both primal and dual reduced basis spaces not only leads to faster convergence of the output approximation error but also improves the sharpness of the a posteriori error bound thanks to the “square effect", thus enhances the computational efficiency for the evaluation of

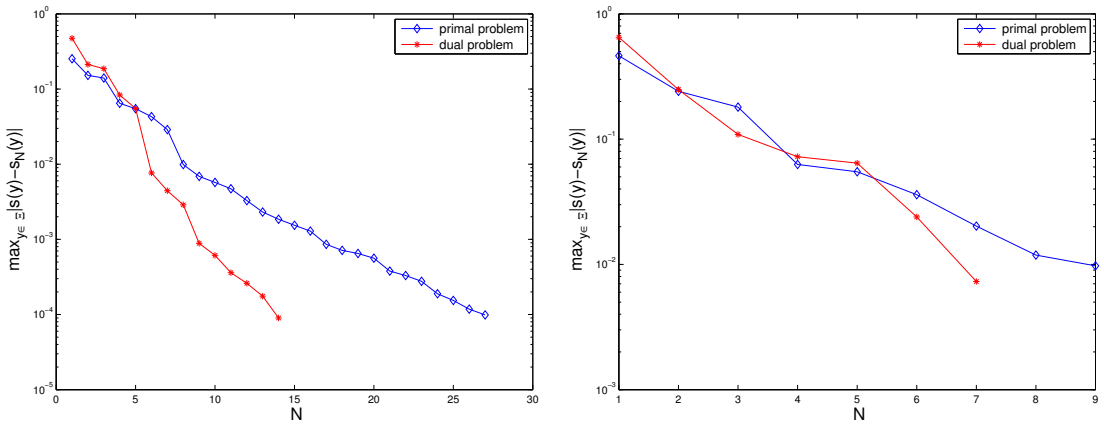

Figure 4.5: Worst primal-dual reduced basis approximation error of hybrid type with $\varepsilon_{tol} = 1 \times 10^{-4}$ (left) and goal-oriented adaptive type with $\varepsilon_{tol} = 1 \times 10^{-2}$ (right) at 100 test samples.

the failure probability.

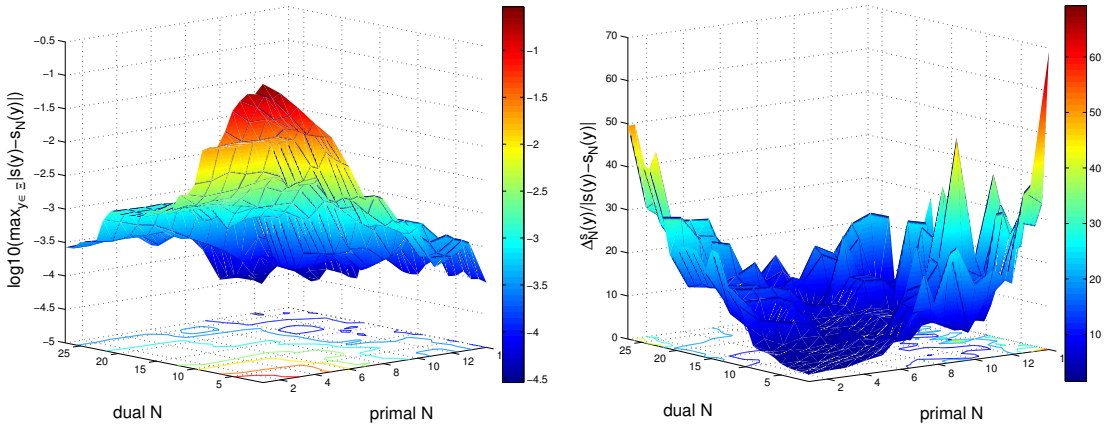

Figure 4.6: The worst primal-dual hybrid reduced basis approximation error with $\varepsilon_{tol} = 1 \times 10^{-4}$ (left) and goal-oriented adaptive type with $\varepsilon_{tol} = 1 \times 10^{-2}$ (right) at 100 test samples.

The error tolerance for the failure probability is set to $\epsilon_{tol} = 1 \times 10^{-4}$ with a critical value $s_0 = 4$. We test both the hybrid and the goal-oriented adaptive approaches, with the results recorded in Table 4.3. Due to the fact that the solution lies in a very low-dimensional stochastic manifold, the fine hybrid reduced basis approximate output is very close to the true value and there are only 52 out of 1365000 samples that cannot be determined; as for the goal-oriented adaptive approach, 48 samples can not be determined and only 21 PDEs are fully solved for adaptation of the primal and dual reduced basis spaces. From this experiment, we can see that we do not gain much more computational efficiency by the goal-oriented adaptive method than by the hybrid method, so that it is sufficient to use the hybrid reduced basis method to compute failure probability for problems with very smooth solution in the probability space.

Number of Monte Carlo samples	$1M_0$	$4M_0$	$16M_0$	$64M_0$	$256M_0$	$1024M_0$	$1365M_0$
Hybrid RBM, # ($\|s-s_N\|<\triangle_N^s$)	1	0	0	1	11	39	52
Adaptive RBM, # ($\|s-s_N\|<\triangle_N^s$)	3	4	7	5	15	14	48
Adaptive RBM, # adapted bases	2	1	3	4	6	5	21
Failure probability $P_0^m=P_0^h=P_0^g$	0.361	0.372	0.3823	0.3864	0.3832	0.3831	0.3831

Table 4.3: Comparison between hybrid RBM and goal-oriented adaptive RBM in terms of the number of samples for which the full PDE have to be solved; $M_0 = 1000$.

4.4.3 Unsteady problems

We consider a heat transfer problem in a thermal fin with the geometry displayed in Figure 4.7, where the thermal conductivity in the main body and the four extended surfaces depends on five independent random variables obeying uniform distribution in $[-2,2]$, i.e.,

$$a_0(x,y) = 1 + 10^{y_0}\mathscr{X}_0(x), \text{ and } a_k(x,y) = 10^{y_k}\mathscr{X}_k(x), 1 \le k \le 4,$$

where the characteristic functions $\mathscr{X}_k, 0 \le k \le 4$ are supported in the sub domains $D_k, 0 \le k \le 4$. Moreover, we consider the Biot number on the Robin boundaries as a random field as

$$b(x,y) = 10^{y_5}\mathscr{X}_{\partial D_r}(x),$$

where the characteristic function $\mathscr{X}_{\partial D_r}(x)$ is supported on the Robin boundaries. The time dependent heat transfer problem is formulated in the strong form as

$$\frac{\partial u(t,x,y)}{\partial t} - \sum_{k=0}^{4} \nabla(a_k(x,y)\nabla u(t,x,y)) = 0, \quad (t,x) \in [0,T]\times D, \text{ a.s. } y \in \Gamma, \tag{4.68}$$

where $\Gamma = [-2,2]^6$; we take $T = 5$ and impose homogeneous initial condition $u(0,x,y) = 0$ everywhere; we also prescribe heat flux $f(x) = 1, x \in \partial D_n^1$ at the bottom edge, homogeneous Neuman condition on the boundary ∂D_n^2 and the following Robin condition on the boundary of the extended surfaces ∂D_r:

$$\sum_{k=0}^{4} a_k(x,y)\nabla u(t,x,y)\cdot\mathbf{n} + b(x,y)u(t,x,y) = 0, \quad (t,x) \in [0,T]\times\partial D_r, \text{ a.s. } y \in \Gamma.$$

By the first order backward Euler scheme for time discretization with time step $\Delta t = 0.05$, we can write the semi-weak formulation of the problem (4.68) as: find $u^i(y) \in X, 1 \le i \le 100$ such that the following equation holds almost surely $y \in \Gamma$:

$$M(u^i(y),v;y) + \Delta t\sum_{k=0}^{4} A_k(u^i(y),v;y) + \Delta t B(u^i(y),v;y) = \Delta t F(v;y) + M(u^{i-1}(y),v;y), \forall v \in X, \tag{4.69}$$

where $B(u^i(y),v;y) = \int_{\partial D_r} b(x,y)u^i(x,y)v(x)dx$ and $F(v;y) = \int_{\partial D_n^1} f(x)v(x)dx$.

We define the compliant output as the integrated temperature over the flux boundary at the final time $T = 5$, i.e., $s(y) = s(T;y) = F(u(T;y);y)$, and consider a critical value $s_0 = 2.3$ with failure probability (ineffective heat transfer) defined as $P_f(\omega \in \Omega : s(y(\omega)) > s_0)$. Figure 4.8 displays temperature distribution at three different samples at the end of the simulation, being the first one very effective for heat transfer and the last one ineffective.

We build the reduced basis space for hybrid method with tolerance $\varepsilon_{tol} = 1\times10^{-4}$, resulting in 93 bases as shown in the left of Figure 4.9; as for goal-oriented adaptive method, we set the tolerance

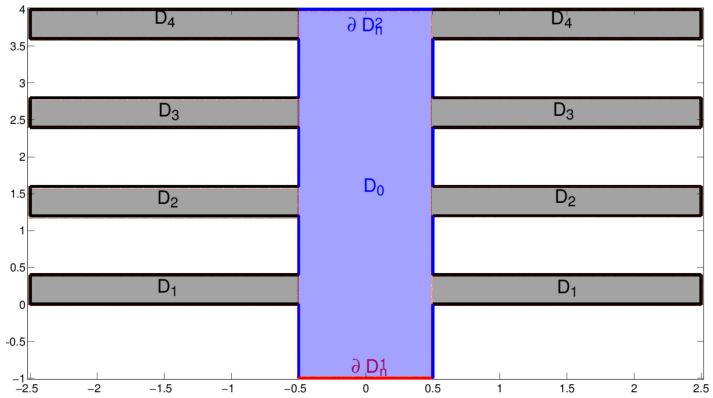

Figure 4.7: Geometry of a thermal fin, with domain D_0 (blue) defined as the main body, $D_k, 1 \le k \le K$ (black) as the extended surfaces, ∂D_n^1 (red) where imposing heat flux, ∂D_n^2 (blue) for homogeneous Neuman condition and the left boundary ∂D_r as Robin boundary.

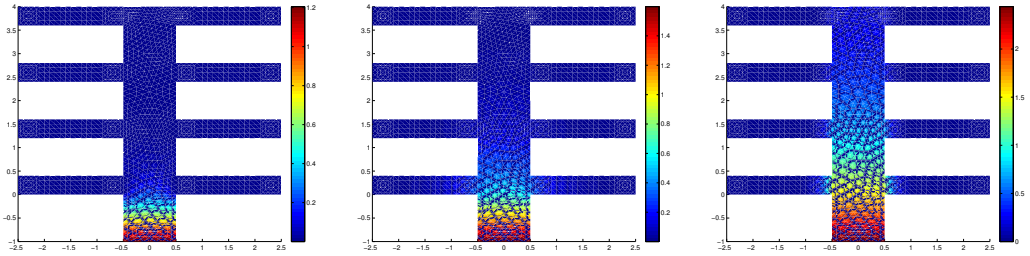

Figure 4.8: Temperature distribution at $T = 5$ for three different samples: left, $y_k = 2$, effective heat transfer; middle, reference $y_k = 0$; right, $y_k = -2, 0 \le k \le 5$, ineffective heat transfer.

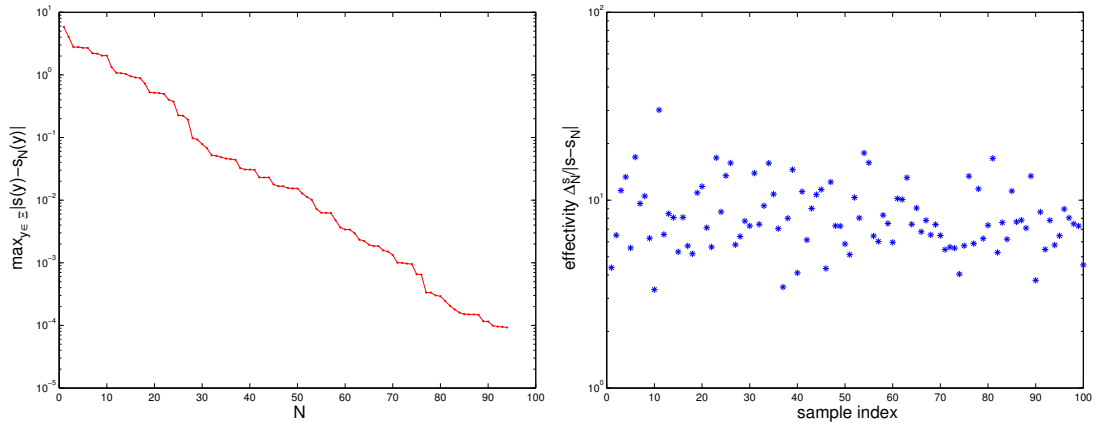

Figure 4.9: Left: decay of worst approximation error $\max_{y \in \Xi_{test}} |s(y) - s_N(y)|$ with respect to the number of reduced basis functions N; right: error bound effectivity $\Delta_N^s / |s - s_N|$ at 100 samples.

$\varepsilon_{tol} = 1 \times 10^{-2}$, leading to 42 initial bases. The effectivity for a posteriori error bound at 100 test samples is displayed in the right of Figure 4.9, which are sharp and distributed close to a constant smaller than

Number of Monte Carlo samples	$1M_0$	$4M_0$	$16M_0$	$64M_0$	$256M_0$	$341M_0$
Hybrid RBM, # ($\|s-s_N\|<\triangle_N^s$)	0	600	1500	3800	12400	18300
Adaptive RBM, # ($\|s-s_N\|<\triangle_N^s$)	700	1200	500	2300	1400	6100
Adaptive RBM, # PDE solves	400	400	200	700	300	2000
Adaptive RBM, # adapted bases	13	39	10	37	28	127
Failure probability $P_0^m = P_0^h = P_0^g$	0.0280	0.0315	0.0288	0.0304	0.0308	0.0308

Table 4.4: Comparison between hybrid RBM and goal-oriented adaptive RBM in terms of the number of samples for which the full PDE have to be solved; $M_0 = 1000$.

10. The results for the evaluation of failure probability with a tolerance $\epsilon_{tol} = 1 \times 10^{-3}$ are shown in Table 4.4, from which we can see that the goal-oriented adaptive approach is much more efficient than the hybrid approach, involving the solution of only 2000 full PDEs (4.69) instead of 18300 in the latter approach. We remark that the number of full PDE solves (2000 in total) is different from the number of adapted bases (127 in total) in goal-oriented adaptive reduced basis method for unsteady problems.

4.4.4 Nonaffine problems

Instead of the affine expansion (4.66) of the random fields a and f, we consider the Karhunen-Loève expansion for the logarithmic function of the random fields a and f, written as follows:

$$\log\big(a(x,y(\omega)) - \mathbb{E}[a]\big) = C_a \sum_{q=1}^{P_a} \sqrt{\lambda_q^a}\, a_q(x)\, y_q(\omega),$$

$$\log\big(f(x,y(\omega)) - \mathbb{E}[f]\big) = C_f \sum_{q=1}^{P_f} \sqrt{\lambda_q^f}\, f_q(x)\, y_q^f(\omega),$$

which are widely used in practical engineering models [149] in that the random fields are guaranteed to be positive, so that the random variables in the Gaussian random field expansion (4.67) are allowed to be standard Gaussian random variables with zero mean and unit variance. We take a correlation length $L = 1/16$ smaller than in section 4.4.2 for both the diffusion random coefficient $a(x_1, y)$ depending only in x_1 and the random force $f(x_2, y)$ depending only in x_2 in the formula (4.67). This leads to $P_a = P_f = 51$ terms to cover 99% of the total randomness, thus yielding a high-dimensional stochastic problem with $P_a + P_f = 102$ independent standard Gaussian random variables in total. The physical domain is set as $D = (0,1)^2$.

We perform an empirical interpolation procedure to affinely decompose the nonaffine random fields a (with $C_a = 50$ and $\mathbb{E}[a] = 0.1$) and f (with $C_f = 20$ and $\mathbb{E}[f] = 0.1$) with error tolerance $\varepsilon_{tol} = 1 \times 10^{-8}$ in Algorithm 3. The decay of the error bound $r_{Q+1}(x^{Q+1}, y^{Q+1})$ and the worst approximation error $\max_{y \in \Xi_{test}} ||g(y) - g_Q(y)||_{L^\infty(D)}$ computed in a test set Ξ_{test} with 100 samples are displayed in Figure 4.10, from which we can see that the empirical interpolation reaches very small error (1×10^{-8}) by only a few affine terms, $Q_a = 33$ for a_{Q_a} and $Q_f = 17$ for f_{Q_f} in (4.32), which are smaller than 51. By setting $\Theta_q^a, 1 \le q \le 33$ and $\Theta_q^f, 1 \le q \le 17$ as new random variables in the affine decomposition formula (4.32), we can view the empirical interpolation as an efficient dimension reduction method in order to alleviate the curse-of-dimensionality, especially when the manifold of stochastic solution is in low-dimensional probability space. Moreover, the error bound $r_{Q+1}(x^{Q+1}, y^{Q+1})$ is accurate and very sharp (close to the worst approximation error) as can be observed from Figure 4.10, so that the cheap a posteriori error bounds constructed in (4.62) and (4.65) are also accurate and sharp.

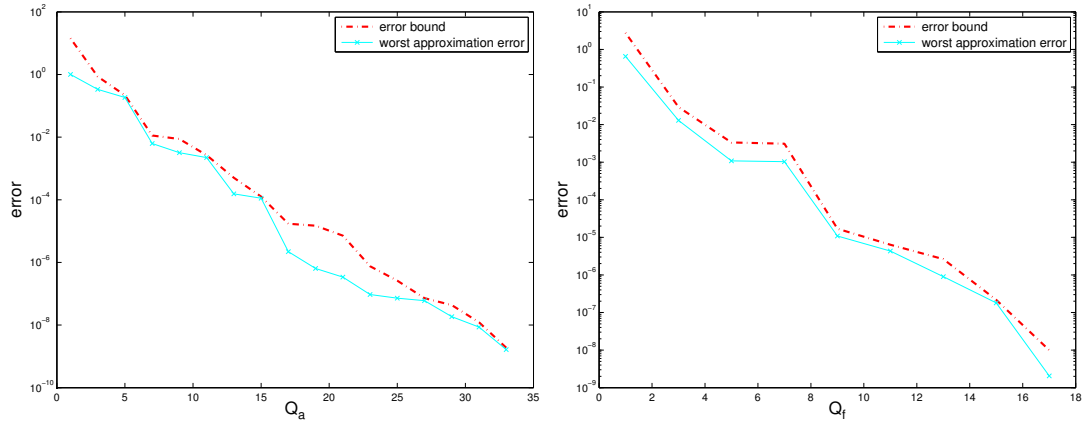

Figure 4.10: Decay of the error bound $r_{Q+1}(x^{Q+1}, y^{Q+1})$ and the worst approximation error $\max_{y\in\Xi_{test}} ||g(y) - g_Q(y)||_{L^\infty(D)}$ for a (left) and f (right) by empirical interpolation method.

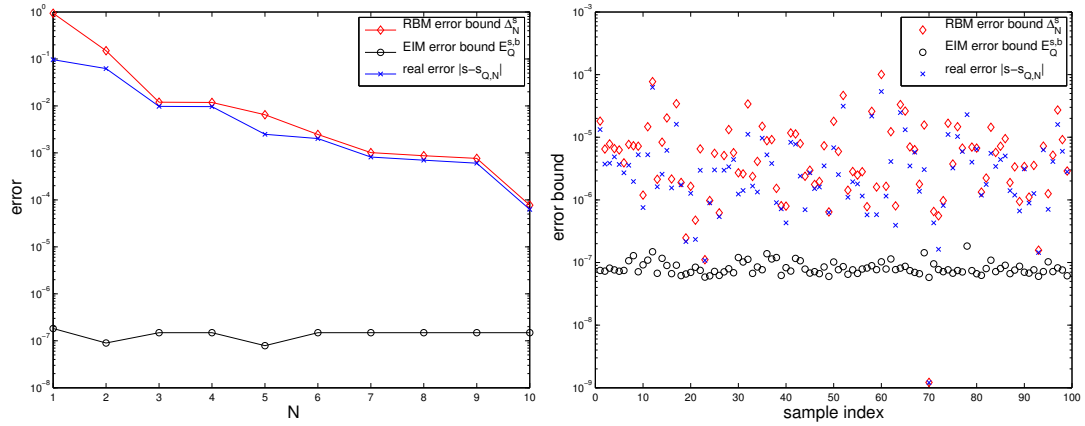

Figure 4.11: Worst approximation error $|s - s_{Q,N}|$, reduced basis error bound $\triangle_N^s$ and empirical interpolation error bound $\mathscr{E}_Q^{s,b}$ for N reduced basis functions (left), and at 100 test samples (right).

To evaluate the a posteriori error bound $\mathscr{E}_Q^{s,b}$ in (4.65) from the contribution of affine decomposition, we first compute $C_1 = 1$, $C_2 = 1$ from (4.52) and bound a almost surely by the estimate $a_{max} = 10$; the bound for the empirical interpolation error are taken from the construction of the affine decomposition as $r^a_{Q_a+1}(x^{Q_a+1}, y^{Q_a+1}) = 1.9 \times 10^{-9}$ and $r^f_{Q_f+1}(x^{Q_f+1}, y^{Q_f+1}) = 9.9 \times 10^{-9}$. We construct the reduced basis space with error tolerance $\varepsilon_{tol} = 1 \times 10^{-4}$, leading to 10 bases as shown on the left of Figure 4.11, where the reduced basis error bound $\triangle_N^s(y)$ as well as the empirical interpolation error bound $\mathscr{E}_Q^{s,b}(y)$ are also shown at the sample y that leads to the worst approximation real error $y = \arg\max_{y\in\Xi_{test}} |s(y) - s_{Q,N}(y)|$. It can be observed that the empirical interpolation error bound $\mathscr{E}_Q^{s,b}$ is much smaller than the reduced basis error bound $\triangle_N^s$, so that we can enrich the reduced basis space in order to obtain better approximation of the output with certified small error from affine decomposition. On the right of Figure 4.11, the different error bounds computed with 10 bases in the reduced basis space as well as the real output error are displayed at 100 test samples, which confirms the observation of the left figure for most of the samples (with one exception where the reduced basis approximation of the output is extremely close to the real output). Moreover, we can see that the reduced basis error bound is accurate and sharp, being very close to the real error at most of the samples. In order to evaluate the

failure probability with critical value $s_0 = 0.3$ and tolerance $\epsilon_{tol} = 1 \times 10^{-3}$, we set the reduced basis construction tolerance $\varepsilon_{tol} = 1 \times 10^{-4}$ for hybrid approach, resulting in 10 bases and $\varepsilon_{tol} = 1 \times 10^{-2}$ for goal-oriented adaptive approach with 4 bases. The results are displayed in Table 4.5, which shows that the reduced basis space is in very low dimensions (only 10 dimensions for the hybrid approach and $4 + 7 = 11$ dimensions for the goal-oriented adaptive approach) due to the fact that the stochastic solution and output live in a very low-dimensional manifold, even though the random inputs are in high dimensions.

Number of Monte Carlo samples	$1M_0$	$4M_0$	$16M_0$	$64M_0$	$256M_0$	$341M_0$
Hybrid RBM, # ($\lvert s - s_N \rvert < \Delta_N^s$)	0	1	1	7	33	42
Adaptive RBM, # ($\lvert s - s_N \rvert < \Delta_N^s$)	13	2	5	1	14	35
Adaptive RBM, # adapted bases	2	1	1	1	2	7
Failure probability $P_0^m = P_0^h = P_0^g$	0.064	0.059	0.062	0.065	0.064	0.064

Table 4.5: Comparison between hybrid RBM and goal-oriented adaptive RBM in terms of the number of samples for which the full PDE has to be solved; $M_0 = 1000$.

4.5 Summary

In this chapter, we developed hybrid and goal-oriented adaptive computational strategies based on the reduced basis method to efficiently and accurately compute the failure probability of partial differential equations with random inputs. In particular, we designed an efficient sampling scheme by the goal-oriented greedy algorithm to construct an accurate reduced basis model to approximate the stochastic output, especially for high-dimensional problems with many random inputs. In order to compute the failure probability of low regularity systems with respect to the random inputs, we developed a hybrid approach with goal-oriented adaptation governed by cheap and sharp a posteriori error bound for both the construction of the reduced basis space and the approximation of the output with certification. Using appropriate techniques, we extended the proposed methods for risk analysis to more general PDE models of noncompliant, unsteady and nonaffine types. In the numerical experiments, we studied different PDEs with uncertainties from physical parameters, external loadings, boundary conditions as random inputs obeying uniform distribution and normal distribution.

At this step, however, our numerical experiments are based on simple academic examples with specific design , with the sole aim of testing the computational properties of our proposed methods. Further research will be devoted to the development and application of our methods in practical engineering problems with more general PDE models and random inputs. It is worth to mention that Monte Carlo error plays a significant role in accurate evaluation of rare failure probability of extreme events. We will address this issue by efficiently combining model reduction with importance sampling techniques in a coming work, where the proposed reduced basis method is used to reduce the cost in solving the underlying PDE model and adaptive cross-entropy method is employed to reduce the number of Monte Carlo samples. We also remark that we did not take temporal and spatial discretization errors into account, which might be important, e.g., in highly nonlinear or advection-dominated problems. A global error analysis and the design of suitable global error bounds would be helpful for more rigorous evaluation of failure probability.

5 Breaking the curse of dimensionality – sparsity and reducibility

Forward uncertainty quantification (UQ) problems can be generally classified into two categories: integration problems, including computation of statistical moments, variance-based sensitivity analysis, etc., which were considered in chapter 2; pointwise evaluation problems, such as evaluation of probability density function, quantile, failure probability, etc., which were studied in chapter 4 in the context of reliability analysis. Despite various computational methods, such as stochastic Galerkin and collocation methods, can be effectively applied in solving forward UQ problems, a common computational challenge is faced by these methods, which is known as curse of dimensionality. When the dimension of the uncertainties becomes high (in the order of 100 and beyond), the number of dictionary projection bases or collocation nodes grows exponentially fast such that the computational burden can not be handled by even the most powerful computers. Another computational challenge stems from the fact that when the solution of the underlying model at one sample is expensive, the available computational resource can only afford the full solve at a few tens or hundreds samples, which is far from the required number (in the order of million or beyond) in a high-dimensional space. Any of the two challenges makes it impossible a direct application of the stochastic computational methods introduced above in solving high-dimensional UQ problems.

An opportunity to tackle this “curse-of-dimensionality" is to take advantage of the sparsity – the importance (or sensitivity) of different dimensions and their interaction/combination is very different for the quantities of interest, so that only a limited number of dimensions play an effective role. This role has lead to the development of the weighted function space based quasi Monte Carlo method [62], a priori and a posteriori analysis based anisotropic sparse grid construction [148], (Sobol) decomposition of function based techniques such as ANOVA (analysis of variance) [88, 78, 73], HDMR (high-dimensional model representation) [128], hierarchical surplus based dimension-adaptive generalized sparse grid techniques [33, 79, 88], and so on [21, 19, 140]. The quasi Monte Carlo method improves the convergence rate of the Monte Carlo method (which is $O(1/\sqrt{M})$ when using M randomly chosen samples) by following some digit rules or lattice rules [62] that explore the “weights" of different dimensions when choosing the samples. A faster convergence rate (typically $O((\log(M))^K/M)$ for K dimensional problems) can be achieved in this way. However, when the functions to be approximated feature smoothness and sparsity in the sense that the effective dimensions are much less than the total or nominal dimensions, the quasi Monte Carlo method is still too slow compared to the stochastic Galerkin method or the stochastic collocation method. Smoothness and sparsity have been exploited by anisotropic sparse grid techniques based on either a priori or a posteriori analysis of the convergence

Reference for this chapter:

P. Chen and A. Quarteroni. A new algorithm for high-dimensional uncertainty quantification problems based on dimension-adaptive and reduced basis methods. Submitted, 2014

rate of the approximate error in each stochastic dimension [148]. This has proved to be more efficient than the isotropic sparse grid in certain test cases. An essential drawback remains for this approach in that the interaction of different dimensions can not effectively be taken into account, leading to either too many useless grid nodes or less accurate approximation for some strongly interacting variables. As the high-dimensional function may be decomposed into a series of low-dimensional additive functions depending on the interaction of different dimensions, the variance based ANOVA (in combination with HDMR) approach has been employed to detect the interactions. Nevertheless, this approach may either be too expensive (more expensive than the original approximation problem based on Lebesgue measure) or not enough accurate (due to arbitrary choice of anchored points based on Dirac measure) and not suitable for high-dimensional interpolation (pointwise evaluation) for stochastic problems with arbitrary probability measure. Another recently developed method under the name of dimension-adaptive tensor-product integration [79] uses a generalized sparse grid construction scheme and employs hierarchical surplus from the construction as error indicators to automatically detect different importance and interaction of different dimensions. Although being essentially equivalent to the anchored ANOVA approach, it is more versatile with different choice of hierarchical surpluses and suitable for interpolation problems. Still, it is to blame for the drawback of running into stagnation phenomenon, where too early stop of the grid construction in some region occurs before arriving at the desired accuracy of approximation. Another drawback is it use one higher level of grid to assess the error indicators, resulting in a very heavy computational cost.

In this chapter, we adopt the more versatile dimension-adaptive algorithm based on hierarchical surpluses and generalized sparse grid construction for both integration and interpolation. However, we propose two remedies in addressing the drawbacks and enhancing both its efficiency and accuracy for solving different UQ problems. As for the first drawback of running into stagnation, a balanced greedy algorithm was suggested in [79] and [110], where a purely greedy criteria of choosing the next index by hierarchical surplus for grid construction is balanced by performing the conventional sparse grid construction. However, it is neither possible to choose an optimal balance weight nor feasible to use the same weight throughout the whole grid construction. Alternatively, we propose to carry out a verification procedure in order to get rid of the stagnation phenomenon. The basic idea is that whenever the construction is stopped at some region by meeting certain criteria, we check whether it should be continued by some verification algorithms specific to different dimensional problems. This approach avoids the difficulty in tuning the balanced weight parameter and works efficiently to get out of the stagnation region for grid construction at the appropriate moment.

The verification remedy has not yet been studied in the literature or applied in practice because it needs additional verification samples besides the ones used for assessing hierarchical surpluses in one higher level. This drawback is critical for large-scale UQ problems that already require large computational efforts in solving the underlying PDE model at one sample, as the second computational challenge mentioned before. In order to harness the computational burden, we employ a reduced basis method, which has been used in combination with ANOVA in [96], and develop an adaptive and weighted algorithm in the framework of the verified hierarchical approximation. The rationale of this computational approach is deeply rooted in probability theory: though the random inputs live in a high-dimensional space, the output of interest (statistics of these random inputs) may only lie in a low-dimensional manifold, for instance the arithmetic mean of a large number of independent random variables fulfilling certain conditions (e.g. having finite variance) converge to a (Gaussian) random variable, as guaranteed by the central limit theorem [64]. This fact enables us to construct a reduced bases space with a few number of bases while achieving high accuracy in approximating the high fidelity solution, e.g. finite element solution and the output of interest. Based on this idea and using the reduced basis method for parametrized PDEs [131, 178, 158, 87, 86, 49], we develop an adaptive greedy algorithm in combination with the verified dimension-adaptive hierarchical grid construction procedure to solve high-dimensional UQ problems. In order to take the arbitrary probability measure into account, we use a weighted a posteriori error bound for guiding the selection of the most representative bases [49]. This proves to be more efficient with much less bases in achieving the same approximation accuracy as

the a posteriori error bound without incorporating the weight.

By the end, an adaptive and reduced computational framework is developed in efficiently and accurately solving high-dimensional UQ problems that feature sparsity and reducibility. Application of the proposed framework in solving high-dimensional UQ problems based on more general PDE models, such as non-affine, non steady, non-compliant, non-coercive and nonlinear problems can be realized by resorting to specific techniques and computational approaches, e.g. empirical interpolation, primal-dual approach, supermizers enrichment, POD-greedy algorithm and Newton iteration, respectively, which will be summarized in this work. A series of numerical experiments featuring various properties for both functions and PDEs are carried out in demonstrating the efficiency and accuracy of our method and in comparing its computational performance to several other techniques.

This chapter is organized as follows. A family of UQ problems is introduced in section 5.1 based on a general formulation. For their numerical solution, two computational challenges are identified and briefly illustrated. Section 5.2 is devoted to the development of the verified dimension adaptive hierarchical approximation based on generalized sparse grid construction. Some remarks regarding the computational effectivity, efficiency and accuracy of this method in comparison with some other techniques are provided at the end of the section. In section 5.3, the adaptive and weighted reduced basis method is presented based on a simple PDE model. A large effort has been devoted to conducting a variety of numerical experiments in section 5.4, including 10 examples in 6 different topics that offer a rich diversity for demonstrating the accuracy and efficiency of the proposed computational framework and comparing them with several other techniques. In the last section, we close this chapter by drawing some conclusions based on the numerical experiments and providing a few further research perspectives for developing and applying the adaptive and reduced computational framework in solving more general high-dimensional UQ problems.

5.1 High-dimensional uncertainty quantification

In this section, we start with the presentation of several uncertainty quantification (UQ) problems that have been largely studied in the literature, and then we identify some common computational challenges in solving the UQ problems.

5.1.1 Formulation of UQ problems

Associated with the general stochastic PDEs (38) under finite dimensional noise assumption as introduced in the preliminary chapter, the quantities of interest may be the solution u, the solution restricted to a certain physical region or to the boundary, some functional, e.g. $s : u(y) \to s(y) \equiv s(u(y))$. Here is a list (far from exhaustive) of uncertainty quantification problems:

1. compute the probability density function or the cumulative distribution function of either u or s [72];
2. evaluate statistical moments, e.g., mean $\mathbb{E}[s]$, variance $\mathbb{V}[s] := \mathbb{E}[s^2] - (\mathbb{E}[s])^2$, etc. [10, 8];
3. perform derivative-based local sensitivity analysis, e.g., compute $du(y)/dy$ or $ds(y)/dy$ [182];
4. perform variance-based global sensitivity analysis, e.g., compute $\mathbb{V}_k[s]/\mathbb{V}[s]$, where $\mathbb{V}_k[s]$ is the variance of s from the contribution of the random variable $Y_k, k = 1, \ldots, K$ [34, 46];
5. perform risk analysis, e.g., for a given a critical value s_0, compute the failure probability [121, 41]

$$P(\omega \in \Omega : s(\omega) < s_0); \tag{5.1}$$

6. solve stochastic optimal control problems, e.g., the following minimization problem [47, 45, 197]

$$f = \arg\min_{f \in \mathscr{U}_{ad}} \mathscr{J}(u, f) \text{ such that } u \text{ satisfies problem (38)}, \tag{5.2}$$

where f is regarded as a deterministic control function living in an admissible space $\mathscr{U}_{ad}$, u_d is a given observation, α is a regularization parameter, and the cost functional is

$$\mathscr{J}(u, f) := ||u - u_d||_{\mathscr{H}^s} + \alpha ||f||_{\mathscr{L}^2}. \tag{5.3}$$

7. estimate a parameter by Bayesian inference, e.g., given experiment data u or s with certain noise η, evaluate the posteriori density ρ_{post} of a random coefficient a of $\mathscr{L}$ based on its priori density ρ_{pr} [18, 188, 211].

Etc...

From a numerical standpoint, the above UQ problems could be classified as follows: for problems 1, 3, 5, 7, we look for pointwise evaluation of the stochastic solution, i.e. compute $u(y)$ or $s(y)$ at many $y \in \Gamma$; problems 2, 4, 6 require the evaluation of statistical moments. Interpolation techniques are requested for the former class, integration techniques for the latter.

5.1.2 Computational challenges

In order to study the UQ problems introduced above, the underlying stochastic PDEs (38) have to be solved. Several methods have attracted large attention in recent years. This includes the non-intrusive Monte Carlo method and several variants, stochastic collocation method based on sparse grid techniques, the intrusive stochastic Galerkin method with generalized polynomial chaos, surrogate models by different model order reduction approaches, etc. [72, 80, 208, 149, 152, 25, 49, 50].

The Monte Carlo method is typically blamed for its slow convergence; on the other side, all the other methods that are expected to feature fast convergence face some common computational challenges. A critical one is *high dimensionality*, which requires an exponentially increasing number of collocation (for interpolation) or quadrature (for integration) nodes with growing stochastic dimensions. Figure 5.1 depicts the total number of nodes in tensor product (left) and sparse grid (right) structures with different probability dimensions. The left one reports the results in dimensions 1, 5, 10, 20, 50 and 100 with the number of nodes in each dimension increasing from 1 to 8, from which we can see that the number of nodes easily overpasses the capacity of computational power in relatively high dimensions, e.g., $2^{100} \approx 10^{30}$ nodes are needed for 100 dimensional case with only 2 nodes in each dimension. The results of sparse grid (Smolyak type with Clenshaw–Curtis nodes [207, 149], corresponding to Chebyshev-Gauss-Lobatto nodes in the context of spectral methods [35]) for dimension going up to 200, 500 and 1000 are displayed on the right of Figure 5.1. Compared to tensor product structure, the sparse grid structure considerably reduces the number of nodes, e.g., around 10^6 and 10^9 nodes are needed with 9 nodes in each dimension at sparse grid level 3 for 100 and 1000 dimensional cases. Nevertheless, only tens or hundreds of nodes are affordable in practical engineering problems when a full solve of the underlying PDEs is very expansive. This requirement prevents a direct use of sparse grid techniques for even moderate dimensional problems. This challenge is particularly relevant to UQ problems 6 and 7, namely optimization and inverse problems, for which many full solves (in the order of tens or hundreds) of the underlying PDEs, using some iteration method [161], have to be performed at each of a large number of nodes [43, 45].

Another computational challenge, which we would like to emphasize again, for solving most of the PDE-based UQ problems is that the numerical solution of the underlying PDE model might require a large computational effort: this is e.g. the case of multiscale and/or multiphysics problems. In these

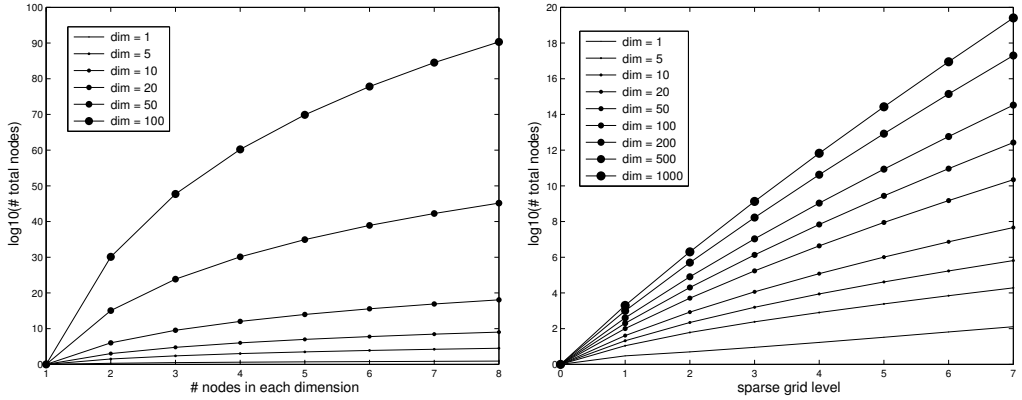

Figure 5.1: Number of collocation (for interpolation) or quadrature (for integration) nodes of tensor product structure (left) and sparse grid structure (right) for different probability dimensions.

circumstances, only a few tens or hundreds of the underlying PDEs can be fully solved, therefore preventing direct application of any method mentioned above in solving high-dimensional UQ problems, for which a large number (in the order of million and beyond) of PDEs have to be solved in order to evaluate the quantity of interest. This computational challenge is critical for UQ analysis in many practical engineering fields. Research in addressing this challenge in the context of high-dimensional UQ problems is still in its infancy [153].

5.2 Verified dimension adaptive hierarchical approximation

In this section, we present the dimension-adaptive tensor-product algorithm for hierarchical approximation of high-dimensional UQ problems based on the work [33, 79, 110]. Our original contribution is to identify the stagnation phenomenon in the hierarchical construction of a generalized sparse grid for this algorithm and propose a verified version of this algorithm in order to cure this undesirable behavior. Suitable error indicators (in particular, a new integration error indicator) are provided for interpolation and integration problems. Some comparisons with several other techniques, e.g. anisotropic sparse grid [148] and variance-based ANOVA (HDMR) [73, 128], are provided at the end of this section.

5.2.1 Hierarchical interpolation and integration in one dimension

For numerical interpolation of function $s : \Gamma \to \mathbb{R}$ in a one dimensional probability domain $\Gamma \subset \mathbb{R}$, we first pick a series of collocation nodes $y^j \in \Gamma$, $j = 0, \ldots, m$, ordered such that $y^1 < y^2 < \cdots < y^m$ and for any given node $y \in \Gamma$, we approximate the function value $s(y)$ by the following interpolation formula

$$s(y) \approx \mathscr{U} s(y) = \sum_{j=1}^{m} s(y^j) l^j(y), \tag{5.4}$$

where $\mathscr{U}$ is an interpolation operator; $l^j, 1 \le j \le m$ are basis functions that, depending on the regularity of the function s with respect to y in Γ, are either piecewise polynomials or global polynomials [164]. For instance, the piecewise linear polynomials most often used in approximating low regularity functions

are defined as

$$l^j(y) = \begin{cases} \dfrac{y - y^{j-1}}{y^j - y^{j-1}}, & \text{if } y \in [y^{j-1}, y^j], \quad j = 2, \ldots, m; \\ \dfrac{y^{j+1} - y}{y^{j+1} - y^j}, & \text{if } y \in [y^j, y^{j+1}], \quad j = 1, \ldots, m-1. \end{cases} \tag{5.5}$$

Though converging very slowly (thus requiring a large number of nodes for accurate approximation), these bases lead to uniform convergence when the nodes become dense in the domain Γ. As for the approximation of smooth functions, more suitable are the globally supported polynomials, for instance Lagrange polynomials defined as

$$l^j(y) = \prod_{l=1, l\neq j}^{m} \frac{y - y^l}{y^j - y^l}, \quad j = 1, \ldots, m, \tag{5.6}$$

for a suitable set of nodes such as Gauss quadrature nodes, Chebyshev or Clenshaw–Curtis nodes [164, 199]. For instance, the Clenshaw–Curtis nodes in the interval $[-1, 1]$ are given by

$$y^j = \cos\left(\frac{j-1}{m-1}\pi\right), \quad 1 \le j \le m. \tag{5.7}$$

Let $i \in \mathbb{N}_+$ denote the grid level, Θ^i denote the set of collocation nodes on the grid of level i, with $m(i)$ being the number of nodes on the grid of level i, for instance

$$m(1) = 1; \quad m(i) = 2^{i-1} + 1, \quad i \ge 1. \tag{5.8}$$

We consider nested set of nodes, i.e. $\Theta^i \subset \Theta^{i+1}, i = 1, 2, \ldots, q$ with $q \in \mathbb{N}_+$. In this way, the hierarchical interpolation formula can be written as [33, 110]

$$s(y) \approx \mathscr{U}^q s(y) = \sum_{i=1}^{q} \triangle^i s(y), \tag{5.9}$$

where $\triangle^i$ is the difference of interpolation operators at two successive levels, defined as

$$\triangle^i = \mathscr{U}^i - \mathscr{U}^{i-1}, \quad 1 \le i \le q, \tag{5.10}$$

being $\mathscr{U}^0 = 0$ and $\mathscr{U}^i$ the interpolation operator supported on Θ^i. For notational convenience, let us define $\Theta^i_\triangle = \Theta^i \setminus \Theta^{i-1}, 1 \le i \le q$ with $\Theta^{i-1} := \emptyset$, and reorder the collocation nodes $y^1, \ldots, y^{m(q)}$ in $\Theta^q = \cup_{i=1}^q \Theta^i_\triangle$ level by level in such a way that $y^i_j \in \Theta^i_\triangle, 1 \le i \le q, 1 \le j \le m(i) - m(i-1)$ with $m(0) = 0$. Corresponding to the reordering of the collocation nodes, we denote the basis functions as $l^i_j, 1 \le i \le q, 1 \le j \le m(i) - m(i-1)$. Thanks to the hierarchical structure $\Theta^{i-1} \subset \Theta^i$, $\mathscr{U}^{i-1} s = \mathscr{U}^i \circ \mathscr{U}^{i-1} s$. Moreover, $s(y^i_j) = \mathscr{U}^{i-1} s(y^i_j)$ for $y^i_j \in \Theta^{i-1}$. Therefore, the interpolation operator (5.9) can be rewritten as

$$\mathscr{U}^q s(y) = \sum_{i=1}^{q} \left(\mathscr{U}^i s(y) - \mathscr{U}^i \circ \mathscr{U}^{i-1} s(y) \right) = \sum_{i=1}^{q} \sum_{y^i_j \in \Theta^i_\triangle} \underbrace{\left(s(y^i_j) - \mathscr{U}^{i-1} s(y^i_j) \right)}_{s^i_j} l^i_j(y). \tag{5.11}$$

The real number s^i_j is called *hierarchical surplus* [33], which provides a measure of the interpolation accuracy of the interpolant $\mathscr{U}^{i-1}$ on the successive grid of level i. When this surplus is small, a relatively accurate interpolation is obtained at the corresponding node and grid level.

The construction of interpolation based on nodal basis (left) and hierarchical basis (right) in the form of piecewise linear polynomials are illustrated in Figure 5.2, from which we can see that the interpolation constructed by the two approaches are equivalent in evaluating function values at any $y \in \Gamma$. However

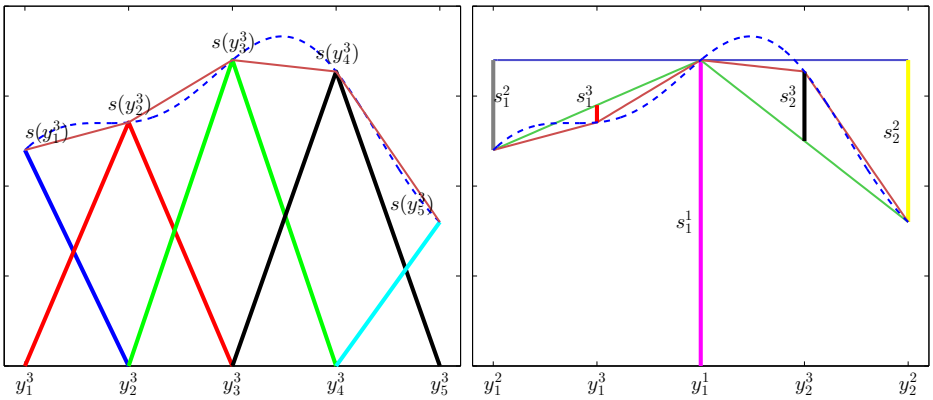

Figure 5.2: Construction of interpolation based on nodal basis (left) and hierarchical basis (right).

the latter also provides an estimate of interpolation error via hierarchical surpluses. For instance, s_1^2 and s_2^2 are the errors of a constant approximation of the function at nodes y_1^2 and y_2^2, which can provide a rough estimate of the interpolation accuracy.

As for numerical integration in evaluating statistical moments, we can take advantage of the interpolation formula (5.11) and assess the accuracy of integration by hierarchical surplus. For instance, the expectation of the function s can be computed by

$$\mathbb{E}[s] \approx \mathbb{E}[\mathscr{U}^q s] = \sum_{i=1}^{q} \sum_{y_j^i \in \Theta_\Delta^i} s_j^i w_j^i, \tag{5.12}$$

where the quadrature weights w_j^i are computed by

$$w_j^i = \int_\Gamma l_j^i(y)\rho(y)dy, \quad 1 \le i \le q, 1 \le j \le m(i) - m(i-1) \tag{5.13}$$

using suitable quadrature rules depending on the choice of different collocation nodes [164]. Similarly, the kth ($k \ge 2$) order statistical moments can be evaluated by setting the hierarchical surpluses as $s_j^i = s^k(y_j^i) - \mathscr{U}^{i-1}s^k(y_j^i), 1 \le i \le q, 1 \le j \le m(i) - m(i-1)$.

Based on the hierarchical surplus s_j^i, we may define the interpolation error $\mathscr{E}_i$ and the integration error $\mathscr{E}_e$ as

$$\mathscr{E}_i := \max_{1 \le j \le m(q)-m(q-1)} |s_j^q|, \mathscr{E}_e := \sum_{y_j^q \in \Theta_\Delta^q} s_j^q w_j^q. \tag{5.14}$$

There quantities can be used as error indicators in adaptively constructing the interpolation formula (5.11) and integration formula (5.12), respectively. However, one drawback of using the hierarchical surplus as error indicator is that the error may be underestimated where the refinement of the grid has stagnated at an early stage. For instance, in the interpolation constructed from hierarchical basis, the interpolated function values coincide with the true function values at the nodes y_1^3 and y_2^3 as shown in Figure 5.3 in two cases – hierarchical interpolation based on locally supported piecewise linear polynomials and globally supported Lagrange polynomials – so that the hierarchical surplus s_1^3 and s_2^3 become zero, leading to the termination of the adaptive construction of the grid to the next level even the approximation is far from accurate in almost all the region.

In order to get rid of this stagnation phenomenon, we propose to check the interpolation accuracy

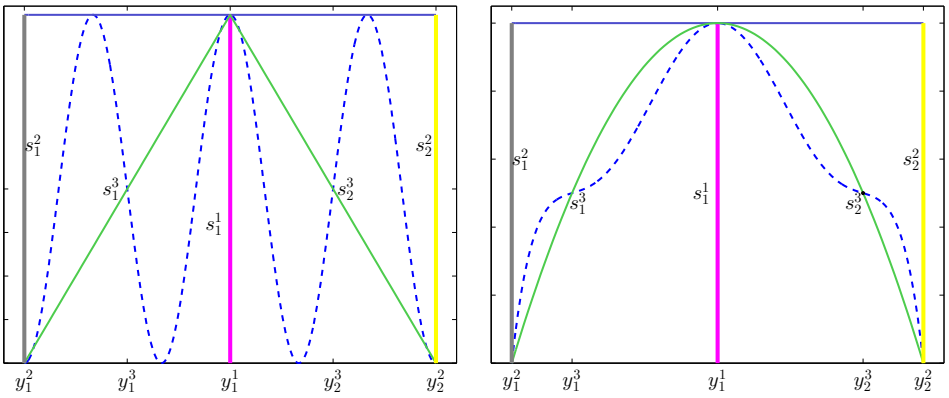

Figure 5.3: Stagnation phenomenons for hierarchical interpolation. Left: piecewise linear polynomials based on equidistant nodes; right: Lagrange polynomials based on Clenshaw–Curtis nodes.

(via hierarchical surplus) at the nodes of the next grid level. If the error indicator is larger than the error tolerance, we continue the construction procedure to the next level. Otherwise, we stop. The construction procedure of hierarchical interpolation stopped by satisfying certain error tolerance is summarized in Algorithm 7, which can also be used for hierarchical integration with the interpolation error indicator $\mathscr{E}_i$ replaced by the integration error indicator $\mathscr{E}_e$.

Remark 5.2.1 *There is the possibility that the error indicator in the next grid level might still be smaller than the error tolerance when the approximation is not good enough somewhere, e.g. for continuous functions displaying high oscillation at some very locally supported region that has not been explored by interpolation nodes. In this case, which is also difficult to handle by other interpolation techniques, we may randomly select a certain number of nodes to perform further verification besides using the nodes in the next grid level, expecting that the region can be touched by these nodes with large possibility. This empirical idea needs to be further investigated to balance computational efficiency and accuracy.*

5.2.2 Hierarchical Smolyak sparse grid in multiple dimensions

In multiple dimensional numerical interpolation, when $\Gamma \subset \mathbb{R}^K, K = 2,3,\ldots$, the univariate interpolation formula (5.9) can be straightforwardly extended as the tensor product interpolation [8]

$$\mathscr{I}_q s(y) := \left(\mathscr{U}_1^q \otimes \cdots \otimes \mathscr{U}_K^q\right) s(y) = \sum_{i_1=1}^{q} \cdots \sum_{i_K=1}^{q} \left(\triangle_1^{i_1} \otimes \cdots \otimes \triangle_K^{i_K}\right) s(y), \tag{5.15}$$

where $\mathscr{U}_k^{q_k}$ and $\triangle_k^{i_k}$ are the univariate interpolation and difference operators in dimension $k = 1,\ldots,K$. Since, as shown in Figure 5.1, the tensor product interpolation needs too many collocation nodes, the Smolyak sparse grid interpolation [191]

$$\mathscr{S}_q s(y) = \sum_{|\mathbf{i}| \le q} \left(\triangle_1^{i_1} \otimes \cdots \otimes \triangle_K^{i_K}\right) s(y) \tag{5.16}$$

is employed to reduce the number of nodes, where the multivariate index $\mathbf{i} = (i_1,\ldots,i_K) \in \mathbb{N}_+^K$ represents the multi-dimensional grid level with *interaction level* $|\mathbf{i}| = i_1 + \cdots + i_K$; $q \ge K$ denotes the *total level* of the isotropic sparse grid. To obtain a hierarchical representation of the sparse grid interpolation (5.16),

Algorithm 7 Verified hierarchical interpolation in one dimension

1: **procedure** INITIALIZATION:
2: specify error tolerance ε_t, type of interpolation bases $l(y)$ and nodes y, specify function $m(i)$;
3: specify maximum level q, set $i = 1$, $\Theta^1 = \{y_j^1, 1 \le j \le m(1)\}$ and evaluate $s_1^1 = s(y_j^1)$;
4: set $\mathscr{E}_i = 2\varepsilon_t$;
5: **end procedure**
6: **procedure** CONSTRUCTION:
7: **while** $\mathscr{E}_i > \varepsilon_t$ and $i \le q$ **do**
8: provide the set of nodes $\Theta_\triangle^i = \{y_j^i, 1 \le j \le m(i) - m(i-1)\}$;
9: for all $y_j^i \in \Theta_\triangle^i$, evaluate function values $s(y_j^i)$ and the interpolation $\mathscr{U}^{i-1}s(y_j^i)$ by (5.11);
10: compute the hierarchical surpluses $s_j^i = s(y_j^i) - \mathscr{U}^{i-1}s(y_j^i)$ and error indicator $\mathscr{E}_i$ by (5.14);
11: **procedure** VERIFICATION:
12: **if** $\mathscr{E}_i \le \varepsilon_t$ **then**
13: go to the next level $i = i + 1$ and repeat the steps in line 8 - line 10;
14: **end if**
15: **end procedure**
16: **if** $\mathscr{E}_i \le \varepsilon_t$ **then**
17: **return** .
18: **else**
19: go to the next level $i = i + 1$;
20: **end if**
21: **end while**
22: **end procedure**

we split it as follows

$$\mathscr{S}_q s(y) = \mathscr{S}_{q-1} s(y) + \triangle\mathscr{S}_q s(y), \text{ with } \triangle\mathscr{S}_q s(y) := \sum_{|\mathbf{i}|=q} \left(\triangle_1^{i_1} \otimes \cdots \otimes \triangle_K^{i_K}\right) s(y). \tag{5.17}$$

A more explicit expansion for $\triangle\mathscr{S}_q s(y)$ is

$$\triangle\mathscr{S}_q s(y) = \sum_{|\mathbf{i}|=q} \sum_{\mathbf{j}} \underbrace{\left(s(y_{j_1}^{i_1}, \ldots, y_{j_K}^{i_K}) - \mathscr{S}_{q-1} s(y_{j_1}^{i_1}, \ldots, y_{j_K}^{i_K})\right)}_{s_{\mathbf{j}}^{\mathbf{i}}} \underbrace{\left(l_{j_1}^{i_1}(y_1) \otimes \cdots \otimes l_{j_K}^{i_K}(y_K)\right)}_{l_{\mathbf{j}}^{\mathbf{i}}}. \tag{5.18}$$

Here, $y_{j_k}^{i_k} \in \Theta_\triangle^{i_k}$ is the j_kth node of grid level i_k in dimension $k = 1, \ldots, K$ and $l_{j_k}^{i_k}$ is the corresponding basis function; $s_{\mathbf{j}}^{\mathbf{i}}$ is the hierarchical surplus at node $\mathbf{j}$ of grid level $\mathbf{i}$, which can be used as an error indicator for the construction of adaptive sparse grid. The hierarchical construction of the two dimensional full grid and sparse grid based on Clenshaw–Curtis nodes is illustrated in Figure 5.4 (the size of markers indicates the level of grid), where 1, 4, 8 nodes are added in the 1st, 2nd and 3rd level of sparse grid corresponding to $|\mathbf{i}| = 2, 3, 4$ for the dimension $K = 2$. Note that the sparse grid contains less nodes than the full grid and achieves the same approximation accuracy by taking advantage of the assumption that the interaction level of different dimensions stays small, especially in high-dimensional case. For instance the interpolation (5.15) based on the full grid and (5.16) on the sparse grid in Figure 5.4 can reconstruct exactly any polynomial in the form $y_1^{m(i_1)-1} y_2^{m(i_2)-1}$ such that $i_1 + i_2 \le 4$. However sparse grid interpolation will produce approximation error when $i_1 + i_2 > 4$, in which case the full grid interpolation is exact as long as $i_1 \le 3$ and $i_2 \le 3$. Problems featuring dimensions independent to each other or small interaction level are called *separable* dimensional problems; for them the sparse grid approximation is more favorable. In order to detect the interaction level of different dimensions and

enrich the nodes accordingly, the hierarchical surplus $s_{\mathbf{j}}^{\mathbf{i}}$ can be employed directly, as we will see in the next sections.

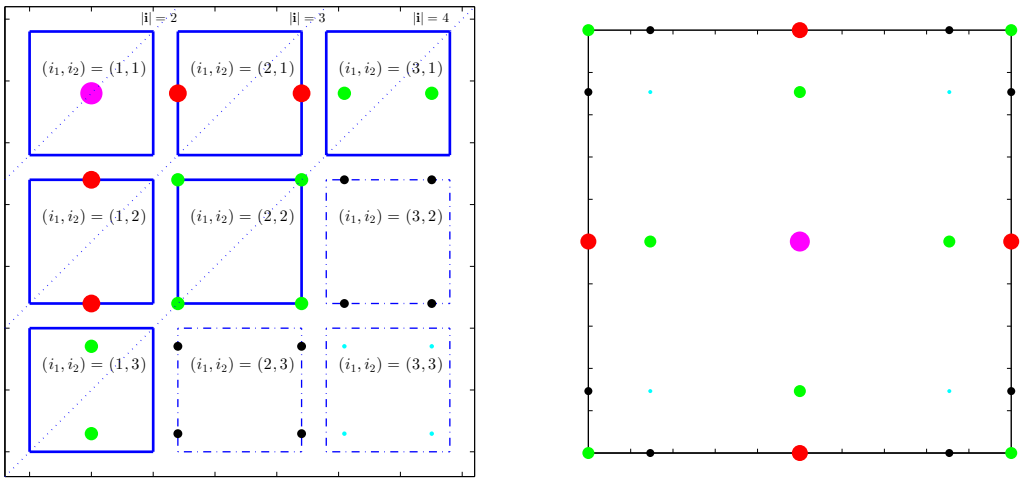

Figure 5.4: Illustration of hierarchical construction of full grid and sparse grid in two dimensions. Left: construction procedure with solid box indicating sparse grid and with the additional dashed box full grid; right: sparse grid (nodes of the first three largest markers in size) and full grid (all nodes).

As for the multivariate numerical integration based on the hierarchical sparse grid interpolation formula (5.16), we obtain the integration formula assembled in a hierarchical form as

$$\mathbb{E}[s] \approx \mathbb{E}[\mathscr{S}_q s] = \sum_{p=K}^{q} \sum_{|\mathbf{i}|=p} \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} w_{\mathbf{j}}^{\mathbf{i}}, \tag{5.19}$$

where the weight

$$w_{\mathbf{j}}^{\mathbf{i}} = \int_{\Gamma} \left(l_{j_1}^{i_1}(y_1) \otimes \cdots \otimes l_{j_K}^{i_K}(y_K) \right) \rho(y) dy, \tag{5.20}$$

is computed approximately by a suitable quadrature rule depending on the choice of nodes. Provided that the probability density function is separable, i.e. $\rho(y) = \prod_{k=1}^{K} \rho_k(y_k)$, we have

$$w_{\mathbf{j}}^{\mathbf{i}} = \prod_{k=1}^{K} w_{j_k}^{i_k}, \text{ with } w_{j_k}^{i_k} = \int_{\Gamma_k} l_{j_k}^{i_k}(y_k) \rho(y_k) dy_k, 1 \le k \le K, \tag{5.21}$$

which can be precomputed and stored for the sake of computational efficiency. We remark that, when the function s is continuous in Γ, the hierarchical surpluses $s_{\mathbf{j}}^{\mathbf{i}} \to 0$ with $|\mathbf{i}| = q$ as the total approximation level $q \to \infty$ for both interpolation and integration. Therefore, we may estimate the sparse grid interpolation error $\mathscr{E}_i$ and integration error $\mathscr{E}_e$ respectively as

$$\mathscr{E}_i = \max_{|\mathbf{i}|=q,\mathbf{j}} |s_{\mathbf{j}}^{\mathbf{i}}| \text{ and } \mathscr{E}_e = \sum_{|\mathbf{i}|=q} \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} w_{\mathbf{j}}^{\mathbf{i}}. \tag{5.22}$$

5.2.3 Dimension adaptation for high-dimensional problems

As we can observe from Figure 5.1, sparse grid introduced in the last section considerably reduces the total number of collocation nodes, making it advantageous to solve moderate (several tens [207, 149]) dimensional approximation problems as well as high (several hundreds or beyond [110]) but separable dimensional problems. However, when the dimensions become too high and the interaction level

of different dimensions becomes big, sparse grid techniques are difficult to be directly applied due to computational constraint, e.g. around 10^{12} nodes are needed to approximate 100 dimensional problems with interaction level 7, see Figure 5.1. In this section, we take advantage of the hierarchical surplus and adopt the dimension-adaptive approach [33, 110] to cope with high-dimensional approximation problems. In particular in Algorithm 8, we will propose a high-dimensional verification procedure to deal with possible stagnation phenomena and a new adaptive criterion more suitable for high-dimensional integration problems. Other techniques are also considered for comparison with our proposed approach in a series of remarks.

The sparse grid interpolation based on the difference operator (5.16) is constructed in an isotropic manner due to the restriction $|\mathbf{i}| \leq q$. For a more general construction of sparse grid interpolation, we break the isotropic restriction and pose only an *admissibility condition* to satisfy the essential property (5.9) of the hierarchical representation [33, 110]. The set of indices $S \subset \mathbb{N}_+^K$ is called admissible if for each $\mathbf{i} \in S$, the indices $\mathbf{i} - \mathbf{e}_k \in S$ for all $k = 1, \ldots, K$ such that $i_k > 1$. Note that $\mathbf{e}_k \in \{0,1\}^K$ with the kth element as one and the other elements zero. The sparse grid constructed from an admissible set is called *generalized sparse grid* [33], which includes both the isotropic sparse grid with the index set $S_i := \{\mathbf{i} \in \mathbb{N}_+^K : |\mathbf{i}| \leq q\}$ and the full tensor product grid with the index set $S_t := \{\mathbf{i} \in \mathbb{N}_+^K : i_k \leq q, 1 \leq k \leq N\}$. In the admissible index set S_m, being m the cardinality of S_m, we can write the generalized sparse grid interpolation formula (5.16) in a hierarchical way as

$$\mathscr{S}_g s(y) = \sum_{\mathbf{i} \in S_m} \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} l_{\mathbf{j}}^{\mathbf{i}}. \tag{5.23}$$

Correspondingly, the generalized sparse grid integration formula (5.19) can be written as

$$\mathbb{E}[s] \approx \mathbb{E}[\mathscr{S}_g s] = \sum_{\mathbf{i} \in S_m} \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} w_{\mathbf{j}}^{\mathbf{i}}. \tag{5.24}$$

At the root level, we set $S_1 = \{\mathbf{1}\}$, in which case the hierarchical surplus $s_{\mathbf{j}}^{\mathbf{i}}$ takes the value of the function s at $y_{\mathbf{j}}^{\mathbf{i}}$. At the next level, we enrich S_1 with the indices of the forward neighborhood of the root index 1, i.e. $S_m = \{\mathbf{1}, \mathbf{1} + \mathbf{e}_k, 1 \leq k \leq K\}$ with $m = K + 1$ and compute the hierarchical surplus $s_{\mathbf{j}}^{\mathbf{i}}$ for $\mathbf{i} \in S_m \setminus \{\mathbf{1}\}$. Afterwards, the index $\mathbf{i}$ is picked corresponding to the largest error indicator defined via $s_{\mathbf{j}}^{\mathbf{i}}$ and enrich S_m with the indices from $\{\mathbf{i} + \mathbf{e}_k, 1 \leq k \leq K\}$ such that S_m remains admissible. Here, we follow [110] to use the averaged hierarchical surplus as the error indicator to pick $\mathbf{i}$

$$\mathbf{i} = \underset{\mathbf{i} \in \mathscr{A}}{\operatorname{argmax}} \mathscr{E}_i(\mathbf{i}) \text{ with } \mathscr{E}_i(\mathbf{i}) := \frac{1}{n(\mathbf{i})} \sum_{\mathbf{j}} |s_{\mathbf{j}}^{\mathbf{i}}|, \tag{5.25}$$

where $n(\mathbf{i})$ is the number of nodes added due to the enrichment of the index $\mathbf{i} \in S_m$; $\mathscr{A} \subset S_m$ is the *active index set* collecting all the indices in S_m whose forward neighbors has not been processed. The complementary of $\mathscr{A}$ is called *old index set* with notation $\mathscr{O} = S_m \setminus \mathscr{A}$. After the enrichment, we move the index $\mathbf{i}$ from $\mathscr{A}$ to $\mathscr{O}$ and add the admissible forward neighbors of $\mathbf{i}$ into $\mathscr{A}$ and S_m. Subsequently, we carry out the same procedure to enrich S_m in an adaptive way until satisfying certain stopping criteria, e.g. error tolerance or maximum number of nodes. As for high-dimensional integration, we propose to build the dimension-adaptive sparse grid based on a new error indicator

$$\mathscr{E}_e(\mathbf{i}) := \frac{1}{n(\mathbf{i})} \left| \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} w_{\mathbf{j}}^{\mathbf{i}} \right|, \tag{5.26}$$

which takes into account three factors: the hierarchical surpluses, the quadrature weights that correspond to arbitrary probability density function and the work contribution by dividing $n(\mathbf{i})$. We remark that the error indicator (5.26) tends to underestimate the integral error since only one index is

considered. We provide a more reasonable estimate for the integral error as

$$\mathscr{E}_e(\mathscr{A}) = \left| \sum_{\mathbf{i}\in\mathscr{A}} \sum_{\mathbf{j}} s_{\mathbf{j}}^{\mathbf{i}} w_{\mathbf{j}}^{\mathbf{i}} \right|. \tag{5.27}$$

The construction of the generalized sparse grid in the above procedure not only automatically detects the importance and interaction of different dimensions but also adaptively builds an anisotropic sparse grid without any a priori knowledge or a posteriori processing. However, as in the one dimensional case, stagnation of the adaptive construction might occur at some index $\mathbf{i} \in \mathscr{A}$, thus preventing accurate approximation at an early stage of the hierarchical construction. To overcome this drawback, several algorithms have been proposed in [33, 110] to keep the balance between the purely greedy adaptive construction and a conservative grid construction. For instance, given a weight parameter $w \in [0,1]$, we add the forward neighbors of the index $\mathbf{i}$, regardless of $\mathscr{E}_i$ or $\mathscr{E}_e$, to the active index set $\mathscr{A}$ as long as [110]

$$\frac{\min_{\mathbf{i}\in\mathscr{A}} |\mathbf{i}|}{\max_{\mathbf{i}\in\mathscr{A}\cup\mathscr{O}} |\mathbf{i}|} \leq (1-w), \tag{5.28}$$

where $w = 1$ corresponds to the purely greedy adaptive construction and $w = 0$ the conservative grid construction. Nevertheless, it is not easy to decide what value the weight parameter w should take, leading to either deterioration of the efficiency of the adaptive construction or possible stagnation persisting until a very fine grid has been built. We propose here, as in one dimensional case in Algorithm 7, to perform the verification for each index in the active index set in order to get out of the stagnation set as well as retain the efficiency of the adaptive construction. Our verified dimension-adaptive hierarchical algorithm for interpolation is summarized in Algorithm 8 for high-dimensional interpolation problems. The same algorithm can be adapted for integration by simply replacing the interpolation error indicator $\mathscr{E}_i$ in (5.25) by the integration error indicator $\mathscr{E}_e$ in (5.26). We remark that for function-based high-dimensional interpolation problems, the verified dimension-adaptive hierarchical interpolation algorithm 8 is employed, while for PDE-based interpolation problems, we propose to apply the certified reduced basis method developed in section 5.3, which produces more accurate approximation results with certification in practice.

As pointed out in [73, 128], in addition to stagnation for the dimension-adaptive hierarchical construction, another drawback is that it involves evaluating the function $s(y)$ at one higher grid level in each dimension in order to assess the error indicator. This is rather costly, especially for high-dimensional uncertainty quantification problems with verification procedure, where the evaluation at each of a large number of nodes requires a full solve of the underlying PDE. Fortunately, this computational burden can be considerably alleviated by using the adaptive reduced basis method that will be developed in section 5.3, where full solve of the underlying PDE model is replaced by a very cheap solve of a reduced model. The corresponding dimension-adaptive approach with verification becomes much more appealing.

5.2.4 Comparison remarks

In order to take the importance of different dimensions into consideration, an anisotropic sparse grid was proposed in [148] by choosing the index set for the construction of the grid as

$$S_{\boldsymbol{\alpha}} = \left\{ \mathbf{i} \in \mathbb{N}_+^K : \sum_{k=1}^{K} (i_k - 1)\alpha_k \leq q \min_{1\leq k\leq K} \alpha_k \right\}. \tag{5.29}$$

The multivariate weight $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K)$ indicates the importance of different dimensions and $q \in \mathbb{N}$ represents the grid level; its choice is a challenging task. The authors suggested two ways to specify $\boldsymbol{\alpha}$

Algorithm 8 Verified dimension-adaptive hierarchical algorithm for interpolation

1: **procedure** INITIALIZATION:
2: specify error tolerance ε_t, types of interpolation bases $l(y)$ and nodes y, specify function $m(i)$;
3: specify maximum number of nodes M, set $\mathbf{i}=\mathbf{1}$, compute $\Theta^{\mathbf{1}}$ and evaluate $s^{\mathbf{1}}_{\mathbf{j}}=s(y^{\mathbf{1}}_{\mathbf{j}}), y^{\mathbf{1}}_{\mathbf{j}}\in\Theta^{\mathbf{1}}$;
4: set $\mathscr{E}_i=2\varepsilon_t$, $m=\#|\Theta^{\mathbf{1}}|$, $\mathscr{A}=\{\mathbf{1}\}$, $\mathscr{O}=\emptyset$, $S_m=\mathscr{O}\cup\mathscr{A}$;
5: **end procedure**
6: **procedure** CONSTRUCTION:
7: **while** $\mathscr{E}_i>\varepsilon_t$ and $m\le M$ **do**
8: set $\mathscr{O}=\mathscr{O}\cup\{\mathbf{i}\}$, $\mathscr{A}=\mathscr{A}\setminus\{\mathbf{i}\}$ and enrich $\mathscr{A}$ by the admissible forward neighbors of $\mathbf{i}$;
9: compute the set of nodes $\Theta_\triangle$ different from old nodes at the newly added indices of $\mathscr{A}$;
10: for all $y^{\mathbf{i}}_{\mathbf{j}}\in\Theta_\triangle$, evaluate function values $s(y^{\mathbf{i}}_{\mathbf{j}})$ and the interpolation $\mathscr{S}_g s(y^{\mathbf{i}}_{\mathbf{j}})$ by (5.23);
11: compute the hierarchical surpluses $s^{\mathbf{i}}_{\mathbf{j}}=s(y^{\mathbf{i}}_{\mathbf{j}})-\mathscr{S}_g s(y^{\mathbf{i}}_{\mathbf{j}})$ and error indicator $\mathscr{E}_i$ by (5.25);
12: increase the number of nodes $m=m+\#|\Theta_\triangle|$, set the total index set $S_m=\mathscr{A}\cup\mathscr{O}$;
13: **procedure** VERIFICATION:
14: **for** $\mathbf{i}_v\in\mathscr{A}$ **do**
15: **if** $\mathscr{E}_i(\mathbf{i}_v)\le\varepsilon_t$ **then**
16: set the admissible forward neighbors of $\mathbf{i}_v$ as $\mathscr{A}_v$;
17: compute the set of added nodes $\Theta_\triangle$ for all indices in $\mathscr{A}_v$;
18: repeat lines 10 and 11 with $\mathscr{A}_v$ in (5.25) to get $\mathscr{E}_i$ in $\mathscr{A}_v$;
19: set $\mathscr{O}=\mathscr{O}\cup\{\mathbf{i}_v\}$, $\mathscr{A}=\mathscr{A}\setminus\{\mathbf{i}_v\}$, $\mathscr{E}_i^m=\max_{\mathbf{i}_m\in\mathscr{A}_v}\mathscr{E}_i(\mathbf{i}_m)$;
20: **if** $\mathscr{E}_i^m>\varepsilon_t$ **then**
21: enrich the active set $\mathscr{A}=\mathscr{A}\cup\mathscr{A}_v$ and repeat line 12;
22: **end if**
23: **end if**
24: **end for**
25: **end procedure**
26: pick the next index $\mathbf{i}$ such that $\mathbf{i}=\operatorname{argmax}_{\mathbf{i}\in\mathscr{A}}\mathscr{E}_i(\mathbf{i})$;
27: **if** $\mathscr{E}_i(\mathbf{i})\le\varepsilon_t$ **then**
28: **return** .
29: **end if**
30: **end while**
31: **end procedure**

in [148]. In those (simple) cases where a priori estimate for the Lagrange interpolation error exist, e.g.

$$\sup_{y_k\in\Gamma_k}|s(y_k)-\mathscr{U}_k s(y_k)|\le C_k e^{-2m_k g(k)},\quad 1\le k\le K, \tag{5.30}$$

being $\mathscr{U}_k$ the Lagrange interpolation operator and m_k the number of interpolation nodes in dimension k, the weights can be set as $\alpha_k=g(k),1\le k\le K$. An alternative way to estimate this weight is to perform a posteriori analysis by computing the outputs of interest at a series of collocation nodes and fitting the convergence rate in each dimension. Nevertheless, a posteriori estimate based on error fitting in each dimension can not identify the interaction effect among different dimensions and thus may lead to either not efficient anisotropic sparse grid construction or not accurate approximation. Moreover, the interpolation error may not decay exponentially with respect to the number of nodes for non smooth problems, and no general rule has been proposed for estimating the weight in these circumstances. In comparison, the dimension-adaptive construction of the sparse grid approximation based on hierarchical surpluses does not need to estimate the weights. Instead, it can automatically

detect the weight as well as the interaction level among different dimensions as a byproduct of the construction procedure [110].

Another technique to deal with high-dimensional approximation problems is based on ANOVA or HDMR, where the output of interest s can be decomposed into a series of additive functions (in total 2^K) incorporating all the 2^K possible interactions of different dimensions [88], written as

$$s(y) = s_0 + \sum_{1\le k_1\le K} s_{k_1}(y_{k_1}) + \sum_{1\le k_1<k_2\le K} s_{k_1,k_2}(y_{k_1},y_{k_2}) + \cdots + s_{k_1,\ldots,k_K}(y_{k_1},\ldots,y_{k_K}), \tag{5.31}$$

with

$$s_0 = \int_\Gamma s(y)d\mu(y), s_{k_1} = \int_{\Gamma^*_{k_1}} s(y)d\mu(y^*_{k_1}) - s_0, s_{k_1,k_2} = \int_{\Gamma^*_{k_1,k_2}} s(y)d\mu(y^*_{k_1,k_2}) - s_0 - s_k, \ldots, \tag{5.32}$$

with $y^*_{k_1} \in \Gamma^*_{k_1}$ in $K-1$ dimensional probability domain except Γ_{k_1}, $y^*_{k_1,k_2} \in \Gamma^*_{k_1,k_2}$ in $K-2$ dimensional probability domain except $\Gamma_{k_1}\times\Gamma_{k_2}$, and so on. Moreover, the variance of the function s admits the same expansion as in (5.31). It is known that there are only a few functions involving a limited number of dimensions play the majority role measured by variance when the function s displays distinctive importance and interaction in different dimensions [88]. Therefore, the high-dimensional approximation problem can be approximated by a series of low-dimensional approximation problems, leading to the development of ANOVA (HDMR) based dimension-adaptive algorithms [98, 71, 73, 128]. However, when the measure μ is the Lebesgue measure, high-dimensional integration has to be carried out in order to evaluate $s_0, s_{k_1}, \ldots$. Alternatively, when μ is a Dirac measure at some anchor point $\bar{y}\in\Gamma$, the expansion (5.31) takes the name of anchored-ANOVA [73] (or cut-HDMR [128]) expansion, which can substantially reduce the computational effort. However, there is no general rule to pick the anchor point, which is critical for accurate approximation and easily results in large error as pointed out in [192]. A single point - centroid of the lowest dimensional tensorial Gaussian quadrature - was suggested as the anchor point in [78]; improvement was also made in [90] by using a screening method, basically selecting several anchor points and taking the average in order to enhance the robustness, which might still not be satisfactory as our numerical examples in section 5.4 will reveal. Moreover, these variance-based techniques are primarily developed for solving integration problems, which may not be suitable when dealing with pointwise interpolation problems. In contrast, these drawbacks are not faced by the verified dimension-adaptive hierarchical Algorithm 8 that can be used for both high-dimensional interpolation and integration by choosing different error indicators. As a matter of fact, the hierarchical grid construction Algorithm 8 governed by different error indicators plays an equivalent role as automatically decomposing the targeted function into a series of additive functions involving limited dimensions indicated by the interaction of grid level among different dimensions, as demonstrated in the numerical experiments in section 5.4.

5.3 Adaptive and weighted reduced basis method

As mentioned in section 5.1.2, solving PDE-based UQ problems faces another critical computational challenge when the underlying PDEs are very expensive to solve. In this circumstance, non of the computational techniques presented in section 5.2 can be directly applied to deal with high-dimensional UQ problems. In order to tackle this difficulty, we exploit the property that the outputs of interest of the underlying PDEs may live in low-dimensional manifold even though the random inputs are from high-dimensional space. This property, which is known as reducibility, is quite common in practice and is essentially supported by *central limit theorem* and *law of large numbers* in the core of probability theory [64]. In this section, we develop an adaptive and weighted reduced basis method in combination with the hierarchical approximation to efficiently solve high-dimensional UQ problems.

The presentation of the reduced basis method follows the same lines as in chapter 1, section 1.3, for a

simple model, which we recall as

$$-\nabla(a(x,y)\nabla u) = f(x,y) \quad (x,y) \in D \times \Gamma, \tag{5.33}$$

where we make the affine assumption that

$$a(x,y) = \sum_{q=1}^{Q_a} \Theta_q^a(y) a_q(x) \text{ and } f(x,y) = \sum_{q=1}^{Q_f} \Theta_q^f(y) f_q(x). \tag{5.34}$$

For the purpose of solving high-dimensional problems, we propose an adaptive greedy algorithm based on the hierarchical approximation in Algorithm 9, where $\mathscr{E}_r : \Gamma \to \mathbb{R}$ is an a posteriori error bound depending on the quantities of interest. For pointwise quantities, we use the a posteriori error bound developed in section 1.3.3 of chapter 1, while for integral quantities, we apply the weighted scheme developed in section 2.1 of chapter 2: more explicitly, the a posteriori error bound and the weighted a posteriori error bound are given by

$$\triangle_N^s(y) := ||\hat{e}(y)||_X^2 / \alpha(y), \tag{5.35}$$

and

$$\triangle_N^{\rho,s}(y) = \rho(y)\triangle_N^s(y), \tag{5.36}$$

respectively, where $\hat{e}$ is the Riesz representation of the residual as given in (2.9) and α is the coercivity constant and can be computed in the way introduced in section 1.3.3 of chapter 1; ρ is the probability density function of the random vector y.

Algorithm 9 Adaptive greedy algorithm

1: **procedure** INITIALIZATION:
2: specify error tolerance ϵ_t, solve (46) at each $y \in \Theta^{\mathbf{1}}$ and construct $X_N = \text{span}\{u(y), y \in \Theta^{\mathbf{1}}\}$;
3: **end procedure**
4: **procedure** CONSTRUCTION:
5: at each step in line 9 of Algorithm 8, specify the set of nodes $\Theta_\triangle^{rb} = \Theta_\triangle$;
6: solve the reduced basis problem (1.22), compute $\mathscr{E}_r(y)$ and $s(y)$ at each $y \in \Theta_\triangle^{rb}$;
7: update $\Theta_\triangle^{rb}$ such that $\mathscr{E}_r(y) > \epsilon_t, \forall\, y \in \Theta_\triangle^{rb}$ (remove well approximated nodes);
8: **while** $\max_{y\in\Theta_\triangle^{rb}} \mathscr{E}_r(y) > \epsilon_t$ **do**
9: pick $y^{N+1} = \text{argmax}_{y\in\Theta_\triangle^{rb}} \mathscr{E}_r(y)$;
10: solve (46) at y^{N+1} and update $X_{N+1} = X_N \oplus \text{span}\{u(y^{N+1})\}$;
11: set $N = N+1$ and repeat steps in line 6 - line 7 with new X_N;
12: **end while**
13: **end procedure**

We remark that the adaptive greedy algorithm 9 for the construction of reduced basis space explores all the nodes in the construction of the dimension-adaptive hierarchical approximation in Algorithm 8 and the outputs of interest s are evaluated based on the surrogate (reduced basis) solution with inexpensive solve of the reduced basis problem in contrast to the full (high-fidelity) solution with expensive solve of the high-fidelity problem. Moreover, error estimates of the surrogate outputs of interest can be obtained based on the reduced basis approximation error $\mathscr{E}_r$ controlled by the error tolerance ϵ_t.

Note that when the number of terms Q_f and Q_a become large, the full online evaluation (2.9) will be expensive. Let us make two observations in order to further reduce the online evaluation cost: the first is that often $\Theta_q^a(y) = y_q$ with Q^a representing the dimension of a high-dimensional probability space for UQ problems; the other is that the nodes inside one set $\Theta_\triangle$ or from neighbor sets are only different from each other in limited dimensions, e.g., the node $(1, 0.5, 0.5, \ldots, 0.5)$ is a neighbor of the node $(0, 0.5, 0.5, \ldots, 0.5)$, which are only different in the first dimension. Based on these two observations,

we may identify the different terms in (2.9) from one node to the next in the adaptively constructed grid and only subtract these terms from $||\hat{e}(y)||_X^2$ at the previous node and add the corresponding new terms to it at the current node, resulting in $O(Q_f + NQ_a)$ operations in average for each evaluation. We remark that this computational reduction is still valid whenever there are only a few terms among $\Theta_q^a(y), 1 \le q \le Q_a$ different from one node to its neighbors.

Remarks on extension to more general PDE models

We presented the adaptive and weighted reduced basis method based on a coercive, steady and linear elliptic equation with affine input and compliant output. However, the method is not constrained by these elementary properties. In fact, it has been developed and extended to deal with many different PDE models, [180, 177, 93, 163, 11, 86, 48, 63], and applied in a variety of physical and engineering fields, [162, 51, 163, 117, 175, 41, 45]. Besides some specific extensions as introduced in section 4.3 of chapter 4 in the context of failure probability evaluation, we provide here a series of remarks for more general extensions with some associated references.

First of all, the coercivity property of the differential operator is used in computing a lower bound for the evaluation of a posteriori error bound (5.35). When the problem fails to be coercive, for instance in Stokes equations, where only an “inf-sup" compatibility condition is satisfied, we can replace the coercivity constant by an *inf-sup constant* and arrive at the same reliable and accurate a posteriori error bound [180, 177, 143]. Moreover, we may even introduce some “surrogates" error bounds based on more rough estimation of the inf-sup stability constant.

Secondly, for unsteady problems, e.g., a parabolic equation, the reduced bases should be explored not only at different samples but also at different time steps. In order to efficiently extract the most representative bases, we may employ proper orthogonal decomposition (POD) to project the solutions at different time steps into a small bases and use a greedy algorithm to choose the samples, leading to a *POD-greedy algorithm* [93, 163, 146] as presented in Algorithm 6 in chapter 4.

Thirdly, in order to deal with nonlinear problems, different approaches can be adopted. Taking Navier-Stokes equations for example, where the nonlinearity is quadratic on the state variable, we may employ Newton iteration to solve the reduced basis system as done for solving the high fidelity system [162]. Another approach is to use the *empirical interpolation for operators* [86, 63] in decomposing the nonlinear operators into linear combination of a series of linear operators.

Fourthly, when the random inputs are not given in affine structure, e.g., log-normal random field, we may reconstruct the nonaffine random field as random field with finite affine terms by *empirical interpolation method* [11, 86, 48]. This reconstruction is very efficient (resulting in a limited number of affine terms) for smooth functions and functions that enjoy the compressibility property, i.e., a function is compressible from a high-dimensional space to a low-dimensional space without losing too much accuracy.

Finally, for noncompliant problems where the output of interest is different from the right hand side of the equations, the error convergence for the approximation of the output depends only linearly on the error convergence for the approximation of the solution. Moreover, the norm of the functional for the output may not be easy to evaluation. In this case, we employ a *primal-dual approach* as in section 4.3.1 of chapter 4, where a dual problem is formulated by setting the output as a right hand side of the dual equations and the output is evaluated with contribution from both the primal and dual problem [178, 163]. The advantages of the primal-dual approach are that it avoids the computation of the functional norm and achieves quadratic error convergence. We will illustrate this approach by one numerical example in section 5.4.6.

5.4 Numerical experiments

This section is devoted to demonstrate the efficiency and accuracy of the adaptive and reduced computational framework and compare it to other methods (anisotropic sparse grid, ANOVA as introduced in section 5.2.4) for high dimension uncertainty quantification problems. We illustrate the computational performance of the proposed Algorithm 8 with verification in two dimensions, and compare it to the algorithm without verification and anisotropic sparse grid scheme (5.29) in section 5.4.1. In section 5.4.2 we illustrate why the ANOVA approach does not work well for functions with strong interaction and arbitrary probability measure, whereas this case can be efficiently dealt with by our proposed approach. In section 5.4.4, we show how the sparsity in high dimensions (from $O(10)$ to $O(1000)$), including different interaction and importance of different dimensions, can be efficiently and accurately captured by the proposed method. The last two sections 5.4.5 and 5.4.6 deal with heat diffusion and groundwater flow problems and demonstrate how the adaptive and weighted reduced basis method can be effectively applied to reduce the computational effort.

5.4.1 Hierarchical construction with verification

In this experiment, we compare the dimension-adaptive hierarchical interpolation Algorithm 8 with the same algorithm without the procedure of verification. The two dimensional function $s : [0,1]^2 \to \mathbb{R}$ is given by

$$s(y) = \cos(2\pi(y_1 - 0.3))\cos(2\pi(y_2 - 0.5)). \tag{5.37}$$

We run the interpolation Algorithm 8 in six different cases. The first three cases include hierarchical construction without verification based on piecewise linear polynomials with equidistant nodes and the weight in (5.28) are set as $w = 1, 0.5, 0$, corresponding to the purely dimension-adaptive grid construction, balanced construction and conservative sparse grid construction, respectively. The fourth case is specified with the same configuration as the first three except that the verification procedure is incorporated. The last two cases use Lagrange polynomials based on Clenshaw–Curtis nodes with verification and the weight $w = 1$ and $w = 0$, respectively. We set the maximum number of nodes adaptively as one larger than the number of nodes in the current grid, with the upper bound $M = 10^4$, and specify the interpolation error tolerance as $\varepsilon_t = 10^{-15}$. We compute the interpolation error as $\max_{y \in \Xi_{test}} |s(y) - \mathscr{I}_g s(y)|$ with the set of testing nodes given by $\Xi_{test} := \{y_1, y_2 = n/2^8, n = 0, \ldots, 2^8\}$, a fine regular grid with step size $1/2^8$. The final index sets S_m for the six different cases are plotted in Figure 5.5, where the active indices are marked with boxes (blue and red) and the index to be processed in the next step is marked with red box. Figure 5.6 reports the interpolation errors for all the six cases.

From the first figure (left-top of Figure 5.5), we can see that the enrichment of active indices has stagnated along y_2 by the purely dimension-adaptive scheme, resulting in large interpolation error (see left of Figure 5.6) since the function is not sufficiently well approximated in the second dimension. The balancing scheme with $w = 0.5$ (see middle-top of Figure 5.5) is able to construct fine grid in the second dimension but fails to capture the interaction of the two dimensions (due to stagnation), and thus still leads to large interpolation error as shown in Figure 5.6 (left). The Smolyak sparse grid construction introduced in section 5.2.2 does not run into the stagnation problem and achieves small interpolation error in this example (see right-top of Figure 5.5), but it can identify neither the important dimension nor the interaction. This drawback can be observed more clearly by comparison of the grid construction in the last two cases, where a full tensor grid is constructed by the adaptive scheme (see middle-bottom of Figure 5.5) and the sparse scheme that produces many more useless nodes in each single dimension (see right-bottom of Figure 5.5). Note that the last two cases result in higher approximation accuracy (see Figure 5.6) than the others because the globally supported Lagrange polynomial basis is more suitable to approximate smooth functions. By using the same locally supported piecewise linear basis as in the first three cases but incorporating the verification procedure, we can get rid of the stagnation problem and adaptively construct the grid with automatic

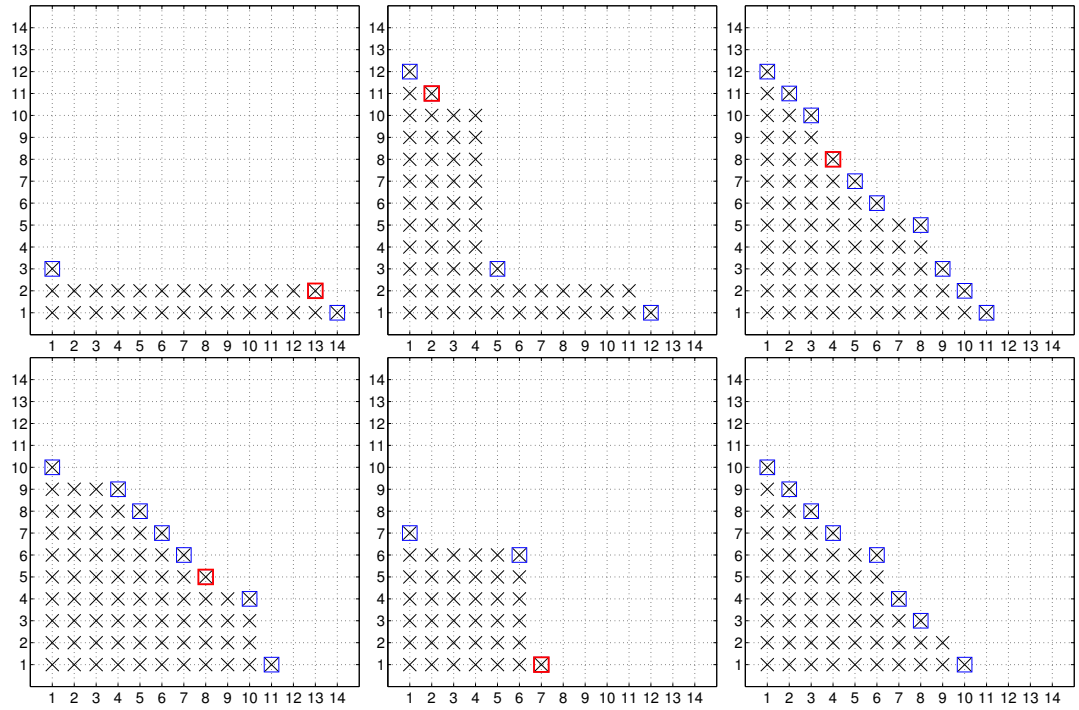

Figure 5.5: Illustration of dimension-adaptive hierarchical construction of the generalized sparse grid in different cases; top row: piecewise interpolation without verification with weight $w = 1$ (left), $w = 0.5$ (middle), and $w = 0$ (right); bottom row: piecewise interpolation with verification and weight $w = 1$ (left), Lagrange interpolation with verification and weight $w = 1$ (middle), and $w = 0$ (right).

identification of the importance and interaction of different dimensions, as shown in Figure 5.5 (left-bottom). From this experiment (two dimensional case for the sake of the illustration), we can see that the verification procedure works efficiently to get rid of the stagnation problem, which is to blame as one drawback of the dimension-adaptive hierarchical construction approach. We remark that the balancing scheme in (5.28) can not effectively avoid stagnation. Moreover, it is not computationally convenient to use since the weight parameter w is not known a priori and it depends on different problems under consideration.

In the second example, we test the efficiency of the verified dimension-adaptive algorithm for interpolation of these anisotropic functions

$$s_1(y) = \exp(y_1/5) + \exp(5y_2), \quad s_2(y) = \exp(y_1 y_2), \quad s_3(y) = \exp(y_1/5) + \exp(5y_2) + \exp(y_1 y_2). \quad (5.38)$$

We run the interpolation Algorithm 8 with the interpolation error tolerance set as $\varepsilon_t = 10^{-15}$. The constructed indices are displayed in Figure 5.7, from which we can see that the verified dimension-adaptive algorithm efficiently and accurately captured the interaction and importance of different variables of the test functions. The first one has no interaction term and y_2 plays a more important role (in terms of function value) than y_1. The second one features strong interaction and equal importance of the two dimensions. The last one has strong interaction and more important dimension y_2 than y_1. We remark that these properties can not be captured by the anisotropic sparse grid construction with weighted index set (5.29) as introduced in [148]. As a matter of fact, such approach either deteriorates efficiency because many useless indices are included or loses accuracy because the necessary indices (for strong interaction term) can not be captured, especially in high dimensions.

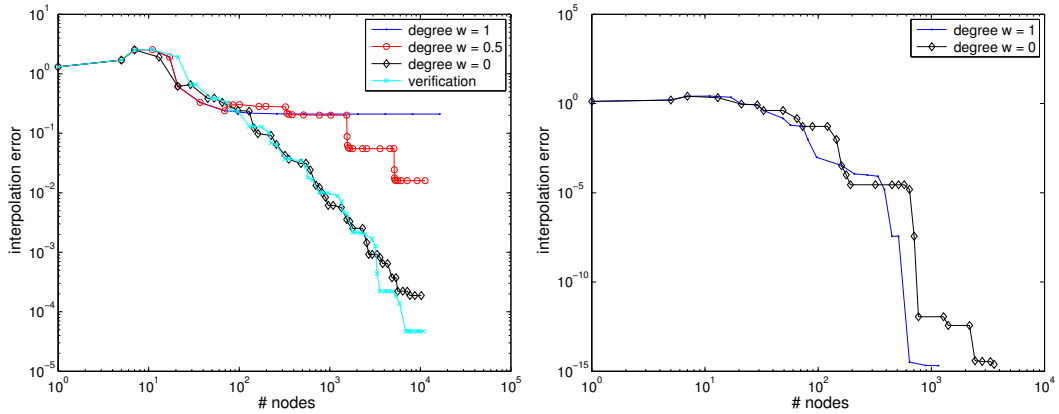

Figure 5.6: Interpolation error corresponding to the grid construction in Figure 5.5; left: picecwise interpolation in the first four cases; right: Lagrange interpolation in the last two cases.

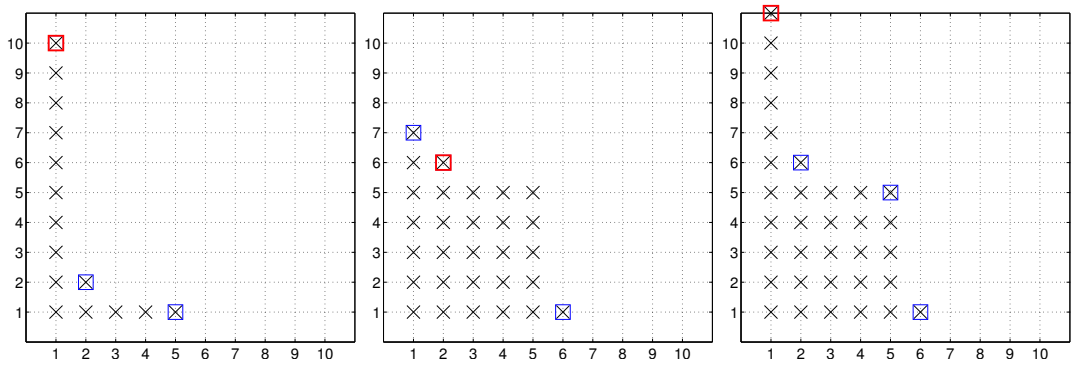

Figure 5.7: Illustration of dimension-adaptive hierarchical construction of generalized sparse grid for anisotropic interpolation; indices for s_1 (left), s_2 (middle), s_3 (right) with tolerance $\varepsilon_t = 10^{-15}$.

5.4.2 Sobol functions featuring strong interaction

In this numerical experiment, we study the functions with separated variables proposed by Sobol [192] to test the accuracy and efficiency of the hierarchical approximation in the extreme case - building minimal full tensor product grid, with comparison to the approximation based on anchored ANOVA (cut-HDMR) [73, 128]. The functions are defined as

$$s_1(y) = \prod_{k=1}^{K} \frac{|4y_k - 2| + p_k}{1 + p_k} \text{ and } s_2(y) = \prod_{k=1}^{K} \frac{1 + 3p_k y_k^2}{1 + p_k}, \tag{5.39}$$

where $y_k \in [0,1], 1 \le k \le K$ and the parameter $p_k, 1 \le k \le K$, is nonnegative for the first function and positive for the second one. Both functions have separated variables, meaning that the total integral (with value 1) can be computed by the product of individual integrals evaluated separately, but all of them are strongly interacting for pointwise evaluation of the function value. Since the first function has singularities ("peaks") at $y_k = 0.5, 1 \le k \le K$ and the second function is smooth, we use piecewise polynomial basis for the first function and global Lagrange polynomial basis for the second one. First of all, let us take a simple low-dimensional function s_2 with $K = 3$ and $p_k = 1, 1 \le k \le K$ and consider the anchored ANOVA approximation with several different anchor points $\bar{y}_k, 1 \le k \le K$

and expansion orders. Let $\bar{s}_i, 0 \le i \le K$ denote the approximated integral with expansion up to i dimensions (see the expansion formula (5.31)). When $i = 0$, the approximated integral is taken as the function value at the anchor point. The approximated integrals for different additive functions in the expansion are computed by tensor product Clenshaw–Curtis quadrature formula with 3 abscissas in each dimension. The results at different settings are reported in Table 5.1, from which we can observe that the approximation results are far from each other at different anchor points before the full expansion with $i = K$ is used. Moreover, the averaged approximations in the last column do not lead to a more accurate approximation as proposed in [90]. The approximations of the integral converge to the exact value with growing expansion order and reach the exact value only when the full expansion with $2^K = 8$ terms has been incorporated in all cases. These observations confirm the drawbacks of the anchored ANOVA approximation as pointed out in section 5.2.4. Similar results can be shown also for the first function s_1 and for higher dimensional integration problems by this approach. In fact, there is no gain in this case but more cost by the anchored ANOVA approximation since not only the last term has to be evaluated in all the K dimensions but also the other $2^K - 1$ terms of the expansion (5.31).

$\bar{y}_k$	0.0000	0.1667	0.3333	0.5000	0.6667	0.8333	1.0000	average
$\bar{s}_0$	0.1250	0.1589	0.2963	0.6699	1.5880	3.6641	8.0000	2.0717
$\bar{s}_1$	0.5000	0.5624	0.7407	0.9570	0.9074	-0.1981	-4.0000	-0.0758
$\bar{s}_2$	0.8750	0.9037	0.9630	0.9980	1.0046	1.1589	2.0000	1.1290
$\bar{s}_3$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 5.1: Approximated values of the integral of the function s_2 by anchored ANOVA expansion (5.31) with different anchor points (in row) and expansion orders (in column).

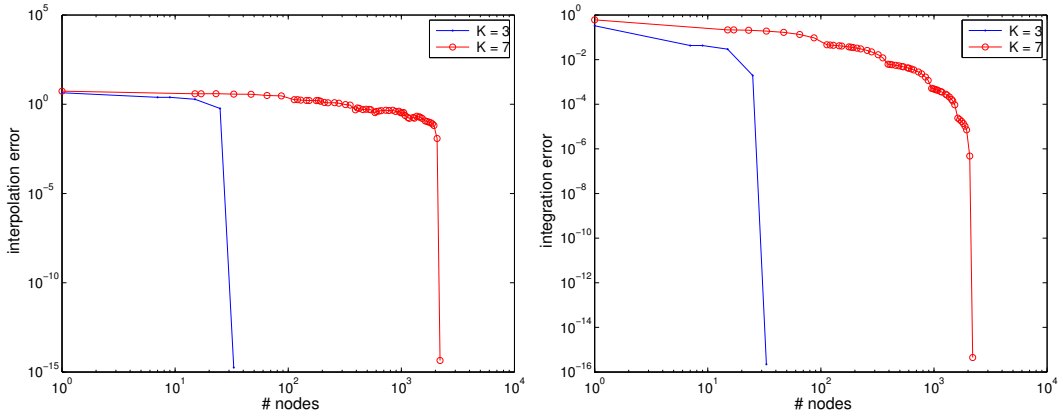

Figure 5.8: Interpolation error (left) and integration error (right) of dimension-adaptive hierarchical approximation of the smooth function s_2 with the dimension $K = 3$ and $K = 7$.

The interpolation and integration errors for the dimension-adaptive hierarchical approximation of the smooth function s_2 are displayed in Figure 5.8, where the interpolation error is defined by $\max_{y \in \Xi_{test}} |s(y) - \mathscr{I}_g s(y)|$ with the testing set Ξ_{test} consisting of 100 randomly selected samples. The decay of both interpolation and integration errors is very slow at the beginning, and fall to about the machine precision when the minimal full tensor product grid with 3 nodes in each dimension has been constructed, requiring in total $3^3 = 27$ and $3^7 = 2187$ nodes, respectively. This decay confirms again the necessity to use all the expansion terms by the anchored ANOVA approximation in order to have accurate integration. The dimension-adaptive hierarchical algorithm successfully detects the full tensor product grid structure and construct it automatically with the ultimate number of nodes 33 and 2201, slightly bigger than those of the full tensor product grid due to the verification procedure. As

for the approximation of the singular function s_1, we reduce the effect of the variation of y_k by setting a large parameter $p_k = 100, 1 \le k \le K$, which leads to the results in dimensions $K = 4$ and $K = 8$ in Figure 5.9. Similar convergence behaviour can be observed for the singular function as that for the smooth function, in particular 89 and 6577 nodes are constructed close to the minimal number of full tensor product grid $3^4 = 81$ and $3^8 = 6561$. Note that in this case, the approximation errors decay more uniformly due to the reduced variation.

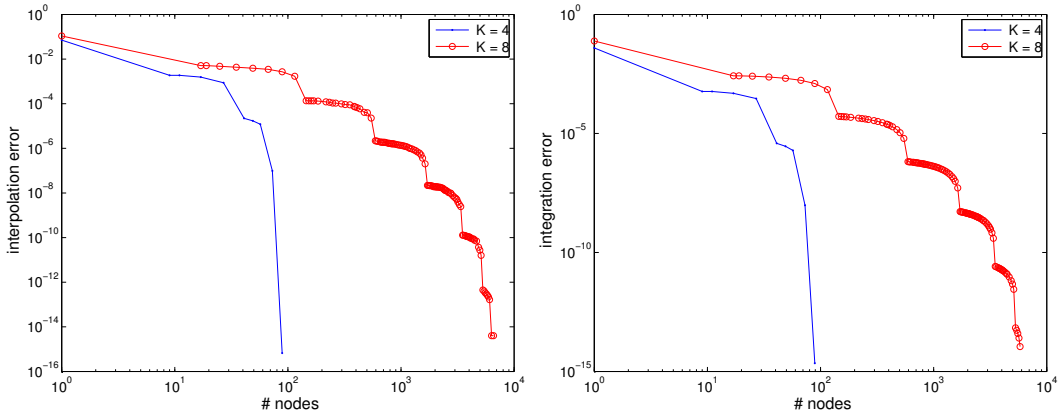

Figure 5.9: Interpolation error (left) and integration error (right) of dimension-adaptive hierarchical approximation the singular function s_1 with dimension $K = 4$ and $K = 8$.

5.4.3 Approximation with arbitrary probability measure

By this experiment, we study the dimension-adaptive hierarchical approximation with arbitrary probability measure in order to demonstrate the efficiency of using the interpolation and integration error indicators (5.25) and (5.11) for interpolation and integration problems, respectively, and illustrate why the variance-based ANOVA (or HDMR) expansion is not suitable for interpolation problems. We use the following exponential function

$$s(y) = \exp\left(-\sum_{k=1}^{K} c_k(y_k - 0.5)\right), \tag{5.40}$$

and set $c_k = 1, 1 \le k \le K$ with dimension $K = 5$. The random variables are set to obey beta distribution as $y_k \sim Beta(\beta k, \beta k), 1 \le k \le K$, being $\beta \in \mathbb{R}_+$ a scaling parameter. The probability density function (PDF) with different parameters is displayed in Figure 5.10, from which we can see that as the parameter becomes bigger, the more concentrated the PDF becomes and the smaller the variance is. Therefore, the importance of different dimensions becomes different as influenced by the given probability measure instead of the parameter c_k.

We run the dimension-adaptive hierarchical approximation Algorithm 8 to compute both the interpolation and the integration of the given function with different error indicators. We employ the nested Kronrod-Patterson quadrature nodes [159] associated with the beta measure at different parameter β. The interpolation error is computed as the maximum at 100 randomly selected samples and the integration error is computed by taking the approximation of the integral in the final step as the “exact" value. The error convergence of the approximation for interpolation and integration with different error indicators is shown for $\beta = 1, 5, 10, 20$ in Figure 5.11. From the left of Figure 5.11, we can see that the interpolation errors obtained with interpolation error indicator converge faster than those obtained with integration error indicators at different values of β. On the other hand, the convergence of the

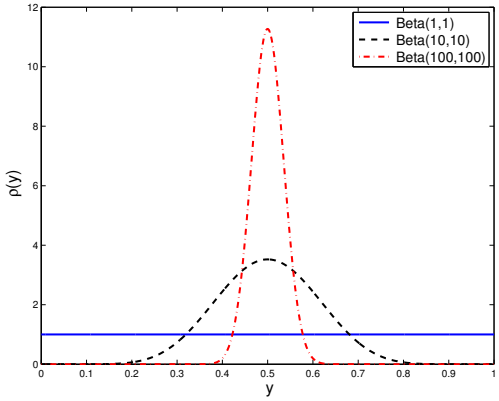

Figure 5.10: Probability density function of beta distributed random variable with different parameters.

integration errors shown on the right of Figure 5.11 highlights that the integration error indicator leads to evidently more accurate approximation of the integral than the interpolation error indicator for the cases $\beta = 5, 10, 20$. These observations confirm that the integration error indicator, closely related by the underlying probability measure to the variance-based ANOVA approximation, works efficiently for integration but may give rise to large errors for interpolation.

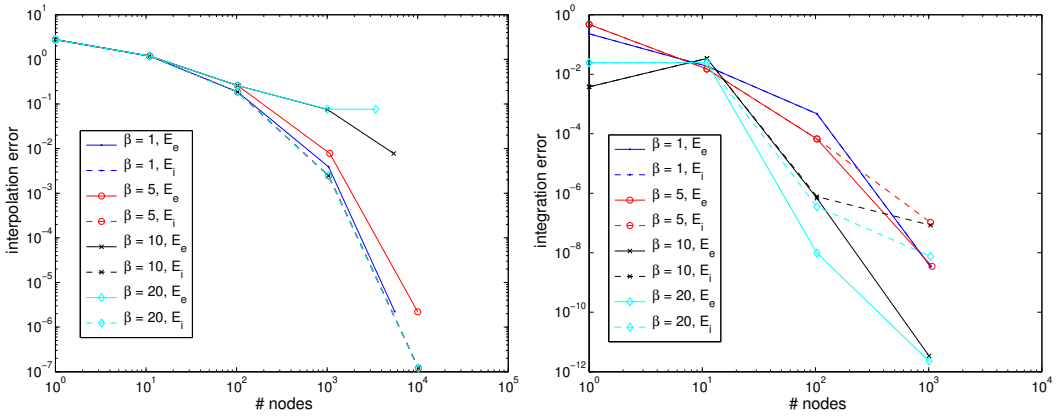

Figure 5.11: Interpolation error (left) and integration error (right) with different scaling parameter $\beta = 1, 5, 10, 20$ (with different markers) and error indicators, $\mathscr{E}_i$ in dashed line and $\mathscr{E}_e$ in solid line.

5.4.4 High-dimensional functions featuring sparsity

In this numerical experiment, we test the performance of the dimension-adaptive hierarchical approximation of high-dimensional functions featuring sparsity, i.e. low interaction or distinct importance of different dimensions. The first function has low interaction property, given by

$$s(y) = \sum_{k=1}^{K} y_k^2 - \sum_{k=1}^{K-1} y_k y_{k+1}, \tag{5.41}$$

which is a polynomial of total degree 2 and interaction level 2 (in the sense of ANOVA expansion (5.31)). We set the dimension as $K = 10$ and 100, and run the dimension-adaptive hierarchical approximation

algorithm with the interpolation error indicator (5.25) for both the interpolation and integration. The interpolation error is computed at 100 randomly selected samples, and the integration error is measured with respect to the value of the exact integral $K/3-(K-1)/4$. Since the function is smooth, we use Lagrange basis with Clenshaw–Curtis nodes. The results of the error indicator, the interpolation and integration errors, as well as the number of nodes are reported in Table 5.2, from which we can see that the second level isotropic sparse grid is sufficient to evaluate the integral accurately up to machine precision (with rounding error) in all the three cases, while for accurate interpolation the third level of sparse grid is needed and sufficient due to the interaction in the second term of s. The dimension-adaptive hierarchical construction algorithm is able to detect the isotropic structure of the sparsity and build automatically the isotropic sparse grid as can be seen from the comparison of the number of nodes in Table 5.2, where a small number of extra nodes are used for checking stopping criterion. We remark that in order to detect the full interaction relation of s by ANOVA expansion, the total number of terms to be explored is $1+K+(K-1)(K-2)/2$, which results in the same number of nodes by sparse grid or two dimensional tensor product grid (3 nodes in each dimension) for each term of the second level.

K	level	# nodes	error indicator $\mathscr{E}_i$	interpolation error	integration error
10	1	1 (1)	∞	1.9499	0.8333
	2	21 (21)	0.5000	0.6306	6.661e-16
	3	233 (221)	3.8858e-16	1.332e-15	6.661e-16
100	1	1 (1)	∞	11.5087	8.3333
	2	201 (201)	0.5000	1.9956	1.243e-14
	3	20213 (20201)	3.5527e-15	4.796e-15	1.066e-14

Table 5.2: Interpolation and Integration errors for dimension-adaptive hierarchical approximation of the low interacting function s_1; the number of nodes in $(\cdot)$ corresponds to an isotropic sparse grid.

We use (5.40) as the second test function, which features the sparsity due to distinct importance even with strong interaction of different dimensions. Here the parameter $c_k \in \mathbb{R}_+$ determines the importance of the dimension $k = 1,\dots,K$; $y_k \in [0,1], 1 \le k \le K$ are independent and uniformly distributed random variables. In the first example, we set $c_k = \alpha^{-k+1}, 1 \le k \le K$, with the scaling parameter $\alpha = 1.1$ and consider the dimension $K = 2^n, 3 \le n \le 6$. Clenshaw–Curtis quadrature is employed for the computation of the integral, where the "exact" value is taken as the approximation at the last step. We set the maximal number of nodes as $M = 10^m, 1 \le m \le 5$. The interpolation and integration error convergence is depicted in Figure 5.12 and the level of interpolation (note that we plot $i_k - 1$ in y axis due to implementation convenience) in the 64 dimensional case is reported in Figure 5.13. From these two figures we can conclude that only the first few dimensions dominate all the other dimensions and the dimension-adaptive hierarchical approximation Algorithm 8 successfully constructed the grid according to the importance of different dimensions. The convergence rate of the integration error for the 64 dimensional problem is around 1, which is faster than that of the Monte Carlo method (rate $= 1/2$) or quasi Monte Carlo method (rate $\in (1/2,1)$) [62].

In the second example, we test the dimension $K = 100, 400, 900, 1600$ and set the parameter $c_k, 1 \le k \le K$ as follows: we randomly select $\sqrt{K}$ dimensions and set c_k in these dimensions as $10^{-y_k^0}$, where $y_k^0 \in [0,1]$ is a sample drawn from uniform distribution, and in the other dimensions we set $c_k = 10^{-y_k^0-6}$. Therefore, the dimensions are divided into two scales. In each scale the importance of different dimensions is determined by a random variable $10^{-y_k^0}$. In another word, the important dimensions randomly distributes from 1 to K with total effective number of dimensions around $\sqrt{K}$. The convergence results for the interpolation and integration error is shown in Figure 5.14, from which we can see that the dimension-adaptive hierarchical approximation works efficiently for high-dimensional problems, with the integration error converging faster than Monte Carlo method with the total dimension as high as 1600. The right of Figure 5.13 demonstrates that both the scales (between

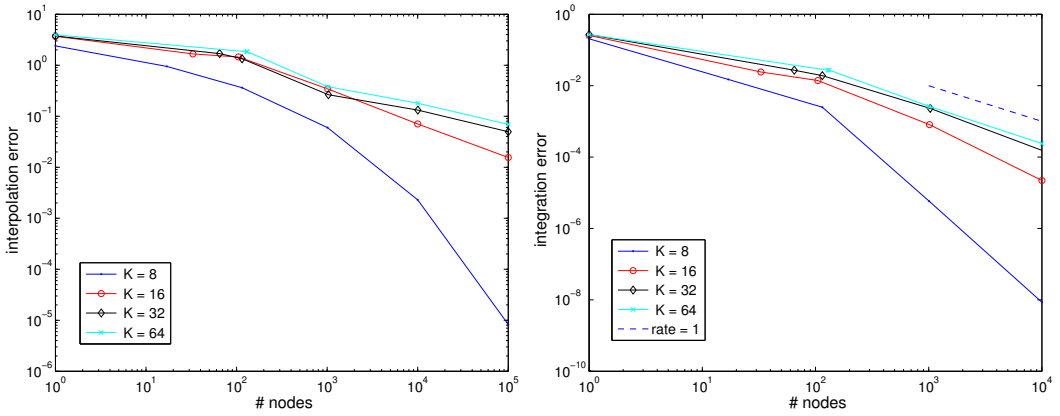

Figure 5.12: Convergence of interpolation error and integration error with dimension $K = 8, 16, 32, 64$.

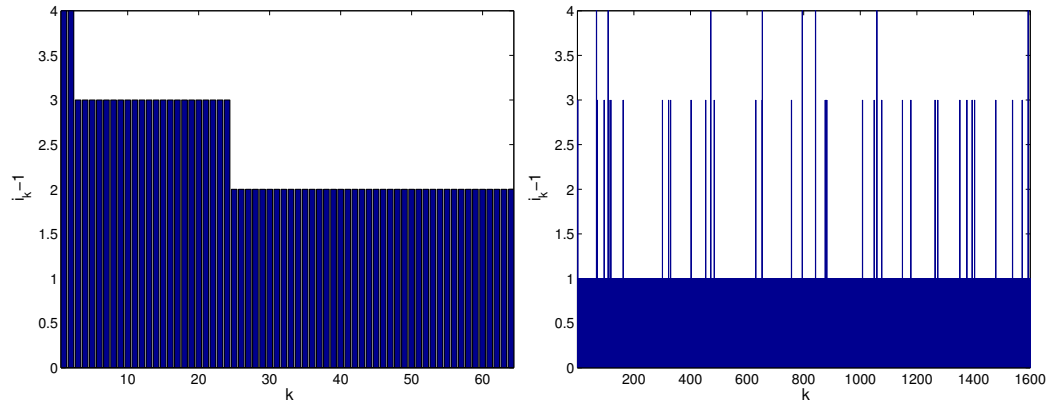

Figure 5.13: Grid level constructed by the dimension-adaptive hierarchical aproximation algorithm 8.

level $i_k = 1+1$ and levels $i_k = 1+3, 1+4$) and the importance in each scale (between level $i_k = 1+3$ and $i_k = 1+4$) of different dimensions are captured effectively by the dimension-adaptive algorithm 8. We remark that the examples in high-dimensional space feature distinct importance of different dimensions. In the case of equal importance of different dimensions in high-dimensional problems, the classical Monte Carlo method would achieve better computational performance.

5.4.5 Heat diffusion in thermal blocks

In this example, we study a heat diffusion problem (5.33) in thermal blocks with the thermal conductivity modeled by random variables. The problem is defined in the physical domain $D = (0,1)^2$ discretized with 101^2 nodes, which can be equally divided into K ($K = n^2, n \in \mathbb{N}_+$) blocks $D_k, 1 \le k \le K$. The thermal conductivity of each block is a random variable. In the first test, we demonstrate the efficiency of the weighted a posteriori error bound (5.36) in the case of arbitrary probability measure for integration problem. We consider the random coefficient a in (5.33) as

$$a(x,y) = \sum_{k=1}^{K} \chi_{D_k}(x) 10^{(y_k - 0.5)}, \tag{5.42}$$

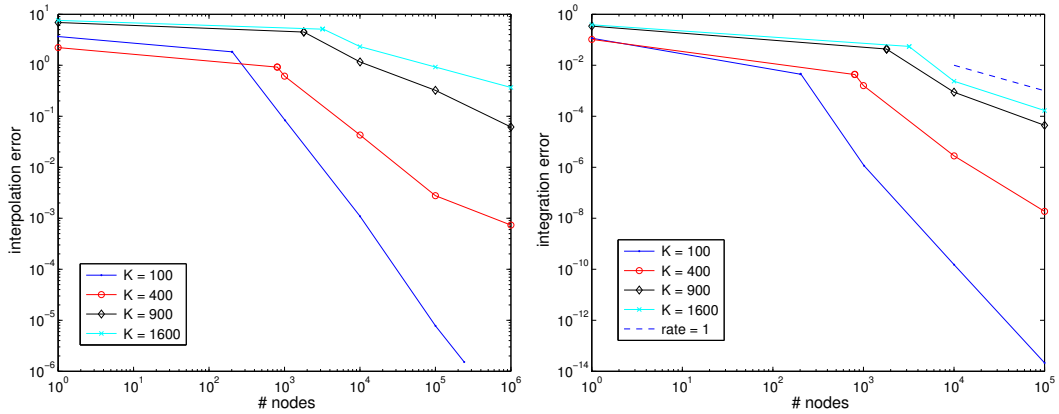

Figure 5.14: Convergence of interpolation error and integration error with dimension $K = 100, 400, 900, 1600$.

where χ_{D_k} is a characteristic function supported on the block D_k and $y_k \in [0,1], 1 \le k \le K$ with $K = 9$, are independent random variables obeying beta distribution $Beta(\beta,\beta)$ with $\beta = 5$, which feature almost equal importance in each of the 9 dimensions. A deterministic force term is considered as $f = 1$. We run the adaptive greedy Algorithm 9 with tolerance $\epsilon_t = 10^{-11}$ to construct the reduced basis space based on the hierarchical construction of the generalized sparse grid by Algorithm 8. For the construction of the generalized sparse grid, the integration error indicator (5.26) is used with the total number of nodes specified as $10^n, 0 \le n \le 4$ and the nested Kronrod-Patterson quadrature nodes are employed corresponding to the beta measure with different parameter β. The quantity of interest is the average temperature over the whole domain $\int_D u dx$, which is a compliant quantity. We apply both the a posteriori error bound (5.35) and the weighted a posteriori error bound (5.36) to construct the reduced basis space, resulting in 211 and 118 bases, respectively.

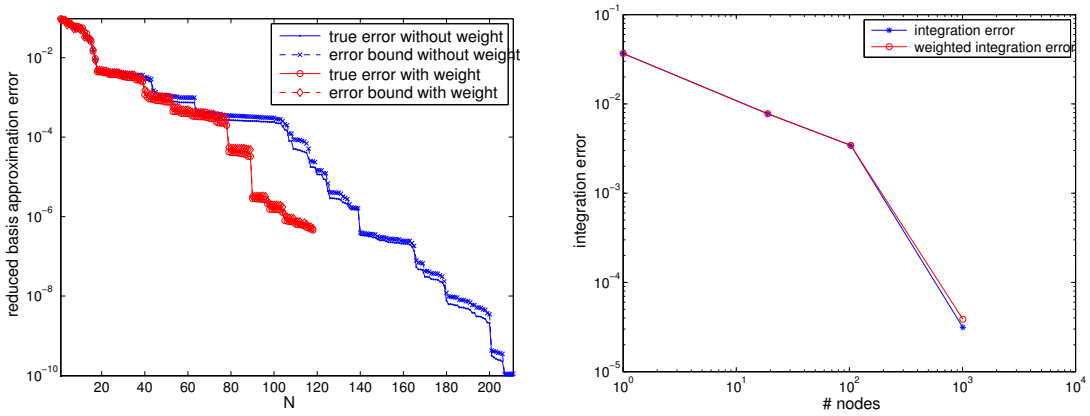

Figure 5.15: Left: true error and error bound of reduced basis approximation constructed by a posteriori error bound without (5.35) and with weight (5.36); right: (weighted) integration error.

The reduced basis approximation error (in the worse scenario case) tested with 100 randomly samples and the integration error (computed with the integral at 10^4 nodes as the reference value) of the two different cases are depicted in Figure 5.15. From the right of this figure we can see that the reduced basis approximation with the weighted a posteriori error bound (5.36) achieves almost the same accuracy for integration as that without the weight (5.35), even using much less bases (118 compared to 211). As

for the pointwise approximation, the weighted scheme results in faster convergence of the reduced basis approximation error than that without the weighted scheme, though does not guarantee the same small error at the end because it makes use of much less reduced bases, see in the left of Figure 5.15. Moreover, from the comparison of the true error and error bound plotted in the left figure, we confirm that the error bound is rather sharp, almost indistinguishable from the true error even if we use a constant $\alpha = 1$ for the lower bound in (5.35) and (5.36).

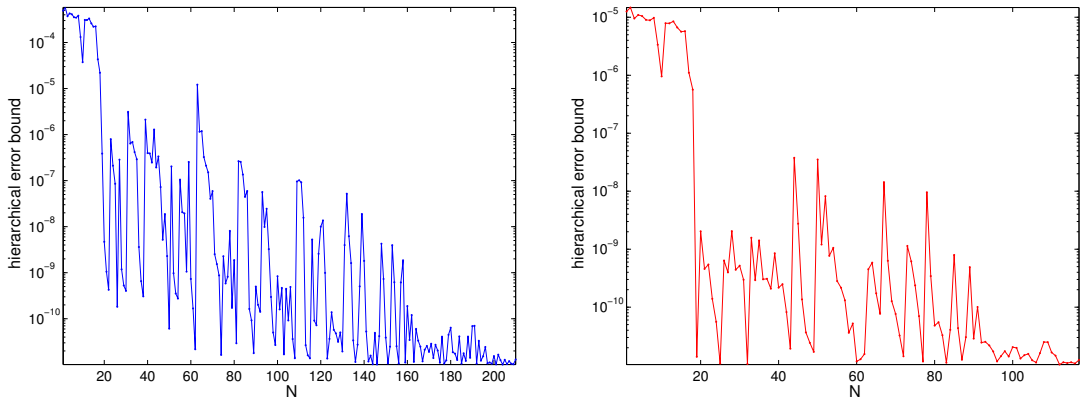

Figure 5.16: Left: the a posteriori error bound (5.35); right: the weighted a posteriori error bound (5.36) during the hierarchical construction of the generalized sparse grid by Algorithm 8.

Figure 5.16 reports the reduced basis error bound during the hierarchical construction process for both the weighted scheme and non weighted scheme. Large oscillation of the worst error bound evaluated at the nodes corresponding to the current active index can be observed for both cases. Both of them decrease to the prescribed tolerance $\epsilon_t = 10^{-11}$ but with different number of bases. In fact, the probability density ρ in (5.36) becomes very small when the node is far away from the center, thus gives rise to very small weighted a posteriori error bound and early stop of the algorithm with less bases. Moreover, this test also demonstrates that the total number of reduced bases is much smaller than the total number of constructed nodes, thus efficiently alleviate the entire computational cost.

In the second test, we consider a high-dimensional heat diffusion problem with 100 thermal blocks. The conductivity coefficient is

$$a(x,y) = \sum_{k=1}^{K} \chi_{D_k}(x) 10^{c_k(y_k - 0.5)} \tag{5.43}$$

where $y_k \in [0,1], 1 \le k \le K$ with $K = 100$, are independent and uniformly distributed random variables; $c_k, 1 \le k \le K$, are taken similarly to the second test of section 5.4.4 in separating the dimensions into two scales: we randomly select $2\sqrt{K}$ dimensions and set $c_k = 4y_k^0$ in these dimensions and $c_k = 10^{-4} \times 4y_k^0$ in the other dimensions, being $y_k^0 \in [0,1], 1 \le k \le K$, samples drawn from uniform distributed random variable. We set the error tolerance for the reduced basis space construction as $\epsilon_t = 10^{-8}$ in the greedy Algorithm 9 and the maximum number of nodes as $M = 10^n, 0 \le n \le 5$ for the hierarchical construction of the generalized sparse grid in Algorithm 8, which result in 161 bases in the reduced basis space. On the right, the integration error computed with different number of nodes are shown, which decays with a rate larger than 1, demonstrating that the dimension-adaptive hierarchical approximation converges much faster than Monte Carlo method for this high-dimensional uncertainty quantification problem.

Figure 5.17 displays both the reduced basis approximation error and the integration error. On the left, the true error and the error bound (in maximum norm) evaluated at 100 randomly selected samples at different number of reduced bases confirm the effectivity of the a posteriori error bound. The a posteriori error bounds at the selected reduced basis samples, most of which are chosen at

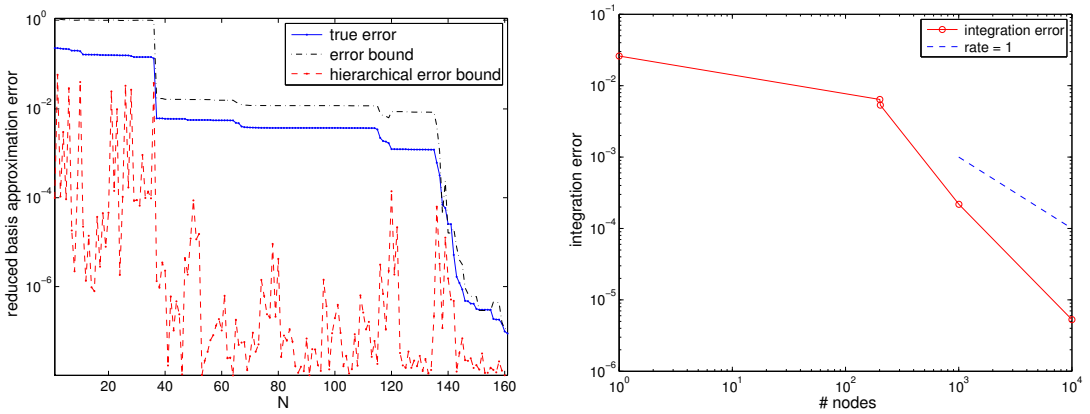

Figure 5.17: Left: true error and error bound of reduced basis approximation; right: integration error.

the beginning of the hierarchical construction process, decrease in an oscillating way to the error tolerance and remain smaller than the maximum error bounds at the 100 samples. Figure 5.18 depicts the effective dimensions and varied importance of different dimensions indicated by the prescribed parameters (on the left) and the level of the generalized sparse grid in different dimensions (on the right). From this figure, we can observe that all the dimensions in the effective scale represented by the characteristic function $\chi_d(k), 1 \leq k \leq K$, (on the left) are correctly identified with the grid level i_k equal or larger than 4 (on the right), and the dimensions in the ineffective scale are approximated mostly by the grid level $i_k = 1 + 1$. Moreover, the varied importance of different dimensions in each scale is also successfully identified as shown in Figure 5.18, where a larger value of y_k^0 leads to a relatively deeper grid level.

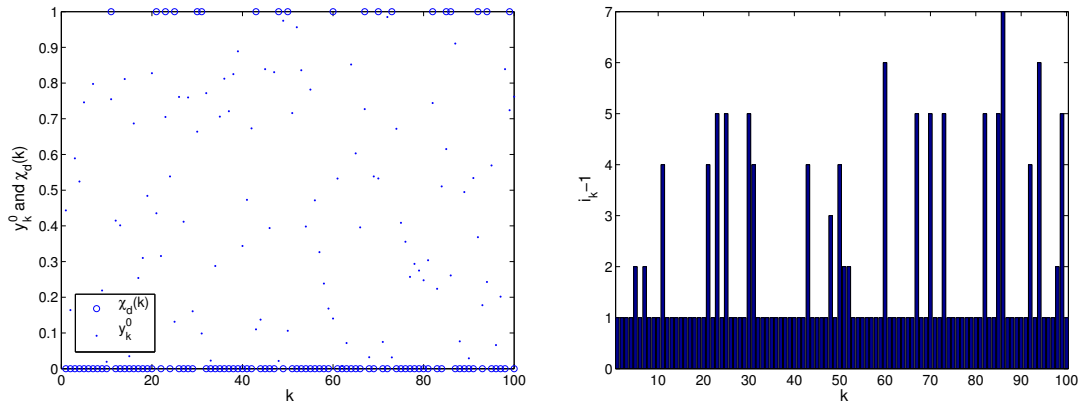

Figure 5.18: Left: true error and error bound of reduced basis approximation; right: integration error.

5.4.6 Groundwater flow through porous medium

This example is devoted to the study of groundwater flow through porous medium described by Darcy's law: find the pressure field $p \in D \times \Gamma$ such that the following equations hold

$$\begin{cases} -\nabla(a\nabla p) &= 0 \quad \text{in } D, \\ p &= 1 \quad \text{on } \partial D_4, \\ p &= 0 \quad \text{on } \partial D_2, \\ a\nabla p \cdot \mathbf{n} &= 0 \quad \text{on } \partial D_1 \cup \partial D_3, \end{cases} \tag{5.44}$$

where the physical domain is the two dimensional square $D = (0,1)^2$, as shown in Figure 5.19, with left and right boundaries ($\partial D_2 \cup \partial D_4$) prescribed of Dirichlet boundary conditions, and the upper and lower boundaries ($\partial D_1 \cup \partial D_3$) homogeneous Neuman boundary conditions. The permeability of the porous medium is given by the random field (with $x = (x_1, x_2)$)

$$a(x,y) = \mathbb{E}[a] + \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1 + \sum_{k=1}^{K} \sqrt{\lambda_k}\left(\sin(k\pi x_1)y_{2k} + \cos(k\pi x_1)y_{2k+1}\right), \tag{5.45}$$

which is a truncated Karhunen-Loève expansion of a Gauss covariance kernel $\exp(-(x_1 - x_1')^2/L^2)$ with correlation length L [149]. The eigenvalues $\lambda_k, 1 \le k \le K$, of this kernel decay exponentially as

$$\lambda_k = \sqrt{\pi}L\exp\left(-\frac{(k\pi L)^2}{4}\right), \tag{5.46}$$

and the random variables $y_k, 1 \le k \le 2K+1$, are assumed to be independent and obey uniform distribution taking values in $[-\sqrt{3}, \sqrt{3}]$ in order to guarantee that a is positive. The quantity of interest is

$$s(y) := L(p;y) = \int_{D_d} a(x,y)\partial_{x_1} p(x,y)dx, \tag{5.47}$$

where the disk region D_d has center $(0.75, 0.5)$ and radius 0.2, see Figure 5.19. This quantity is not compliant with the right hand side of equation $(5.44)_1$. Therefore, we adopt the primal-dual approach introduced in section 5.3. We first write the weak formulation of the Darcy equation (5.44) as: find $p \in H^1(D)$ such that

$$A(p,q;y) = 0 \quad \forall q \in H^1_{dir}(D), \tag{5.48}$$

where $H^1_{dir}(D) := \{q \in H^1(D) : q = 0 \text{ on } \partial D_2 \cup \partial D_4\}$ and the bilinear form A is given by

$$A = \sum_{k=0}^{2K+1} A_k(p,q)y_k, \tag{5.49}$$

being A_k defined corresponding to the terms in the expansion of the permeability coefficient a in (5.45) and $y_0 = 1$ for notational convenience. The dual problem associated with the primal problem (5.48) for the quantity of interest s is formulated as: find $\varphi \in H^1_{dir}$ such that

$$A(q,\varphi;y) = -L(q;y) \quad \forall q \in H^1_{dir}(D). \tag{5.50}$$

We construct reduced basis space $X^{pr}_{N_{pr}}$ with N_{pr} bases and $X^{du}_{N_{du}}$ with N_{du} bases to approximate the primal and dual weak problems (5.48) and (5.50) and define the residual of each problem as

$$R^{pr}(q;y) = -A(p_{N_{pr}}, q; y) \text{ and } R^{du}(q;y) = -L(q;y) - A(q, \varphi_{N_{du}}; y), \tag{5.51}$$

where $p_{N_{pr}}$ and $\varphi_{N_{du}}$ are the reduced basis approximations of the primal and dual solutions, respectively. We apply piecewise finite element basis to approximate these solutions in the physical space and denote the approximation space as $X \subset H^1_{dir}(D)$ and its dual as X'. After solving the primal and

dual reduced basis problems, we can approximate the quantity of interest s defined in (5.47) by

$$s_N(y) = L(p_{N_{pr}}(y); y) - R^{pr}(\varphi_{N_{du}}(y); y), \tag{5.52}$$

whose error can be bounded as (see details in [178])

$$|s(y) - s_N(y)| \le \Delta_N^s(y) := \frac{||R^{pr}(\cdot; y)||_{X'} ||R^{du}(\cdot; y)||_{X'}}{\alpha(y)}. \tag{5.53}$$

For the approximation in physical space, we use piecewise linear finite element basis on a regular triangular mesh with 17361 vertices, leading to a relatively large-scale algebraic system. We run Algorithm 8 for the dimension-adaptive hierarchical construction of the generalized space grid with integration error indicator (5.26) at a series of maximum number of nodes $10^n, 0 \le n \le 5$. The error tolerance for the reduced basis construction for approximating the non-compliant quantity of interest s is set as $\epsilon_t = 10^{-8}$. The a posteriori error bound (5.53) can be efficiently evaluated by an offline-online decomposition procedure for both the primal and dual problems with error tolerance $\epsilon_t = 10^{-4}$ for both problems. We set the correlation length $L = 1/16$ in (5.46) and K as $8, 16, 32, 64$, which lead to $17, 33, 65, 129$ dimensions taking $59\%, 89\%, 99\%$ and 100% percent of the total randomness measured by the L^∞-norm of the coefficient in (5.45). A set of typical solutions of the primal problem (5.48) and dual problem (5.50) at a randomly selected sample are depicted in Figure 5.19 (middle and right), where the dual solution, with evident bigger values near the disk D_d plays the role to correct the reduced basis approximation of the quantity of interest s_N by formula (5.52).

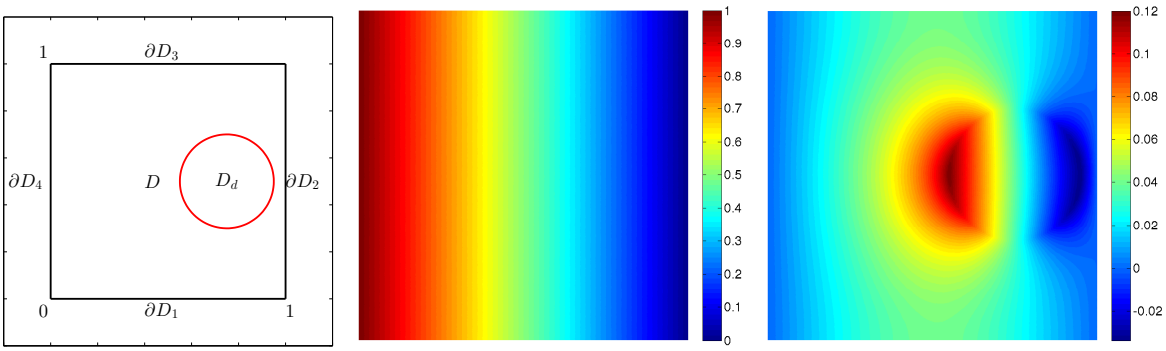

Figure 5.19: Left: physical domain and boundaries; middle and right: primal and dual solutions.

In the 17 dimensional case ($K = 8$), 37 primal bases and 39 dual bases are constructed. We test the convergence of the worst reduced basis approximation error with respect to the number of primal bases and dual bases computed with 100 randomly selected samples, which is displayed in Figure 5.20. From the left figure, we can observe that both the approximation error and the error bound decrease with growing number of primal and dual bases, leading to quadratically fast decrease with N_{pr} and N_{du} increasing simultaneously as shown in the right figure on the path $N_{pr} = N_{du}$. Moreover, the error bound shown in this figure is rather sharp (close to the real approximation error), demonstrating the efficiency of the primal-dual approach using the a posteriori error bound (5.53). The interpolation errors by the hierarchical interpolation formula (5.23) with 10^5 interpolation nodes are evaluated at the same test samples, where the worst approximation error is 4.0603×10^{-5}, much larger than that of the reduced basis approximation error 1.3333×10^{-10}. This large difference is due to fact that the interpolation approach adopts Lagrange basis to approximate the pointwise quantities blind to the underlying PDE model, while the reduced basis approach performs the pointwise evaluation by solving the underlying PDE model with cheap cost in the reduced framework. Therefore, we always use the reduced basis approximation to evaluate pointwise value of quantity of interest s.

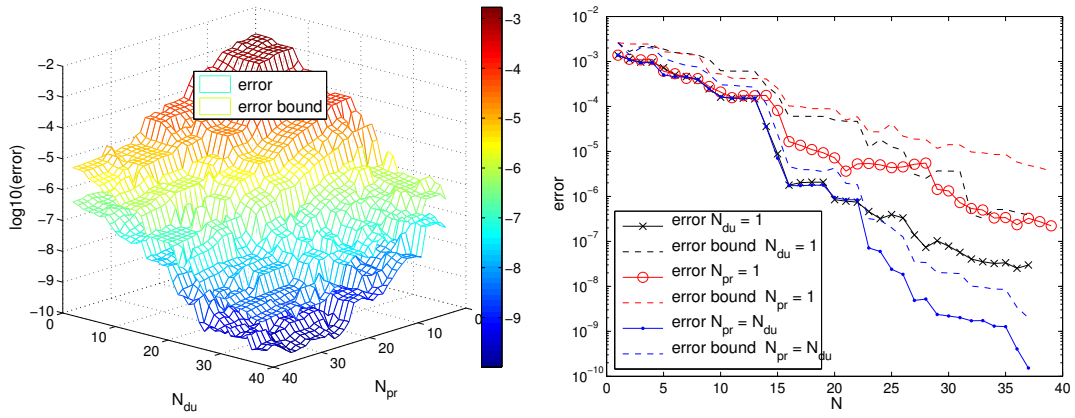

Figure 5.20: Left: reduced basis approximation error and error bound w.r.t. the number of primal bases N_{pr} and the number of dual bases N_{du}; right: three different settings of N_{pr} and N_{du}.$K = 8$.

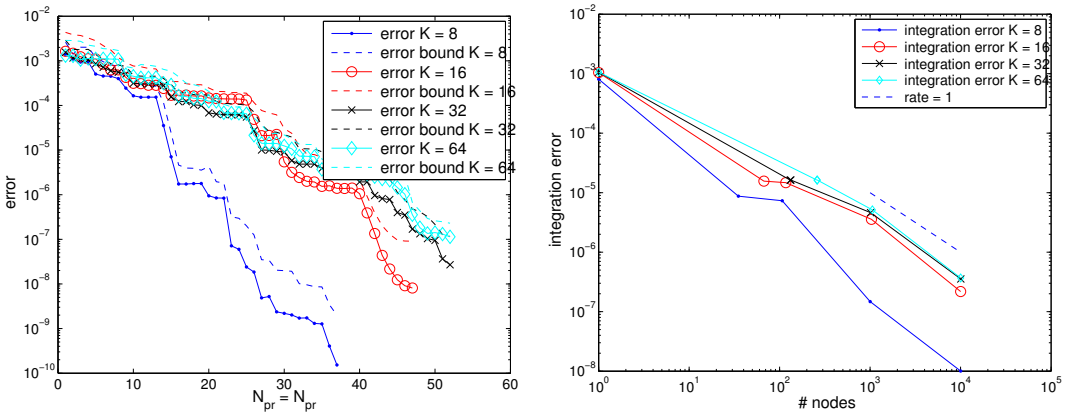

Figure 5.21: Left: reduced basis approximation error and error bound with $N_{pr} = N_{du}$; right: integration error. $K = 8, 16, 32, 64$, corresponding to 17, 33, 65, 129 dimensional problems.

The worst approximation error and integration error for the cases $K = 8, 16, 32, 64$, corresponding to $17, 33, 65, 129$ dimensional problems, are reported in Figure 5.21. The number of primal and dual bases increases with the dimension in order to achieve the same accuracy of approximation for pointwise evaluation and integration. However, the increase is rather small when the dimension becomes high because the important dimensions have been captured by the reduced basis approximation and dimension-adaptive hierarchical integration in the first few dimensions, and all the other dimensions play negligible role in contributing to the approximation error. It is worth to point out the remarkable fact that only a few tens (about 50) of reduced bases have been constructed to approximate the high dimension uncertainty quantification problems as shown in this example, thus requiring only a few tens of full solves of the underlying PDE model compared to a really large number (10^5 in this example) of full solves that would be needed without using the reduced basis method. Furthermore, as shown in the right of Figure 5.21 that the integration error converges with rate larger than 1, which demonstrates that the adaptive and reduced computational strategy for integration in high dimensions is very promising.

5.5 Summary

In this chapter we have developed an adaptive reduced computational framework for solving high-dimensional UQ problems. Two critical computational challenges were identified and illustrated for various UQ problems: curse of dimensionality and heavy computational burden. In order to tackle the first challenge, we adopted the approach for dimension adaptive tensor product integration and developed a verified algorithm based on generalized sparse grid construction to deal with one drawback of this approach – the stagnation phenomenon, and designed different error indicators suitable for integration and interpolation problems based on the hierarchical surpluses. To overcome the second challenge, we developed an adaptive and weighted reduced basis method, using an adaptive greedy algorithm in combination with the dimension adaptive hierarchical grid construction and a weighted a posteriori error bound to alleviate the computational cost in building the reduced basis space. The numerical experiments demonstrated that the verified dimension adaptive algorithm worked effectively in getting rid of the stagnation phenomenon and in automatically detecting the importance and interaction of different dimensions, which converged faster than the Monte Carlo and quasi Monte Carlo methods for high-dimensional integration problems with dimension as high as 1600. Moreover, the integration error indicator incorporating hierarchical surpluses, work contributions as well as quadrature weights was proved to be very efficient for UQ problems with arbitrary probability measures. As for pointwise evaluation of output of interest depending on PDE solution, the reduced basis approximation certified by the a posteriori error bound was demonstrated to be more accurate than the interpolation scheme based on Lagrange polynomials, one kind of dictionary bases without taking into account the underlying PDE models. Furthermore, only a few bases, a few hundreds (about 100 - 200) for heat diffusion in thermal blocks and a few tens (about 40 - 50) for groundwater flow through porous medium compared to 10^5 full solves, were constructed by the reduced basis method in order to achieve high accuracy for the high-dimensional approximation problems. This reduction will dramatically alleviate the prohibitive computational effort to the affordable level in solving large-scale PDE models (with large degrees of freedom) that consume considerable computational power.

Several further topics are worth to be investigated in applying the adaptive and reduced computational framework to solve high-dimensional UQ problems. The first is that low regularity points may exist in the high-dimensional space, for instance, the points featuring discontinuity or singularity. Therefore, efficient low regularity detection algorithms need to be incorporated in this framework, e.g., by checking the pointwise hierarchical surpluses instead of an averaged or maximum value at one index [128]. In addition to the detection algorithm, we remark that the reduced basis approximation may essentially get rid of the low regularity problems since it does not apply any family of dictionary bases but project the new solution into the reduced basis space spanned by solutions at some selected samples [41]. Another research topic is to develop more specific and goal-oriented model order reduction techniques in order to circumvent the “irreducible" PDE models, such as locally supported traveling waves, compressible flows that feature shocks, and so on. Moreover, when the effective dimensions become so high that no quadrature rule is feasible due to computational constraint, we have to turn to other approaches for computing the integration, such as Monte Carlo method, and detect when it is more suitable to apply these approaches than the proposed dimension adaptive quadrature rule. Since the reduced basis method is still applicable for Monte Carlo method, the adaptation may be carried out for sampling set with successive enrichment of new samples and elimination of well approximated samples, as shown in the previous chapter for risk analysis.

Part II
Analyses and Fast Solvers for Stochastic Optimal Control Problems

6 Stochastic elliptic optimal boundary control with random advection field

Design and optimization of physical and engineering systems can be formulated as optimal control problems. The latter usually aim at the determination of the forces or boundary conditions in a system of partial differential equations, through the minimization of suitable objective or cost functionals. Deterministic optimal control problems constrained by partial differential equations have been well developed and investigated for several decades (see, e.g., [123, 84, 200]), while the development of stochastic optimal control problem constrained by stochastic partial differential equations can still be considered to be in its infancy; see some very recent work, e.g. [99, 91, 172]. In [99], a stochastic optimal control problem constrained by a stochastic steady diffusion problem with deterministic distributed control function is introduced, and an error estimate for the Galerkin approximation of the optimality system in both physical space and stochastic space is provided. The work [91] deals with deterministic Neumann boundary control with error estimate for the same numerical approximation based on stochastic steady diffusion problem. The existence of a local optimal solution has also been demonstrated. However, the global existence as well as uniqueness of the optimal solution remain to be investigated. In [172], numerical experiments are conducted with 'pure' stochastic control function as well as 'semi' stochastic control function for an optimal control problem constrained by a stochastic steady diffusion problem.

Robin boundary conditions are a weighted combination of Dirichlet and Neumann boundary conditions, which are very versatile and useful in mathematical modelling [165, 161]. In this chapter, a stochastic Robin optimal control problem constrained by an advection-diffusion-reaction equation with advection-dominated term is studied. In order to analyze the existence and uniqueness of the optimal solution as well as the convergence of numerical approximation, saddle point formulation for linear-quadratic type of optimal control problem in the deterministic case has been developed and fully analyzed in [23, 28] and more recently in a deterministic reduced order modelling setting [144]. We take advantage of this formulation in the stochastic Robin optimal control problem to study the theoretical properties of the optimal solution and the numerical properties of approximation in both physical and stochastic spaces. We first derive a stochastic saddle point system [28, 23] and prove that it is equivalent to the first order optimality system for the stochastic Robin boundary control problem. The global existence and uniqueness of the optimal solution is obtained by Brezzi's theorem [27] for the saddle point formulation. Moreover, the optimal solution of the stochastic saddle point system is proved to depend regularly on the random variables. Thanks to this regularity, we are able to use stochastic collocation approximation [8] for the discretization of random variables and obtain an a

Reference for this chapter:

P. Chen, A. Quarteroni and Gianluigi Rozza. Stochastic optimal Robin boundary control problems of advection-dominated elliptic equations. SIAM Journal on Numerical Analysis, 51(5):2700–2722, 2013

priori error estimate of the numerical approximation. As for the discretization of the physical domain, we apply stabilized finite element approximation [165, 97, 94] and provide a priori error estimate. Based on these two approximations, a global error estimate for their combination is derived. Finally, we verify the correctness of our theoretical error estimates by numerical experiments in both low (of order $O(1)$) and high (of order $O(100)$) stochastic dimensions.

This chapter is organized as follows. In section 6.1, the stochastic Robin boundary control problem constrained by a stochastic advection dominated elliptic equation is introduced. We derive the stochastic saddle point system and prove it to be equivalent to the optimality system. In the following section 6.2, the stochastic regularity of the solution is obtained by recursively applying Brezzi's theorem for the saddle point system. Section 6.3 is attributed to the stabilized finite element approximation in physical space and stochastic collocation approximation in stochastic space as well as the error estimates of these approximations, followed by section 6.4 with numerical experiments of the approximation. Some summary remarks are given in the last section 6.5.

6.1 Stochastic Robin boundary control problem

6.1.1 Problem definition

Our stochastic Robin boundary control consists in finding a stochastic Robin boundary condition $g \in \mathscr{L}^2(\partial D)$ (the control function) in order to minimize the quadratic cost functional

$$\mathscr{J}(u,g) := \frac{1}{2}||u-u_d||^2_{\mathscr{L}^2(D)} + \frac{\alpha}{2}||g||^2_{\mathscr{L}^2(\partial D)} \tag{6.1}$$

constrained by the stochastic elliptic problem featuring a stochastic advection-dominated term

$$\begin{cases} -\nabla\cdot(a(x)\nabla u(x,\omega)) + \boldsymbol{b}(x,\omega)\cdot\nabla u(x,\omega) + c(x)u(x,\omega) = f(x) \text{ in } D\times\Omega, \\ a(x)\nabla u(x,\omega)\cdot\boldsymbol{n} + k(x)u(x,\omega) = g(x,\omega) \text{ on } \partial D\times\Omega, \end{cases} \tag{6.2}$$

where $u_d \in \mathscr{L}^2(D)$ is the observation, $\alpha > 0$ is a regularization coefficient, $a, \boldsymbol{b}, c$ are the diffusion, advection, and reaction coefficients, respectively, f is a force term, k is Robin coefficient, and $\boldsymbol{n}$ is the unit outward normal direction along the boundary. We make the following assumptions for $a, \boldsymbol{b}, c, k$.

Assumption 6.1 *The uncertainty is presented on the advection-dominated term through the random coefficient $\boldsymbol{b}: D\times\Omega\to\mathbb{R}^d$, which satisfies $\boldsymbol{b}\in\left(\mathscr{L}^\infty(\bar{D})\right)^d$, $\nabla\cdot\boldsymbol{b}(x,\omega)\in\mathscr{L}^\infty(D)$ and can be written as a linear function of finite random variables by, e.g., truncation of the Karhunen-Loève expansion [189] as given by (34) in the preliminary chapter*

$$\boldsymbol{b}(x,\omega) = \boldsymbol{b}_0(x) + \sum_{n=1}^{N} \boldsymbol{b}_n(x)\, y_n(\omega), \tag{6.3}$$

where $y_n:\Omega\to\Gamma_n$, $n=1,\ldots,N$ are uncorrelated bounded real-valued random variables with zero mean and unit variance.

Assumption 6.2 *There exist positive constants $0<r<R<\infty$ such that the diffusion coefficient satisfies*

$$r < a(x) < R \quad \text{a.e. in } \bar{D}. \tag{6.4}$$

As is customary, a.e. stands for almost everywhere, meaning everywhere except for a possible set with zero measure, and $\bar{D} = D\cup\partial D$. Moreover, we assume that $c\in L^\infty(\bar{D})$, $f\in L^2(D)$, and $k\in L^2(\partial D)$ as well

as the relations

$$-\frac{1}{2}\nabla\cdot\boldsymbol{b}(x,\omega)+c(x)\geq r'>0 \quad a.e.\ \ in\ D\times\Omega\ with\ r'<r \tag{6.5}$$

and

$$k(x)+\frac{1}{2}\boldsymbol{b}(x,\omega)\cdot\boldsymbol{n}(x)\geq 0 \quad a.e.\ \ on\ \partial D\times\Omega. \tag{6.6}$$

Let us introduce the bilinear form $\mathscr{B}(\cdot,\cdot):\mathscr{H}^1(D)\times\mathscr{H}^1(D)\to\mathbb{R}$, defined as

$$\begin{aligned}\mathscr{B}(u,v)&:=(a\nabla u,\nabla v)+(\boldsymbol{b}\cdot\nabla u,v)+(cu,v)+(ku,v)_{\partial D}\\&\equiv\int_\Omega\int_D a\nabla u\cdot\nabla v dxdP+\int_\Omega\int_D(\boldsymbol{b}\cdot\nabla u)vdxdP\\&\quad+\int_\Omega\int_D cuvdxdP+\int_\Omega\int_{\partial D}kuvd\gamma dP,\end{aligned} \tag{6.7}$$

and the linear functional $\mathscr{F}(\cdot):\mathscr{H}^1(D)\to\mathbb{R}$, defined as

$$\mathscr{F}(v):=(f,v)+(g,v)_{\partial D}\equiv\int_\Omega\int_D fvdxdP+\int_\Omega\int_{\partial D}gvd\gamma dP, \tag{6.8}$$

and then the weak formulation of problem (6.2) can be written as: find $u\in\mathscr{H}^1(D)$ such that

$$\mathscr{B}(u,v)=\mathscr{F}(v)\quad\forall v\in\mathscr{H}^1(D). \tag{6.9}$$

Theorem 6.1.1 *Provided that all the data satisfy Assumptions 1 and 2, we have that there exists a unique solution $u\in\mathscr{H}^1(D)$ to problem (6.2), and for a suitable constant C, it holds that*

$$||u||_{\mathscr{H}^1(D)}\leq C\Big(||f||_{L^2(D)}+||g||_{\mathscr{L}^2(\partial D)}\Big). \tag{6.10}$$

The proof follows the same lines as in the deterministic case [165] and is omitted here for simplicity.

6.1.2 Stochastic saddle point formulation

We apply Lagrangian approach for the derivation of an optimality system to solve optimal control problem (6.1) subject to the constraint (6.9). The Lagrangian functional is defined as [200]

$$\mathscr{L}(u,g,p)=\mathscr{J}(u,g)+\mathscr{B}(u,p)-\mathscr{F}(p), \tag{6.11}$$

where p is the Lagrangian multiplier or adjoint variable in $\mathscr{H}^1(D)$.

Lemma 6.1.2 *The first order necessary optimality conditions of the Robin boundary control problem are equivalent to the following stochastic optimality system: to find $u\in\mathscr{H}^1(D)$, $p\in\mathscr{H}^1(D)$, $g\in\mathscr{L}^2(\partial D)$, such that (s.t.)*

$$\begin{cases}\mathscr{B}(u,\tilde{u})=\mathscr{F}(\tilde{u})\quad\forall\tilde{u}\in\mathscr{H}^1(D),\\ \mathscr{B}'(p,\tilde{p})=(u_d-u,\tilde{p})\quad\forall\tilde{p}\in\mathscr{H}^1(D),\\ \alpha(g,\tilde{g})_{\partial D}=(p,\tilde{g})_{\partial D}\quad\forall\tilde{g}\in\mathscr{L}^2(\partial D),\end{cases} \tag{6.12}$$

where $\mathscr{B}'(p,\tilde{p})=\mathscr{B}(\tilde{p},p)$ is the adjoint bilinear form.

The stochastic optimality system (6.12) is obtained by taking the Gâteaux or directional derivative of the Lagrangian functional with respect to the variables p, u, and g, respectively, and setting them to be zero, which employs the same procedure as in the deterministic case; see [114], for instance.

This optimality system has also been studied in [99, 91, 172], with only local existence of the optimal solution obtained. In the following, we derive a stochastic saddle point formulation of the optimal control problem (6.1) and demonstrate the global existence and uniqueness of the optimal solution.

First of all, let us introduce new variables $\underline{u} = (u, g) \in \mathscr{U}$ and $\underline{v} = (v, h) \in \mathscr{U}$, where the stochastic tensor product space $\mathscr{U} = \mathscr{H}^1(D) \times \mathscr{L}^2(\partial D)$ is equipped with the graph norm $||\underline{u}||_{\mathscr{U}} = ||u||_{\mathscr{H}^1(D)} + ||g||_{\mathscr{L}^2(\partial D)}$. Define a bilinear form $\mathscr{A}(\cdot,\cdot) : \mathscr{U} \times \mathscr{U} \to \mathbb{R}$

$$\mathscr{A}(\underline{u}, \underline{v}) := (u, v) + \alpha(g, h)_{\partial D} \quad \forall \underline{u}, \underline{v} \in \mathscr{U}, \tag{6.13}$$

which is related to the cost functional (6.1) as follows,

$$\mathscr{J}(u, g) = \frac{1}{2}\mathscr{A}(\underline{u}, \underline{u}) - (u_d, u) + \frac{1}{2}(u_d, u_d). \tag{6.14}$$

Write $\underline{u}_d = (u_d, 0) \in \mathscr{U}$ as the new observation variable, and we have the equivalence $(\underline{u}_d, \underline{u}) = (u_d, u)$, so that minimizing the cost functional (6.1) is not different, up to a constant $(u_d, u_d)/2$, than minimizing the following cost functional (still denoted by $\mathscr{J}$)

$$\mathscr{J}(\underline{u}) := \frac{1}{2}\mathscr{A}(\underline{u}, \underline{u}) - (\underline{u}_d, \underline{u}). \tag{6.15}$$

Furthermore, introduce the affine form by slight abuse of notation $\mathscr{B}(\cdot,\cdot) : \mathscr{U} \times \mathscr{H}^1(D) \to \mathbb{R}$

$$\mathscr{B}(\underline{u}, q) := \mathscr{B}(u, q) - (g, q)_{\partial D}, \quad \forall \underline{u} \in \mathscr{U} \forall q \in \mathscr{H}^1(D). \tag{6.16}$$

By this new definition, we have the following minimization problem equivalent to the original one of minimizing the cost functional (6.1) subject to the stochastic constraint (6.9), which is

$$\begin{cases} \min\limits_{\underline{u} \in \mathscr{U}_{ad}} \mathscr{J}(\underline{u}) = \dfrac{1}{2}\mathscr{A}(\underline{u}, \underline{u}) - (\underline{u}_d, \underline{u}), \\ \text{s.t. } \mathscr{B}(\underline{u}, q) = (f, q) \quad \forall q \in \mathscr{H}^1(D). \end{cases} \tag{6.17}$$

Moreover, the equivalence between minimization problem (6.17) and the saddle point problem: to find $(\underline{u}, p) \in \mathscr{U} \times \mathscr{H}^1(D)$ such that

$$\begin{cases} \mathscr{A}(\underline{u}, \underline{v}) + \mathscr{B}(\underline{v}, p) = (\underline{u}_d, \underline{v}) \quad \forall \underline{v} \in \mathscr{U}, \\ \mathscr{B}(\underline{u}, q) = (f, q) \quad \forall q \in \mathscr{H}^1(D), \end{cases} \tag{6.18}$$

is established by the following proposition extended from deterministic case to the stochastic case

Proposition 6.1.3 *(see [23, 28]) Assume that the bilinear form $\mathscr{A}$ is symmetric, continuous, nonnegative, and strongly coercive on the kernel space $\mathscr{U}_0 := \{\exists\, \underline{u} \in \mathscr{U} : \mathscr{B}(\underline{u}, q) = 0 \quad \forall\, q \in \mathscr{H}^1(D)\}$. Assume also that the bilinear form $\mathscr{B}$ is continuous and satisfies the compatibility condition (inf-sup condition) (6.22). Then the minimization problem (6.17) and the saddle point formulation (6.18) are equivalent.*

From the above proposition, we immediately obtain the following lemma.

Lemma 6.1.4 *The minimization problem (6.1) subject to the stochastic problem (6.9) is equivalent to the saddle point problem (6.18).*

Proof We only need to verify the assumptions in Proposition 6.1.3. By definition (6.13), we have $\mathscr{A}(\underline{u}, \underline{v}) = \mathscr{A}(\underline{v}, \underline{u})$ and $\mathscr{A}(\underline{u}, \underline{u}) \geq 0$ so that $\mathscr{A}$ is symmetric and nonnegative. The continuity of $\mathscr{A}$ on

$\mathscr{U} \times \mathscr{U}$ is evident with the following estimate:

$$\mathscr{A}(\underline{u}, \underline{v}) \leq ||u||_{\mathscr{H}^1(D)} ||v||_{\mathscr{H}^1(D)} + \alpha ||g||_{\mathscr{L}^2(\partial D)} ||h||_{\mathscr{L}^2(\partial D)} \leq C_\alpha ||\underline{u}||_{\mathscr{U}} ||\underline{v}||_{\mathscr{U}}, \tag{6.19}$$

where C_α is a constant depending on α. On $\mathscr{U}_0$, we have $\mathscr{B}(\underline{u}, q) = 0$ so that $\mathscr{B}(u, q) = (g, q)_{\partial D} \, \forall \, q \in \mathscr{H}^1(D)$. Hence, it holds by the Lax–Milgram theorem and trace theorem [165] that $||u||_{\mathscr{H}^1(D)} \leq R'/r' ||g||_{\mathscr{L}^2(\partial D)}$, where R' is a positive constant. With this estimate, the coercivity of $\mathscr{A}$ follows:

$$\begin{aligned} \mathscr{A}(\underline{u}, \underline{u}) &= ||u||^2_{\mathscr{L}^2(D)} + \alpha ||g||^2_{\mathscr{L}^2(\partial D)} \\ &\geq \frac{\alpha r'^2}{2R'^2} ||u||^2_{\mathscr{H}^1(D)} + \frac{\alpha}{2} ||g||^2_{\mathscr{L}^2(\partial D)} \geq \frac{\alpha r'^2}{2R'^2} ||\underline{u}||^2_{\mathscr{U}}. \end{aligned} \tag{6.20}$$

As for the continuity of the bilinear form $\mathscr{B}$ on $\mathscr{U} \times \mathscr{H}^1(D)$, we have by definition (6.16) that

$$\begin{aligned} \mathscr{B}(\underline{u}, q) &\leq R' ||u||_{\mathscr{H}^1(D)} ||q||_{\mathscr{H}^1(D)} + ||g||_{\mathscr{L}^2(\partial D)} ||q||_{\mathscr{L}^2(\partial D)} \\ &\leq \max(R', 1) ||\underline{u}||_{\mathscr{U}} ||q||_{\mathscr{H}^1(D)}. \end{aligned} \tag{6.21}$$

The compatibility (inf-sup) condition of $\mathscr{B}$ on $\mathscr{U} \times \mathscr{H}^1(D)$ is shown by the following estimate

$$\begin{aligned} \sup_{0 \neq \underline{v} \in \mathscr{U}} \frac{\mathscr{B}(\underline{v}, q)}{||\underline{v}||_{\mathscr{U}}} &= \sup_{0 \neq (v,h) \in \mathscr{U}} \frac{\mathscr{B}(v, q) - (h, q)_{\partial D}}{||v||_{\mathscr{H}^1(D)} + ||h||_{\mathscr{L}^2(\partial D)}} \\ &\geq \sup_{(v,0) \in \mathscr{U}} \frac{\mathscr{B}(v, q)}{||v||_{\mathscr{H}^1(D)}} \geq r' ||q||_{\mathscr{H}^1(D)}. \end{aligned} \tag{6.22}$$

□

Thanks to the equivalence between the original minimization problem and the saddle point formulation established in Lemma 6.1.4, we can also obtain the global existence of a unique solution to the minimization problem, according to the following Brezzi's theorem for saddle point problem (6.18). (For the proof, see e.g., [27] or [165].)

Theorem 6.1.5 *(Brezzi) Provided that the assumptions in Lemma 6.1.4 are satisfied, the saddle point problem (6.18) admits a unique solution* $(\underline{u}, p) \in \mathscr{U} \times \mathscr{H}^1(D)$ *or* $(u, g, p) \in \mathscr{H}^1(D) \times \mathscr{L}^2(\partial D) \times \mathscr{H}^1(D)$. *Furthermore, we have the following estimate:*

$$\begin{aligned} ||\underline{u}||_{\mathscr{U}} &\leq \alpha_1 ||u_d||_{\mathscr{L}^2(D)} + \beta_1 ||f||_{L^2(D)}, \\ ||p||_{\mathscr{H}^1(D)} &\leq \alpha_2 ||u_d||_{\mathscr{L}^2(D)} + \beta_2 ||f||_{L^2(D)}, \end{aligned} \tag{6.23}$$

where

$$\alpha_1 = \frac{2R'^2}{\alpha r'^2}, \quad \beta_1 = \frac{\alpha r'^2 + 2R'^2}{\alpha r'^3}, \quad \alpha_2 = \frac{2R'^2 + \alpha r'^2}{\alpha r'^3}, \quad \beta_2 = \frac{\alpha r'^2 + 2R'^2}{\alpha r'^4}. \tag{6.24}$$

Lemma 6.1.6 *The saddle point problem (6.18) is equivalent to the optimality system (6.12).*

Proof Equation (6.18) amounts to finding $(u, g, p) \in \mathscr{H}^1(D) \times \mathscr{L}^2(\partial D) \times \mathscr{H}^1(D)$, such that

$$\begin{cases} (u, v) + \alpha (g, h)_{\partial D} + \mathscr{B}(v, p) - (h, p)_{\partial D} = (u_d, v) \quad \forall \, v \in \mathscr{H}^1(D), \forall \, h \in \mathscr{L}^2(\partial D), \\ \mathscr{B}(u, q) - (g, q)_{\partial D} = (f, q) \quad \forall \, q \in \mathscr{H}^1(D). \end{cases} \tag{6.25}$$

As we can observe, $(6.25)_2$ coincides with the state equation $(6.12)_1$. Moreover, we can recover the adjoint equation $(6.12)_2$ by setting $h = 0$ in $(6.25)_1$ (notice $\mathscr{B}(v, p) = \mathscr{B}'(p, v)$) and the optimality

condition $(6.12)_3$ by setting $v = 0$ in $(6.25)_1$. Conversely, $(6.25)_1$ is retrieved by adding $(6.12)_2$ and $(6.12)_3$. □

Remark 6.1.1 *Lemmas 6.1.4 and 6.1.6 establish the equivalence between the minimization problem (6.1) subject to the stochastic problem (6.9), the saddle point problem (6.18), and the optimality system (6.12). In particular, the optimality system also admits a unique global optimal solution (6.1) according to Theorem 6.1.5. Moreover, other properties for the saddle point problem are also shared by the optimality system, in particular, the regularity properties of the optimal solution.*

6.2 Stochastic regularity

The convergence properties of the numerical approximation to the stochastic optimality system (6.12) (or to the stochastic saddle point problem (6.18)) in the stochastic space are determined by the regularity of the stochastic solution (u, g, p) or $(\underline{u}, p)$ and the choice of the approximation scheme. Since (6.12) is equivalent to (6.18) by Lemma 6.1.6, we only need to study the regularity of the stochastic solution to the latter with respect to the random variables $y = (y_1, y_2, \ldots, y_N) \in \Gamma := \prod_{n=1}^{N} \Gamma_n$. Our results are provided in Theorem 6.2.1, whose proof is based on recursively applying Brezzi's theorem 6.1.5 in high-dimensional stochastic space, adopting a similar approach as in [54]. An analytical extension of the solution to a certain complex domain is obtained as a consequence to this theorem in Corollary 6.2.2, whose proof follows using Taylor expansion and a Newton binomial formula together with several elementary inequalities extended in high-dimensional stochastic space.

Theorem 6.2.1 *Holding the assumptions in Theorem 6.1.1 and Theorem 6.1.5, we can estimate the partial derivatives of the solution to the stochastic saddle point problem (6.18) with respect to the variables $y = (y_1, \ldots, y_N)$ as follows: $\forall \nu = (\nu_1, \ldots, \nu_N) \in \mathbb{N}^N$,*

$$\begin{aligned} ||\partial_y^{\nu} \underline{u}(y)||_U \leq & \sum_{0 \leq \mu \leq \nu} C_{\nu-\mu}^{\underline{u},u_d} |\nu-\mu|! ||\boldsymbol{b}||_{(L^\infty(D))^d}^{\nu-\mu} ||\partial_y^{\mu} u_d(y)||_{L^2(D)} \\ & + C_{\nu}^{\underline{u},f} |\nu|! ||\boldsymbol{b}||_{(L^\infty(D))^d}^{\nu} ||f||_{L^2(D)}, \end{aligned} \tag{6.26}$$

while for the adjoint variable we obtain the estimate

$$\begin{aligned} ||\partial_y^{\nu} p(y)||_{H^1(D)} \leq & \sum_{0 \leq \mu \leq \nu} C_{\nu-\mu}^{p,u_d} |\nu-\mu|! ||\boldsymbol{b}||_{(L^\infty(D))^d}^{\nu-\mu} ||\partial_y^{\mu} u_d(y)||_{L^2(D)} \\ & + C_{\nu}^{p,f} |\nu|! ||\boldsymbol{b}||_{(L^\infty(D))^d}^{\nu} ||f||_{L^2(D)}. \end{aligned} \tag{6.27}$$

Here, $\mu \leq \nu$ means that $\mu_n \leq \nu_n \forall n = 1, 2, \ldots, N$, and the constant $C_{\nu-\mu}^{\underline{u},u_d} = C_{\nu-\mu}^{\underline{u},u_d}(\alpha_1, \alpha_2, \beta_1, \beta_2)$ is the sum of $2^{|\nu-\mu|}$ basic elements in the form of $\alpha_1^{n_1} \alpha_2^{n_2} \beta_1^{m_1} \beta_2^{m_2}$ such that $n_1 + n_2 + m_1 + m_2 = |\nu - \mu| + 1$. The meaning holds the same for the other constants $C_{\nu-\mu}^{p,u_d}, C_{\nu}^{\underline{u},f}, C_{\nu}^{p,f}$ as coefficients for different terms.

Proof First of all, let us introduce the following pointwise saddle point formulation corresponding to (6.18) as follows: to find $(\underline{u}(y), p(y)) \in U \times H^1(D)$ with $U = H^1(D) \times L^2(\partial D)$, such that

$$\begin{cases} \mathscr{A}(\underline{u}(y), \underline{v}) + \mathscr{B}(\underline{v}, p(y)) = (\underline{u}_d(y), \underline{v}) & \forall \underline{v} \in U, \\ \mathscr{B}(\underline{u}(y), q) = (f, q) & \forall q \in H^1(D), \end{cases} \tag{6.28}$$

where we still denote $\mathscr{A}$ and $\mathscr{B}$ as the pointwise bilinear forms by slight abuse of notation. The properties of continuity for $\mathscr{A}$ and $\mathscr{B}$, coercivity for $\mathscr{A}$ and compatibility condition for $\mathscr{B}$ hold the

same as in Lemma 6.1.4. Moreover, Brezzi's theorem verifies with the same parameters for the stability results (6.23). Explicitly, we have the pointwise stability for $y \in \Gamma$

$$||\underline{u}(y)||_U \leq \alpha_1 ||u_d(y)||_{L^2(D)} + \beta_1 ||f||_{L^2(D)} \tag{6.29}$$

and

$$||p(y)||_{H^1(D)} \leq \alpha_2 ||u_d(y)||_{L^2(D)} + \beta_2 ||f||_{L^2(D)}. \tag{6.30}$$

For $|\nu| = \nu_1 + \nu_2 + \cdots + \nu_N = 0$, we obtain the estimate in the above stability results (6.29) and (6.30). For $|\nu| \geq 1$, by taking partial derivative of the pointwise saddle point problem (6.28) with respect to the random vector y up to order ν, we claim that the general recursive equation is given by

$$\begin{cases} \mathscr{A}(\partial_y^\nu \underline{u}(y), \underline{v}) + \mathscr{B}(\underline{v}, \partial_y^\nu p(y)) = (\partial_y^\nu \underline{u}_d(y), \underline{v}) \\ \qquad\qquad - \sum\limits_{j:\nu_j \neq 0} \nu_j (\boldsymbol{b}_j \cdot \nabla v, \partial_y^{\nu - e_j} p(y)) \quad \forall \underline{v} \in U, \\ \mathscr{B}(\partial_y^\nu \underline{u}(y), q) = - \sum\limits_{j:\nu_j \neq 0} \nu_j (\boldsymbol{b}_j \cdot \nabla \partial_y^{\nu - e_j} u(y), q) \quad \forall q \in H^1(D), \end{cases} \tag{6.31}$$

where $\boldsymbol{b}_j$, $j = 1, 2, \ldots, N$ is the j-th basis in the linear expansion (6.3) and $e_j = (0, \cdots, 1, \cdots, 0)$ is an unit vector with the j-th element set to 1 and all the others 0.

Indeed, if we suppose that $|\tilde{\nu}| = |\nu| - 1$ and $|\tilde{\nu}|$ takes the form $\nu - e_k$ for some k, by hypothesis, (6.31) holds for $\tilde{\nu}$, and then we verify that it also holds for ν. Taking the derivative of (6.31) with respect to y_k and replacing ν by $\nu - e_k$, we have

$$\begin{cases} \mathscr{A}(\partial_y^\nu \underline{u}(y), \underline{v}) + \mathscr{B}(\underline{v}, \partial_y^\nu p(y)) + (\boldsymbol{b}_k \cdot \nabla v, \partial_y^{\nu - e_k} p(y)) = (\partial_y^\nu \underline{u}_d(y), \underline{v}) \\ - \sum\limits_{j \neq k:\nu_j \neq 0} \nu_j (\boldsymbol{b}_j \cdot \nabla v, \partial_y^{\nu - e_j} p(y)) - (\nu_k - 1)(\boldsymbol{b}_k \cdot \nabla v, \partial_y^{\nu - e_k} p(y)) \quad \forall \underline{v} \in U, \\ \mathscr{B}(\partial_y^\nu \underline{u}(y), q) + (\boldsymbol{b}_k \cdot \nabla \partial_y^{\nu - e_k} u(y), q) = - \sum\limits_{j \neq k:\nu_j \neq 0} \nu_j (\boldsymbol{b}_j \cdot \nabla \partial_y^{\nu - e_j} u(y), q) \\ \qquad -(\nu_k - 1)(\boldsymbol{b}_k \cdot \nabla \partial_y^{\nu - e_k} u(y), q) \quad \forall q \in H^1(D). \end{cases} \tag{6.32}$$

By cancelling the same terms in both sides, we retrieve the recursive equation (6.31). Applying Brezzi's theorem to the recursive equation (6.31), we have that there exist unique partial derivatives of the stochastic functions $\partial_y^\nu \underline{u}$ and $\partial_y^\nu p$ such that

$$\begin{aligned} ||\partial_y^\nu \underline{u}(y)||_U \leq \alpha_1 \left(||\partial_y^\nu u_d(y)||_{L^2(D)} + \sum_{j:\nu_j \neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} ||\partial_y^{\nu - e_j} p(y)||_{L^2(D)} \right) \\ + \beta_1 \sum_{j:\nu_j \neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} ||\partial_y^{\nu - e_j} u(y)||_{L^2(D)}, \end{aligned} \tag{6.33}$$

and

$$\begin{aligned} ||\partial_y^\nu p(y)||_{H^1(D)} \leq \alpha_2 \left(||\partial_y^\nu u_d(y)||_{L^2(D)} + \sum_{j:\nu_j \neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} ||\partial_y^{\nu - e_j} p(y)||_{L^2(D)} \right) \\ + \beta_2 \sum_{j:\nu_j \neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} ||\partial_y^{\nu - e_j} u(y)||_{L^2(D)}. \end{aligned} \tag{6.34}$$

When $|\nu| = 1$, i.e., for some $j \in \mathbb{N}, \nu = e_j$, using (6.29) and (6.30), the above recursive estimates (6.33) and (6.34) become

$$\begin{aligned} ||\partial_y^\nu \underline{u}(y)||_U \le \alpha_1 ||\partial_y^\nu u_d(y)||_{L^2(D)} + (\alpha_1\alpha_2 + \alpha_1\beta_1)|\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||u_d(y)||_{L^2(D)} \\ + (\alpha_1\beta_2 + \beta_1\beta_1)|\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||f||_{L^2(D)}, \end{aligned} \tag{6.35}$$

and

$$\begin{aligned} ||\partial_y^\nu p(y)||_{H^1(D)} \le \alpha_2 ||\partial_y^\nu u_d(y)||_{L^2(D)} + (\alpha_2\alpha_2 + \alpha_1\beta_2)|\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||u_d(y)||_{L^2(D)} \\ + (\alpha_2\beta_2 + \beta_1\beta_2)|\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||f||_{L^2(D)}, \end{aligned} \tag{6.36}$$

where $||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} = \Pi_{n=1}^N ||\boldsymbol{b}_n||^{\nu_n}_{(L^\infty(D))^d}$. For a general ν such that $|\nu| \ge 2$, we claim that the estimates (6.26) and (6.27) hold. Note that $||\partial_y^\nu u(y)||_{L^2(D)} \le ||\partial_y^\nu \underline{u}(y)||_U$, $||\partial_y^\nu p(y)||_{L^2(D)} \le ||\partial_y^\nu p(y)||_{H^1(D)}$, and by substituting (6.26) and (6.27) into the recursive formulae (6.33) with ν replaced by $\nu - e_j$, we have

$$\begin{aligned} &||\partial_y^\nu \underline{u}(y)||_U \le \alpha_1 ||\partial_y^\nu u_d(y)||_{L^2(D)} \\ &+ \alpha_1 \sum_{j:\nu_j\neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} \left(\sum_{0\le\mu\le\nu-e_j} C^{p,u_d}_{\nu-e_j-\mu} (|\nu-\mu|-1)!||\boldsymbol{b}||^{\nu-e_j-\mu}_{(L^\infty(D))^d} ||\partial_y^\mu u_d(y)||_{L^2(D)} \right) \\ &+ \alpha_1 \sum_{j:\nu_j\neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} \left(C^{p,f}_{\nu-e_j} (|\nu|-1)!||\boldsymbol{b}||^{\nu-e_j}_{(L^\infty(D))^d} ||f||_{L^2(D)} \right) \\ &+ \beta_1 \sum_{j:\nu_j\neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} \left(\sum_{0\le\mu\le\nu-e_j} C^{\underline{u},u_d}_{\nu-e_j-\mu} (|\nu-\mu|-1)!||\boldsymbol{b}||^{\nu-e_j-\mu}_{(L^\infty(D))^d} ||\partial_y^\mu u_d(y)||_{L^2(D)} \right) \\ &+ \beta_1 \sum_{j:\nu_j\neq 0} \nu_j ||\boldsymbol{b}_j||_{(L^\infty(D))^d} \left(C^{\underline{u},f}_{\nu-e_j} (|\nu|-1)!||\boldsymbol{b}||^{\nu-e_j}_{(L^\infty(D))^d} ||f||_{L^2(D)} \right) \\ &= \alpha_1 ||\partial_y^\nu u_d(y)||_{L^2(D)} + \sum_{0\le\mu\le\nu-e_j} \left(\alpha_1 C^{p,u_d}_{\nu-e_j-\mu} + \beta_1 C^{\underline{u},u_d}_{\nu-e_j-\mu} \right) \times \\ &\qquad \left(\sum_{j:\nu_j\neq 0} \nu_j \right) (|\nu-\mu|-1)!||\boldsymbol{b}||^{\nu-\mu}_{(L^\infty(D))^d} ||\partial_y^\mu u_d(y)||_{L^2(D)} \\ &+ \left(\alpha_1 C^{\underline{u},f}_{\nu-e_j} + \beta_1 C^{p,f}_{\nu-e_j} \right) \left(\sum_{j:\nu_j\neq 0} \nu_j \right) (|\nu|-1)!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||f||_{L^2(D)} \\ &= \sum_{0\le\mu\le\nu} C^{\underline{u},u_d}_{\nu-\mu} |\nu-\mu|!||\boldsymbol{b}||^{\nu-\mu}_{(L^\infty(D))^d} ||\partial_y^\mu u_d(y)||_{L^2(D)} + C^{\underline{u},f}_\nu |\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||f||_{L^2(D)}, \end{aligned} \tag{6.37}$$

where the new coefficients read

$$C^{\underline{u},u_d}_0 = \alpha_1, \quad C^{\underline{u},u_d}_{\nu-\mu} = \left(\alpha_1 C^{p,u_d}_{\nu-e_j-\mu} + \beta_1 C^{\underline{u},u_d}_{\nu-e_j-\mu} \right) \frac{|\nu|}{|\nu-\mu|} \quad \forall\, 0 \le \mu \le \nu - e_j \tag{6.38}$$

and

$$C^{\underline{u},f}_0 = \beta_1, \quad C^{\underline{u},f}_\nu = \left(\alpha_1 C^{p,f}_{\nu-e_j} + \beta_1 C^{\underline{u},f}_{\nu-e_j} \right). \tag{6.39}$$

Carrying out the same procedure for $||\partial_y^\nu p(y)||_{H^1(D)}$, we obtain the estimate

$$\begin{aligned} ||\partial_y^\nu p(y)||_{H^1(D)} \le \sum_{0\le\mu\le\nu} C^{p,u_d}_{\nu-\mu} |\nu-\mu|!||\boldsymbol{b}||^{\nu-\mu}_{(L^\infty(D))^d} ||\partial_y^\mu u_d(y)||_{L^2(D)} \\ + C^{p,f}_\nu |\nu|!||\boldsymbol{b}||^\nu_{(L^\infty(D))^d} ||f||_{L^2(D)}, \end{aligned} \tag{6.40}$$

where the coefficients are

$$C_0^{p,u_d} = \alpha_2, \quad C_{\nu-\mu}^{p,u_d} = \left(\alpha_2 C_{\nu-e_j-\mu}^{p,u_d} + \beta_2 C_{\nu-e_j-\mu}^{\underline{u},u_d}\right)\frac{|\nu|}{|\nu-\mu|} \quad \forall\, 0 \preceq \mu \preceq \nu - e_j \tag{6.41}$$

and

$$C_0^{p,f} = \beta_2, \quad C_\nu^{p,f} = \left(\alpha_2 C_{\nu-e_j}^{p,f} + \beta_2 C_{\nu-e_j}^{\underline{u},f}\right). \tag{6.42}$$

By the above recursive formulae, we have that the constant $C_{\nu-\mu}^{\underline{u},u_d} = C_{\nu-\mu}^{\underline{u},u_d}(\alpha_1,\alpha_2,\beta_1,\beta_2)$ is the sum of $2^{|\nu-\mu|}$ basic elements in the form of $\alpha_1^{n_1}\alpha_2^{n_2}\beta_1^{m_1}\beta_2^{m_2}$ such that $n_1+n_2+m_1+m_2 = |\nu-\mu|+1$. The same structure holds for the constants $C_{\nu-\mu}^{p,u_d}$, $C_\nu^{\underline{u},f}$, and $C_\nu^{p,f}$. Notice the difference that $C_{\nu-\mu}^{\underline{u},u_d}$ and $C_{\nu-\mu}^{p,u_d}$ are modified by some constant related to $|\nu|/|\nu-\mu| \to 1$ as $|\nu| \to \infty$ for fixed μ. □

A direct consequence of the regularity given in Theorem 6.2.1 is the analyticity property of $(\underline{u}, p)$ with respect to $y \in \Gamma$, provided the following conditions are satisfied.

Corollary 6.2.2 *Holding all the assumptions for Theorem 6.2.1 the conditions*

$$2M\sum_n^N ||\boldsymbol{b}_n||_{(L^\infty(D))^d}|y_n - \bar{y}_n| < 1, \tag{6.43}$$

where $M = \max(\alpha_1,\alpha_2,\beta_1,\beta_2)$, *and*

$$\sum_\mu \frac{|\mu||y-\bar{y}|^\mu}{\mu!}||\partial_y^\mu u_d(\bar{y})||_{L^2(D)} < \infty, \tag{6.44}$$

we have the existence of an analytic expansion of the stochastic solution $(\underline{u}, p)$ *to the saddle point problem (6.28) around* $\bar{y} \in \Gamma$. *Therefore,* $(\underline{u}, p)$ *can be analytically extended to the set*

$$\Sigma = \left\{y \in \mathbb{R}^N : \exists\, \bar{y} \in \Gamma \text{ such that } (6.43) \text{ and } (6.44) \text{ hold}\right\}, \tag{6.45}$$

and we define $\Sigma(\Gamma;\tau) := \{z \in \mathbb{C} : dist(z,\Gamma) \le \tau\} \subset \Sigma$ *for the largest possible vector* $\tau = (\tau_1,\ldots,\tau_N)$.

Proof The Taylor expansion of $\underline{u}(y)$, $y \in \Gamma$ around $\bar{y} \in \Gamma$ is given by

$$\underline{u}(y) = \sum_\nu \frac{\partial_y^\nu \underline{u}(\bar{y})}{\nu!}(y-\bar{y})^\nu, \tag{6.46}$$

where $\nu! = \nu_1!\cdots\nu_N!$. By the bound of Theorem 6.2.1, we have the estimate

$$\begin{aligned}\left|\left|\sum_\nu \frac{\partial_y^\nu \underline{u}(\bar{y})}{\nu!}(y-\bar{y})^\nu\right|\right|_U \le \sum_\nu \frac{|y-\bar{y}|^\nu}{\nu!}\left(\sum_{0\le\mu\le\nu} C_{\nu-\mu}^{\underline{u},u_d}|\nu-\mu|!||\boldsymbol{b}||_{(L^\infty(D))^d}^{\nu-\mu}||\partial_y^\mu u_d(\bar{y})||_{L^2(D)}\right)\\ + \sum_\nu \frac{|y-\bar{y}|^\nu}{\nu!}C_\nu^{\underline{u},f}|\nu|!||\boldsymbol{b}||_{(L^\infty(D))^d}^\nu||f||_{L^2(D)},\end{aligned} \tag{6.47}$$

where $|y-\bar{y}| = (|y_1-\bar{y}_1|,\ldots,|y_N-\bar{y}_N|)$. Let us consider the second term at first, for which we introduce the generalized Newton binomial formula: for any $\eta = (\eta_1,\ldots,\eta_N) \in \mathbb{R}^N$ and $k = 0,1,2,\ldots$, we have

$$\sum_{|\nu|=k}\frac{k!}{\nu!}\eta^\nu = \left(\sum_{n=1}^N \eta_n\right)^k. \tag{6.48}$$

By applying (6.48), the second term of (6.47) becomes

$$
\begin{aligned}
&\sum_{\nu} \frac{|y-\bar{y}|^{\nu}}{\nu!} C_{\nu}^{\underline{u},f} |\nu|! ||\boldsymbol{b}||^{\nu}_{(L^{\infty}(D))^d} ||f||_{L^2(D)} \\
&= ||f||_{L^2(D)} \sum_{k=0}^{\infty} \sum_{|\nu|=k} C_{\nu}^{\underline{u},f} \frac{|\nu|!}{\nu!} \left(||\boldsymbol{b}||_{(L^{\infty}(D))^d} |y-\bar{y}| \right)^{\nu} \\
&\leq M||f||_{L^2(D)} \sum_{k=0}^{\infty} \sum_{|\nu|=k} \frac{|\nu|!}{\nu!} \left(2M||\boldsymbol{b}||_{(L^{\infty}(D))^d} |y-\bar{y}| \right)^{\nu} \\
&= M||f||_{L^2(D)} \sum_{k=0}^{\infty} \left(2M \sum_{n}^{N} ||\boldsymbol{b}_n||_{(L^{\infty}(D))^d} |y_n - \bar{y}_n| \right)^{k},
\end{aligned}
\tag{6.49}
$$

where $M = \max(\alpha_1, \alpha_2, \beta_1, \beta_2)$ and the inequality comes from the estimate for the coefficient $C_{\nu}^{\underline{u},f}$

$$
C_{\nu}^{\underline{u},f} \leq 2^{|\nu|} M^{|\nu|+1} = M(2M)^{\nu}, \tag{6.50}
$$

which is valid by definition. Therefore, the convergence condition for (6.49) is

$$
2M \sum_{n}^{N} ||\boldsymbol{b}_n||_{(L^{\infty}(D))^d} |y_n - \bar{y}_n| < 1. \tag{6.51}
$$

As for the first term of the estimate (6.47), we have

$$
\begin{aligned}
&\sum_{\nu} \frac{|y-\bar{y}|^{\nu}}{\nu!} \left(\sum_{0 \leq \mu \leq \nu} C_{\nu-\mu}^{\underline{u},u_d} |\nu-\mu|! ||\boldsymbol{b}||^{\nu-\mu}_{(L^{\infty}(D))^d} ||\partial_y^{\mu} u_d(\bar{y})||_{L^2(D)} \right) \\
&\leq \sum_{\mu} \frac{|y-\bar{y}|^{\mu}}{\mu!} ||\partial_y^{\mu} u_d(\bar{y})||_{L^2(D)} \left(\sum_{\nu \geq \mu} C_{\nu-\mu}^{\underline{u},u_d} \frac{|\nu-\mu|!}{(\nu-\mu)!} \left(||\boldsymbol{b}||_{(L^{\infty}(D))^d} |y-\bar{y}| \right)^{\nu-\mu} \right) \\
&\leq M \sum_{\mu} \frac{|\mu||y-\bar{y}|^{\mu}}{\mu!} ||\partial_y^{\mu} u_d(\bar{y})||_{L^2(D)} \sum_{k=0}^{\infty} \left(2M \sum_{n}^{N} ||\boldsymbol{b}_n||_{(L^{\infty}(D))^d} |y_n - \bar{y}_n| \right)^{k},
\end{aligned}
\tag{6.52}
$$

where for the first inequality we employ the equality

$$
\sum_{\nu} \sum_{0 \leq \mu \leq \nu} \cdot = \sum_{\mu} \sum_{\nu \geq \mu} \cdot \tag{6.53}
$$

and the bound

$$
\frac{1}{\nu!} \leq \frac{1}{\mu!} \frac{1}{(\nu-\mu)!}. \tag{6.54}
$$

For the second inequality, we replace all $\nu - \mu$ by ν, bound the coefficient $C_{\nu-\mu}^{\underline{u},u_d}$ by

$$
C_{\nu-\mu}^{\underline{u},u_d} \leq \frac{|\nu|}{|\nu-\mu|} 2^{|\nu-\mu|} M^{|\nu-\mu|+1} \leq M|\mu|(2M)^{|\nu-\mu|}, \tag{6.55}
$$

and use the result obtained for the second term (6.49). Hence, the convergence of the first term (6.52) is guaranteed by the condition (6.51) as well as the condition (6.44), which implies that u_d is analytic around $\bar{y}$. The same procedure holds for the Taylor expansion of

$$
p(y) = \sum_{\nu} \frac{\partial_y^{\nu} p(\bar{y})}{\nu!} (y-\bar{y})^{\nu} \tag{6.56}
$$

with estimate in the space of $H^1(D)$. □

6.3 Approximation and error estimates

Numerical approximation in both the physical space and the stochastic space will be studied in this section. More specifically, we apply stabilized finite element approximation in physical space in order to address the advection dominated problem [56, 15, 97, 126, 6, 77] and employ sparse grid stochastic collocation approximation in stochastic space [207, 8, 148, 149, 12] to deal with the computational reduction for the high-dimensional stochastic problem; see chapters 1 for the development of the sparse grid stochastic collocation method. Convergence properties of the approximations in both physical and stochastic space will be provided separately. Finally, we derive a global error estimate for a combined stabilized finite element–adaptive stochastic collocation approximation.

6.3.1 Finite element approximation in physical space

Let us introduce a regular triangulation $\mathcal{T}_h$ of the physical domain $D \subset \mathbb{R}^d, d = 2,3$, such that $\bar{D} = \cup_{K\in\mathcal{T}_h} K$ and $diam(K) \le h$. Based on this triangulation, we define a finite element space X_h^k

$$X_h^k := \left\{ v_h \in C^0(\bar{D}) \middle| v_{h|K} \in \mathbb{P}_k \quad \forall\, K \in \mathcal{T}_h \right\}, \quad k \ge 0, \tag{6.57}$$

where $\mathbb{P}_k, k \ge 0$ is the space of polynomials of total degree less than or equal to k in the variables $x_1, \ldots, x_d$. Therefore, the element v_h in X_h^k is simply a piece-wise polynomial, and we have that $X_h^k \subset H^1(D)$ [165]. Since both the state equation and the adjoint equation are advection dominated, a proper stabilization method is needed. Let us introduce the operator for the pointwise state equation as follows: $\forall\, y \in \Gamma$, define

$$Lu(y) := -\nabla\cdot(a\nabla u(y)) + \boldsymbol{b}(y)\cdot\nabla u(y) + cu(y), \tag{6.58}$$

which can be separated into a symmetric part and a skew-symmetric part $L = L_s + L_{ss}$, defined as

$$L_s u(y) = -\nabla\cdot(a\nabla u(y)) + (c - \nabla\cdot\boldsymbol{b}(y)/2)u(y); \quad L_{ss}u(y) = (\boldsymbol{b}(y)\cdot\nabla u(y) + \nabla\cdot(\boldsymbol{b}(y)u(y)))/2. \tag{6.59}$$

Corresponding to the adjoint equation, we define the adjoint operator: $\forall\, y \in \Gamma$, define

$$L'p(y) := -\nabla\cdot(a\nabla p(y)) - \boldsymbol{b}(y)\cdot\nabla p(y) + (c - \nabla\cdot\boldsymbol{b}(y))p(y), \tag{6.60}$$

and we split it into a symmetric part and a skew-symmetric part, $L' = L'_s + L'_{ss}$, and we have

$$L'_s p(y) = -\nabla\cdot(a\nabla p(y)) + (c - \nabla\cdot\boldsymbol{b}(y)/2)p(y); \quad L'_{ss}p(y) = -(\boldsymbol{b}(y)\cdot\nabla p(y) + \nabla\cdot(\boldsymbol{b}(y)p(y)))/2. \tag{6.61}$$

Substituting the optimality condition (6.12) into the state equation (6.2) and taking the following stabilized weak formulation for both the state equation and adjoint equation, we obtain the stabilized and reduced optimality system in finite element space X_h^k as follows [165]:

$$\begin{cases} \mathscr{B}(u_h(y), v_h) + \sum\limits_{K\in\mathcal{T}_h} \delta_K \left(Lu_h(y), \dfrac{h_K}{|\boldsymbol{b}(y)|}(L_{ss} + \theta L_s)v_h \right) = \dfrac{1}{\alpha}(p_h(y), v_h) \\ \qquad +(f, v_h) + \sum\limits_{K\in\mathcal{T}_h} \delta_K \left(f, \dfrac{h_K}{|\boldsymbol{b}(y)|}(L_{ss} + \theta L_s)v_h \right) \quad \forall\, v_h \in X_h^k, \\ \mathscr{B}'(p_h(y), v_h) + \sum\limits_{K\in\mathcal{T}_h} \delta_K \left(L'p_h(y), \dfrac{h_K}{|\boldsymbol{b}(y)|}(L'_{ss} + \theta L'_s)v_h \right) = (u_d(y) - u_h(y), v_h) \\ \qquad + \sum\limits_{K\in\mathcal{T}_h} \delta_K \left(u_d(y) - u_h(y), \dfrac{h_K}{|\boldsymbol{b}(y)|}(L'_{ss} + \theta L'_s)v_h \right) \quad \forall\, v_h \in X_h^k, \end{cases} \tag{6.62}$$

where $|\boldsymbol{b}(y)|$ is the modulus of $\boldsymbol{b}(y)$ and the parameter δ_K is left to be chosen, for instance,

$$\delta_K \equiv \delta(Pe_K) := \coth(Pe_K) - \frac{1}{Pe_K}, \text{ where } Pe_K := \frac{|\boldsymbol{b}(y)|h_K}{2\min_K a(x)} \quad \forall K \in \mathscr{T}_h. \tag{6.63}$$

Different stabilization methods result from the choice of θ. If $\theta = 0$, it corresponds to streamline upwind/Petrov–Galerkin (SUPG) stabilization; if $\theta = 1$, Galerkin/least-squares stabilization is obtained. For these different stabilization methods, it has been proved that if the parameter δ is small enough and the solution of the state equation (6.2) is regular enough, e.g., $u(y) \in H^{k+1}(D)$, the priori error of the approximation is bounded by the estimate $||u-u_h||_V \leq h^{k+1/2}||u||_{H^{k+1}(D)}$, where the norm $||\cdot||_V$ is defined according to different methods. More details about the strong consistency and accuracy of these stabilization methods are provided in [165]. As for the convergence property of the optimality system, we have the following pointwise results by optimize-then-discretize procedure for SUPG stabilization; see similar proof in [56] for distributed optimal control problem.

Lemma 6.3.1 *Let $k,l,m \geq 1$, and suppose that $\forall\, y \in \Gamma$ the solution $(u(y),g(y),p(y))$ satisfies $u(y) \in H^{k+1}(D)$, $g(y) \in H^{m+1/2}(\partial D)$, and $p(y) \in H^{l+1}(D)$. If the stabilization parameter satisfies*

$$\delta_K \leq \min\left(\frac{h_K^2}{\varepsilon C_K^2}, \frac{r'}{||c||_{L^\infty(K)}}, \frac{r'}{||c-\nabla\cdot\boldsymbol{b}(y)||_{L^\infty(K)}}\right) \quad \forall K \in \mathscr{T}_h, \tag{6.64}$$

where $\varepsilon = a_{max} \leq R$, r' is the coefficient defined in (6.5) and C_K is the constant for the inverse inequality $||\nabla v_h||_{L^2(K)} \leq C_K h_K^{-1}||v_h||_{L^2(K)}\ \forall K \in \mathscr{T}_h$, and we take for positive constant $\zeta_1, \zeta_2 > 0$

$$\delta_K = \zeta_1 \frac{h_K^2}{\varepsilon} \text{ for } Pe_K \leq 1, \text{ or } \delta_K = \zeta_2 h_K \text{ for } Pe_K > 1, \tag{6.65}$$

then the error estimate for the discretized optimal solution $(u_h(y),g_h(y),p_h(y))\ \forall\, y \in \Gamma$ is obtained as

$$\begin{aligned} &||u(y)-u_h(y)||_V + ||g(y)-g_h(y)||_{L^2(\partial D)} + ||p(y)-p_h(y)||_V \\ &\leq C\Big((\varepsilon^{1/2}+h^{1/2})(h^k|u(y)|_{k+1} + h^l|p(y)|_{l+1}) + h^{m+1}|g(y)|_{m+1/2,\partial D}\Big), \end{aligned} \tag{6.66}$$

where the norm $||\cdot||_V$ is defined for SUPG stabilization as

$$||v||_V^2 = \varepsilon|v|_1^2 + r'||v||_{L^2(K)}^2 + \sum_{K\in\mathscr{T}_h} \delta_K||\boldsymbol{b}(y)\cdot\nabla v||_{L^2(K)}^2, \tag{6.67}$$

and $|v|_k, k \geq 1$ is the seminorm in the Hilbert space $H^k(D), k \geq 1$.

Remark 6.3.1 *Lemma 6.3.1 provides a convergence result for the error between the solution of the original and that of the discretized optimal control problem over the entire domain D. The constants in the estimates of the global error depend on regularity of the optimal solution (u,g,p). Similar results have also been obtained recently in [15, 26, 97].*

6.3.2 Collocation approximation in stochastic space

In this section, we apply the stochastic collocation method (introduced in chapter 1, section 1.2) for the approximation of the optimal solution (u,g,p) in the stochastic space. The tensor-product approximation error and the sparse grid approximation error are summarized in the next two lemmas.

Lemma 6.3.2 *The following convergence estimate holds for the multidimensional full tensor product*

interpolation operator $\mathscr{I}_q$

$$||u-\mathscr{I}_q u||_{\mathcal{V}(D)}+||g-\mathscr{I}_q g||_{\mathscr{L}^2(\partial D)}+||p-\mathscr{I}_q p||_{\mathcal{V}(D)}\leq C\sum_{n=1}^{N}e^{-\ln(r_n)q_n}, \tag{6.68}$$

where the norm $\mathcal{V}(D):=L^2_\rho(\Gamma;V)$*; C is a positive constant independent of N;* $q_n, n=1,\ldots,N$ *are the polynomial orders of the univariate Lagrange interpolation formula; the constants* $r_n, n=1,\ldots,N$*, are defined via* τ_n *and* Γ_n *as*

$$r_n=\frac{2\tau_n}{|\Gamma_n|}+\sqrt{1+\frac{4\tau_n^2}{|\Gamma_n|^2}}>1,\quad n=1,\ldots,N. \tag{6.69}$$

Proof The proof is obtained immediately by combining the analytical regularity result of the stochastic optimal solution in Corollary 6.2.2 and the convergence result of the tensor-product stochastic collocation method in Proposition 1.4.3 (up to a difference from L^∞ norm to L^2_ρ norm). □

Lemma 6.3.3 *The approximation error of the stochastic optimal solution by the isotropic sparse grid stochastic collocation method with Clenshaw–Curtis collocation nodes is bounded by*

$$||u-\mathscr{I}_q u||_{\mathcal{V}(D)}+||g-\mathscr{I}_q g||_{\mathscr{L}^2(\partial D)}+||p-\mathscr{I}_q p||_{\mathcal{V}(D)}\leq CN_q^{-r}, \tag{6.70}$$

where C is a constant independent of N_q *and r,* $N_q=\#H(q,N)$ *is the number of collocation nodes, r is defined as* $r=\min(\log(\sqrt{r_1}),\ldots,\log(\sqrt{r_N}))/(1+\log(2N))$ *with* $r_1,\ldots,r_N$ *defined in (6.69). Then using the anisotropic Smolyak sparse grid, still with Clenshaw–Curtis collocation nodes, we have*

$$||u-\mathscr{I}_q u||_{\mathcal{V}(D)}+||g-\mathscr{I}_q g||_{\mathscr{L}^2(\partial D)}+||p-\mathscr{I}_q p||_{\mathcal{V}(D)}\leq CN_q^{-r(\boldsymbol{\alpha})}, \tag{6.71}$$

where $r(\boldsymbol{\alpha})=\min(\boldsymbol{\alpha})(\log(2)e-1/2)/\left(\log(2)+\sum_{n=1}^{N}\min(\boldsymbol{\alpha})/\alpha_n\right)$ *and* $\alpha_n=\log(\sqrt{r_n})$*,* $n=1,\ldots,N$*.*

Proof The results are a direct consequence of the analytic regularity of the stochastic optimal solution in Corollary 6.2.2 and the convergence results in Proposition 1.4.4 (up to a difference from L^∞ norm to L^2_ρ norm). □

6.3.3 Convergence for approximating stochastic optimal control problem

In this section, we provide some convergence results for the stabilized finite element approximation in the physical space and stochastic collocation approximation in the stochastic space for the optimality system (6.12), or equivalently, the saddle point system (6.18). Let us denote the fully approximated solution in both the physical space and the stochastic space as $(u_{h,q},g_{h,q},p_{h,q})$. We summarize in Theorems 6.3.4 and 6.3.5 the error estimates for tensor product grid collocation approximation and sparse grid collocation approximation for stochastic Robin boundary control problem, respectively.

Theorem 6.3.4 *Provided that the assumptions made in Corollary 6.2.2 and Lemma 6.3.1 are satisfied, the following global error estimate for stabilized finite element approximation in the physical space and full tensor product grid collocation approximation in the stochastic space for the stochastic optimality system (6.12) (or equivalently the saddle point system (6.18)) holds:*

$$\begin{aligned}&||u-u_{h,q}||_{\mathcal{V}(D)}+||g-g_{h,q}||_{\mathscr{L}^2(\partial D))}+||p-p_{h,q}||_{\mathcal{V}(D)}\\&\leq C\sum_{n=1}^{N}e^{-\ln(r_n)q_n}+C_p(\varepsilon^{1/2}+h^{1/2})h^k\left(|u|_{\mathscr{H}^{k+1}(D)}+|p|_{\mathscr{H}^{k+1}(D)}+h|g|_{\mathscr{H}^{k+1/2}(\partial D)}\right).\end{aligned} \tag{6.72}$$

Here, $C, (r_n, q_n), n = 1, \ldots, N$ are the constants for approximation in the stochastic space inherited from Lemma 6.3.2 and C_p is the constant for approximation in the physical space inherited from Lemma 6.3.1. The quantity $|u|_{\mathscr{H}^{k+1}(D)}, |p|_{\mathscr{H}^{k+1}(D)}$ and $|g|_{\mathscr{H}^{k+1/2}(\partial D)}$ are the semi-norm of u, p, g in the stochastic Hilbert spaces.

Proof Recall the interpolation operator $\mathscr{I}_q : (u, g, p) \to \mathscr{I}_q(u, g, p) \equiv (u_q, g_q, p_q)$ in the stochastic space. Denoting by $P_h^s : (u_q, g_q, p_q) \to P_h^s(u_q, g_q, p_q) \equiv (u_{h,q}, g_{h,q}, p_{h,q})$ the pointwise projection operator for the stabilized finite element approximation in the physical space, which projects the pointwise solution $(u_q(y), g_q(y), p_q(y))$ for any $y \in \Gamma$ from the Hilbert space $H^{k+1}(D) \times H^{k+1/2}(\partial D) \times H^{k+1}(D)$ to the finite element space $X_h^k \times X_h^k|_{\partial D} \times X_h^k$, we conclude the convergence result for the combined approximation

$$
\begin{aligned}
&||u - u_{h,q}||_{\mathcal{V}(D)} + ||g - g_{h,q}||_{\mathscr{L}^2(\partial D))} + ||p - p_{h,q}||_{\mathcal{V}(D)} \\
&\equiv ||u - u_{h,q}||_{L^2_\rho(\Gamma;V)} + ||g - g_{h,q}||_{L^2_\rho(\Gamma;L^2(\partial D))} + ||p - p_{h,q}||_{L^2_\rho(\Gamma;V)} \\
&= ||u - P_h^s I_q u||_{L^2_\rho(\Gamma;V)} + ||g - P_h^s I_q g||_{L^2_\rho(\Gamma;L^2(\partial D))} + ||p - P_h^s I_q p||_{L^2_\rho(\Gamma;V)} \\
&\le ||u - I_q u||_{L^2_\rho(\Gamma;V)} + ||I_q u - P_h^s I_q u||_{L^2_\rho(\Gamma;V)} \\
&\quad + ||g - I_q g||_{L^2_\rho(\Gamma;L^2(\partial D))} + ||I_q g - P_h^s I_q g||_{L^2_\rho(\Gamma;L^2(\partial D))} \\
&\quad + ||p - I_q p||_{L^2_\rho(\Gamma;V)} + ||I_q p - P_h^s I_q p||_{L^2_\rho(\Gamma;V)} \\
&= ||u - I_q u||_{L^2_\rho(\Gamma;V)} + ||g - I_q g||_{L^2_\rho(\Gamma;L^2(\partial D))} + ||p - I_q p||_{L^2_\rho(\Gamma;V)} \\
&\quad + ||I_q u - P_h^s I_q u||_{L^2_\rho(\Gamma;V)} + ||I_q g - P_h^s I_q g||_{L^2_\rho(\Gamma;L^2(\partial D))} + ||I_q p - P_h^s I_q p||_{L^2_\rho(\Gamma;V)} \\
&\le C \sum_{n=1}^{N} e^{-\ln(r_n) q_n} + C_p(\varepsilon^{1/2} + h^{1/2}) h^k \left(|u|_{\mathscr{H}^{k+1}(D)} + |p|_{\mathscr{H}^{k+1}(D)} + h|g|_{\mathscr{H}^{k+1/2}(\partial D)} \right).
\end{aligned}
\tag{6.73}
$$

The first inequality is due to the triangular inequality, and the second one follows from using the converge results of the stabilized finite element approximation and the stochastic collocation approximation. □

Using similar arguments, we have the following convergence result for the isotropic or anisotropic sparse grid stochastic collocation approximation.

Theorem 6.3.5 *If the assumptions in Corollary 6.2.2 and Lemma 6.3.1 are satisfied, we have the following global error estimate for stabilized finite element approximation in the physical space and isotropic or anisotropic sparse grid collocation approximation in the stochastic space:*

$$
\begin{aligned}
&||u - u_{h,q}||_{\mathcal{V}(D)} + ||g - g_{h,q}||_{\mathscr{L}^2(\partial D))} + ||p - p_{h,q}||_{\mathcal{V}(D)} \\
&\le C N_q^{-r(\alpha)} + C_p(\varepsilon^{1/2} + h^{1/2}) h^k \left(|u|_{\mathscr{H}^{k+1}(D)} + |p|_{\mathscr{H}^{k+1}(D)} + h|g|_{\mathscr{H}^{k+1/2}(\partial D)} \right),
\end{aligned}
\tag{6.74}
$$

where C_p is the constant for approximation in physical space inherited from Lemma 6.3.1, and C, N_q and $r(\alpha)$ are the constants for approximation in stochastic space inherited from Lemma 6.3.3.

6.4 Numerical results

In this section, we demonstrate by numerical experiments our error estimates for the approximation of the stochastic optimal Robin boundary control problem obtained in the last section. Specifically, we test the error estimates for stabilized finite element approximation in physical space and sparse grid collocation approximation in stochastic space, respectively.

The computational domain is a two-dimensional unit square $x = (x_1, x_2) \in D = (0,1)^2$; the coefficients $a = 0.01$, $c = 1$, $k = 1$ and the force term $f = 0.1$, all constants, are fixed; the advection field $\boldsymbol{b} = (b_{x_1}, b_{x_2})^T$ is a stochastic vector function, with the second component $b_{x_2} = 0$ and the first component b_{x_1} as a random field with finite second moment, with expectation and correlation

$$\mathbb{E}[b_{x_1}](x) = x_2(1-x_2); \mathbb{C}ov[b_{x_1}](x,x') = \frac{x_2^2(1-x_2)^2}{10^2}\exp\left(-\frac{(x_1-x_1')^2}{L^2}\right), x, x' \in D \tag{6.75}$$

where L is the correlation length. By Karhunen-Loève expansion as introduced in (27) of the preliminary chapter, b_{x_1} can be written as

$$\begin{aligned} b_{x_1}(x,\omega) = x_2(1-x_2) &+ \frac{x_2(1-x_2)}{10}\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1(\omega) \\ &+ \frac{x_2(1-x_2)}{10}\sum_{n=1}^{\infty}\sqrt{\lambda_n}\left(\sin(n\pi x_1)y_{2n}(\omega) + \cos(n\pi x_1)y_{2n+1}(\omega)\right), \end{aligned} \tag{6.76}$$

where the uncorrelated random variables $y_n, n \geq 1$ have zero mean and unit variance, and the eigenvalues $\lambda_n, n \geq 1$ decay as follows:

$$\sqrt{\lambda_n} = \left(\sqrt{\pi}L\right)^{1/2}\exp\left(-\frac{(n\pi L)^2}{8}\right) \quad \forall\, n \geq 1. \tag{6.77}$$

As for Robin boundary condition g, we assume that its expectation and correlation function are given on the same segment of the boundary,

$$\mathbb{E}[g](x) = 1; \mathbb{C}ov[g](x,x') = \frac{1}{2^2}\exp\left(-\frac{(x_1-x_1')^2 + (x_2-x_2')^2}{L^2}\right), x, x' \in \partial D. \tag{6.78}$$

The Karhunen-Loève expansion of the stochastic Robin boundary condition is written, e.g., on $0 \times [0,1]$

$$g(x,\omega) = 1 + \frac{1}{2}\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1(\omega) + \frac{1}{2}\sum_{n=1}^{\infty}\sqrt{\lambda_n}\left(\sin(n\pi x_2)y_{2n}(\omega) + \cos(n\pi x_2)y_{2n+1}(\omega)\right), \tag{6.79}$$

where $\lambda_n, n \geq 1$ are the same as in (6.77). In the numerical examples, we truncate the expansion up to N terms and assume that the random variables are independent and obey the same uniform distribution $y_n \sim \mathscr{U}(-\sqrt{3}, \sqrt{3}), n = 1, \ldots, N$ with zero mean and unit variance. For the sake of simplicity, we do not consider the contribution of the truncation error and focus on the stochastic collocation approximation error.

As the first test example, let us choose the correlation length $L = 1/4$ for both velocity $\boldsymbol{b}$ and Robin boundary condition g, for which we only need 7 terms in both of the truncations and therefore 15 independent random variables. Using piecewise linear function space X_h^1, $h = 0.025$ for stabilized finite element approximation and isotropic sparse grid collocation approximation with Clenshaw–Curtis collocation nodes as $\mathscr{S}_q, q = 19$ in (1.15), we can compute the solution for the stochastic advection dominated elliptic problem (6.2) on each of 2792 unstructured finite element nodes in D and each of 40001 collocation nodes in Γ. The expectation and standard deviation of the solution, which may represents the temperature distribution of a heat transfer problem, can also be evaluated by quadrature formula introduced in chapter 1; see the results in Figure 6.1.

Taking the solution as our objective function $u_d = u$, and solving the stabilized optimality system (6.62), we obtain the optimal solution (u, g, p). The expectation and standard deviation of the stochastic Robin boundary control function g is displayed on the left of Figure 6.2, which is very close to the theoretical value $\mathbb{E}[g] \equiv \mu = 1$ and $\mathbb{V}ar[g] \equiv \sigma = 0.4876$ computed directly from (6.79).

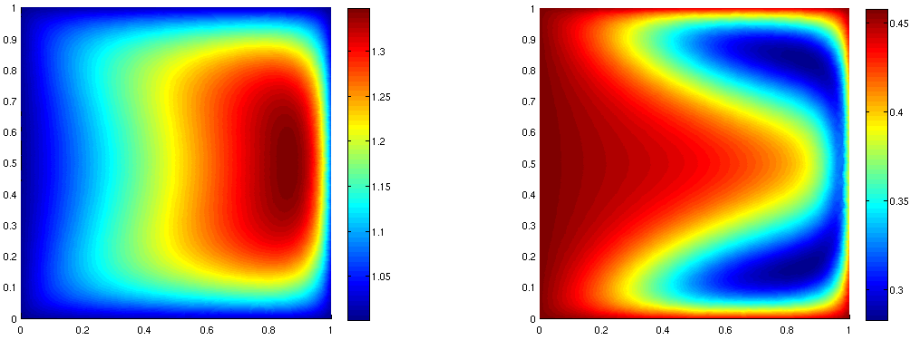

Figure 6.1: Expectation (left) and standard deviation (right) of the solution of problem (6.2)

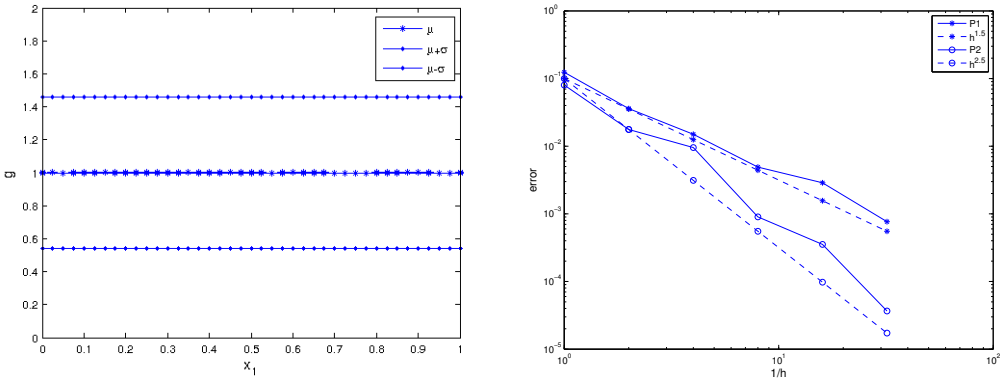

Figure 6.2: Expectation μ and standard deviation σ of the Robin boundary condition g (left), and convergence rate of the error of the solution in stabilized finite element space X_h^1 and X_h^2 (right)

In order to verify the theoretical convergence rate in different finite element spaces, we choose X_h^1 and X_h^2, where for the second one we replace h_K by h_K^2 in the specification of Péclet number Pe_K in (6.63) in order to have an approximately quadratic decay of the parameter δ_K with respect to h_K in (6.65) when $Pe_K \leq 1$ for small h. In fact, from (6.63) we have $\delta_K = \coth(Pe_K) - 1/Pe_K = \coth(O(h_K^2)) - 1/O(h_K^2) \approx O(h_K^2)$. The series of h are $h = 1, 1/2, 1/2^2, 1/2^3, 1/2^4, 1/2^5$. The error is defined as

$$error = ||u - u_{h,q}||_{\mathcal{V}(D)} + ||g - g_{h,q}||_{\mathcal{L}^2(\partial D))} + ||p - p_{h,q}||_{\mathcal{V}(D)}, \tag{6.80}$$

where u is computed by setting $h = 1/2^6$ and $q = 19$, g is given by formula (6.79), the adjoint variable p is set as 0, and $(u_{h,q}, g_{h,q}, p_{h,q})$ is computed by solving the optimality system (6.62). The convergence results is shown on the right of Figure 6.2, which implies that the error decays approximately with order $h^{1.5}$ for X_h^1 and order $h^{2.5}$ for X_h^2, consistently with our theoretical result in Theorem 6.3.5.

For simplicity, we use the same set of random variables for the expansion of b_{x_1} and g in order to test the convergence rate of the collocation approximation. The same error defined in (6.80) is used. For the test of isotropic sparse grid collocation approximation, we use the series of different levels of interpolation $1, 2, 3, 4, 5, 6, 7$ and set the approximated value in the deepest level as the true solution. The correlation length is set as $L = 1/4$ and the number of random variables $\#rv = 3, 5, 7$. The step size for the stabilized finite element approximation h is set to be a relatively large value 0.25 to accelerate the computation.

The error against the number of collocation nodes is displayed on the left of Figure 6.3, from which we can see that the convergence rate decreases as the number of random variables increases, and the comparison of the convergence rate with $O(1/N^2)$ and $O(1/N)$ shows that the isotropic sparse grid collocation approximation is faster than Monte Carlo method whose convergence rate is $O(1/N^{1/2})$.

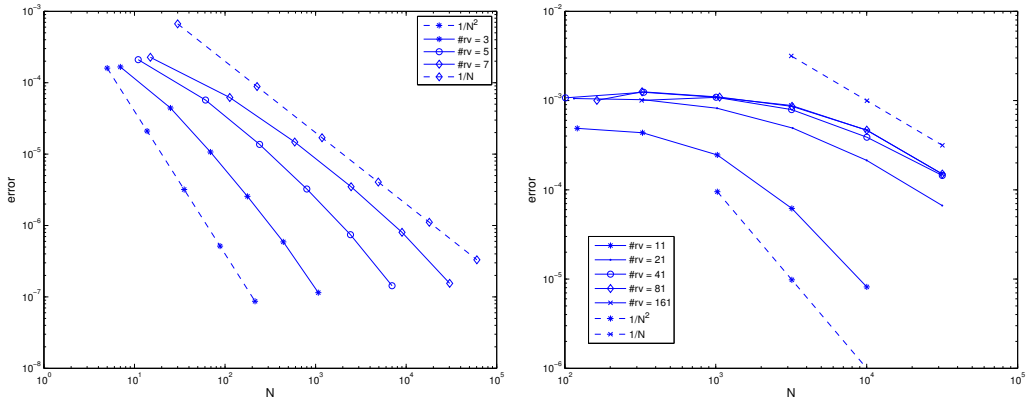

Figure 6.3: Comparison of convergence rate by isotropic sparse grid collocation approximation (left) and anisotropic sparse grid collocation approximation (right) for different dimensions N

However, when the number of random variables becomes very large, this potential advantage will fade down. In this case, we need to approximate with high interpolation level in those dimensions that are more important than the others, using the anisotropic sparse grid with dimension-adaptive tensor-product quadrature [79]. On the right of Figure 6.3, we show the convergence rate with even smaller correlation length $L = 1/16$ for high-dimensional approximation $\#rv = 11, 21, 41, 81, 161$. We set a series of collocation nodes with the cardinality as $10^2, 10^{2.5}, 10^3, 10^{3.5}, 10^4, 10^{4.5}, 10^5$ and use the solution u computed with 10^5 collocation nodes as the true solution. The same stepsize $h = 0.25$ is used. From the figure we can see that the anisotropic sparse grid breaks the curse of dimensionality in the sense of being able to taking care of very high-dimensional stochastic anisotropic problems. Moreover, the convergence rate can be compared to $O(1/N^2)$ for 11 dimensions and $O(1/N)$ for over 41 dimensions, which are both higher than $O(1/N^{1/2})$. The convergence rate becomes almost the same for dimensions over 41 since the randomness is captured over 99% by $n = 26$ terms truncation for $L = 1/16$, so that the left random variables play a very little role.

From the above numerical results, we can conclude that the theoretical results obtained in the last section are very well verified. Meanwhile, the isotropic sparse grid stochastic collocation approximation is very efficient for stochastic optimal control problems with moderate dimensions, and the application of the anisotropic sparse grid is able to deal with high-dimensional stochastic problems with different weights in different dimensions (up to the order $O(10^2)$).

6.5 Summary

In this chapter, we presented a stochastic optimal Robin boundary control problem constrained by an advection dominated elliptic equation. The particular uncertainties we considered arise from the background velocity of the advection term, the objective function, as well as the stochastic optimal control function. We introduced the stochastic saddle point formulation and proved its equivalence to the first order necessary optimality system for the stochastic optimal control problem. The stochastic regularity with respect to the random variables was obtained thanks to Brezzi's theorem for the saddle point system. We applied stabilized finite element approximation in physical space and stochastic

collocation approximation in stochastic space to discretize the optimality system. A global error estimate was obtained for the approximation. In the last part, the error estimate is verified by numerical experiments, with anisotropic sparse grid collocation approximation being highlighted for treating very high-dimensional stochastic problems. Further analysis of other approximations, e.g., adaptive hierarchical stochastic collocation approximation and weighted reduced basis approximation [49, 48], and applications of them to more general distributed and boundary stochastic optimal control problems are promising, e.g., stochastic optimal control constrained by evolution equations. As we will see in the next two chapters, the reduced basis method can be efficiently applied to solve stochastic optimal control problems constrained by, e.g. elliptic equations (chapter 7) and Stokes equations (chapter 8).

7 Reduced basis method for stochastic elliptic optimal control problems

As shown in the last chapter and on some recent works [172, 197, 112, 47], the sparse grid stochastic collocation method can be effectively applied to solve stochastic optimal control problems. However, when the solution of the optimality system becomes computationally very expensive, only a few tens or hundreds of high-fidelity solutions are affordable, which prohibits direct application of the stochastic collocation method since a much larger number (hundreds of thousands as shown in the last chapter) of the full optimality system should be solved. More efficient techniques are therefore needed in order to alleviate the whole computational effort in the many-query optimization context. Model order reduction techniques such as proper orthogonal decomposition [167] or reduced basis method are promising in this perspective; see [104, 58, 59, 85, 132, 144] for the application of the latter method in solving parametrized optimal control problems.

In this chapter, we consider the stochastic optimal control problem constrained by a linear elliptic equation with distributed stochastic control function. After providing an analysis of well-posedness (existence, uniqueness and stability of the stochastic optimal solution), we use finite element method with (optimal) preconditioning techniques for deterministic approximation of the optimal solution in physical space, and stochastic collocation method for stochastic approximation in the probability space. We tailor the weighted reduced basis method (Chapter 2) to reduce the computational cost when solving a considerable number of optimality systems, leading to a reduced optimality system that enables many-query solutions with a posteriori error estimate. Convergence results and remarks on computational efficiency are illustrated by numerical tests in multidimensional probability space.

This chapter is organized as follows. In section 7.1 we state the stochastic optimal control problems with elliptic PDE constraints and random inputs. Section 7.2 is devoted to the study of the mathematical properties of the stochastic optimal control problems. In section 7.3, we present numerical approximation to solve the stochastic optimality system, consisting of finite element method, tensor-product and sparse-grid stochastic collocation method and weighted reduced basis method. Numerical tests for verification and illustration of our method are reported in section 7.4. We close this chapter by a summary of the computational methodology and indicating possible future developments in the last section 7.5.

Reference for this chapter:

P. Chen and A. Quarteroni. Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraints. To appear in SIAM/ASA Journal on Uncertainty Quantification, 2014.

7.1 Problem statement

Recall the definitions of the random vector, stochastic Hilbert space and PDE in the preliminary chapter, and we consider the following elliptic homogeneous Dirichlet boundary value problem with distributed control function

$$\begin{cases} -\nabla\cdot(a(x,y)\nabla u(x,y)) &= f(x,y)+g(x,y) \qquad \forall (x,y)\in D\times\Gamma, \\ u(x,y) &= 0 \qquad\qquad\qquad\qquad\quad \forall (x,y)\in \partial D\times\Gamma, \end{cases} \tag{7.1}$$

where a is a random coefficient field, f is a random force field and g is a random field representing a distributed control. Recall the Assumptions 0.2 and 0.3 for the random data. We make the following assumptions with slight modifications.

Assumption 7.1 *The random coefficient a is uniformly bounded from above and below, i.e., there exist positive constants $0<r<R<\infty$ such that*

$$P\left(\omega\in\Omega : r||v||^2_{H^1(D)} \le (a(\cdot,y(\omega))\nabla v,\nabla v) \le R||v||^2_{H^1(D)}\right)=1, \quad \forall v\in H^1(D). \tag{7.2}$$

The random force term f and random control function g have bounded second moment

$$\int_\Gamma\int_D f^2(x,y)\rho(y)dxdy<\infty \text{ and } \int_\Gamma\int_D g^2(x,y)\rho(y)dxdy<\infty. \tag{7.3}$$

Moreover, recall that the random fields a and f admit the linear expansion as given by (31)

$$a(x,y(\omega))=a_0(x)+\sum_{k=1}^{K}a_k(x)y_k(\omega) \text{ and } f(x,y(\omega))=f_0(x)+\sum_{k=1}^{K}f_k(x)y_k(\omega). \tag{7.4}$$

Let us denote $y_0=1$ for ease of notation. Under Assumption 7.1, we have the following weak formulation for problem (7.1): find $u\in\mathscr{H}^1_0(D)$ such that

$$\mathscr{B}(u,v)=\mathscr{F}(v)+\mathscr{G}(v) \quad \forall v\in\mathscr{H}^1_0(D), \tag{7.5}$$

where $\mathscr{H}^1_0(D):=\{v\in\mathscr{H}^1(D), v=0 \text{ on } \partial D\}$, $\mathscr{G}(v)=(g,v)$ and the bilinear form $\mathscr{B}$ and the linear functional $\mathscr{F}$ are defined as

$$\mathscr{B}(u,v)=\sum_{k=0}^{K}\int_\Gamma\left(B_k(u,v)y_k\right)\rho(y)dy \text{ and } \mathscr{F}(v)=\sum_{k=0}^{K}\int_\Gamma\left(F_k(v)y_k\right)\rho(y)dy, \tag{7.6}$$

with $B_k(u,v)=(a_k\nabla u,\nabla v); F_k(v)=(f_k,v), k=0,\dots,K$.

Theorem 7.1.1 *Provided that the data satisfy Assumption 7.1, there exists a unique solution $u\in\mathscr{H}^1_0(D)$ of problem (7.5) such that*

$$||u||_{\mathscr{H}^1_0(D)} \le (C_R/r)\left(||f||_{\mathscr{L}^2(D)}+||g||_{\mathscr{L}^2(D)}\right), \tag{7.7}$$

where C_R is the constant of Pointcaré inequality.

Proof The proof follows directly from that of the deterministic case [165, 161]. □

7.1.1 Constrained optimal control problems

The distributed optimal control problems constrained by stochastic elliptic PDEs consist in finding a stochastic optimal distributed control function $g^* \in \mathscr{L}^2(D)$ that minimizes a cost functional $\mathscr{J}(u,g)$ under an elliptic PDE constraint: find $(u^*,g^*) \in \mathscr{U}$ such that

$$\mathscr{J}(u^*,g^*) = \min_{(u,g)\in\mathscr{U}} \mathscr{J}(u,g) \text{ subject to problem (7.5)}, \tag{7.8}$$

where $\mathscr{U}$ is an admissible solution space defined without loss of generality as $\mathscr{U} = \mathscr{H}_0^1(D) \otimes \mathscr{L}^2(D)$, and the quadratic cost functional is defined as [91, 47]

$$\mathscr{J}(u,g) = \mathbb{E}\left[\frac{1}{2}\int_D |u-u_d|^2 dx + \frac{\alpha}{2}\int_D |g|^2 dx\right], \tag{7.9}$$

in which $u_d \in L^2(D)$ is provided as an observation function, e.g., the mean of a sequence of experimental measures, α is a positive regularization parameter.

Theorem 7.1.2 *There exists an optimal solution $(u^*,g^*) \in \mathscr{U}$ to problem (7.8).*

Proof The proof is straightforward by following Lions' argument for deterministic optimal control problems [123]; see also similar proof in [91] for stochastic cases. □

Remark 7.1.1 *When higher moments, e.g., variance, of the observational data u_d or the control function g, or the probability distribution of u_d are incorporated into the cost functional in more general settings as considered in [197], we face essentially nonlinear and fully coupled stochastic problems, which will be addressed in [44].*

7.2 Saddle point formulation

We introduce now the stochastic optimality system in order to derive a saddle point formulation of the optimal control problem (7.8) following the same procedure as in the previous chapter. For the sake of simplicity, we provide the

7.2.1 Stochastic optimality system

Let us first derive the stochastic optimality system to the optimal control problem (7.8) by Lagrangian approach [200]. Define the following stochastic Lagrangian functional associated to problem (7.8) as

$$\mathscr{L}(u,g,p;y) = \mathscr{J}(u,g) + \mathscr{B}(u,p) - \mathscr{F}(p) - \mathscr{G}(p), \tag{7.10}$$

where $p \in \mathscr{H}_0^1(D)$ is named the adjoint variable or Lagrangian parameter [200]. By taking Gâteaux derivative (see [200]) of the Lagrangian functional (7.10) with respect to the variables p,g,u evaluated at q,h,v, we obtain the first order necessary optimality conditions of the stochastic optimal control problem (7.8) - the stochastic optimality system:

$$\begin{cases} \mathscr{B}(u,q) - \mathscr{G}(q) &= \mathscr{F}(q) \qquad \forall q \in \mathscr{H}_0^1(D), \\ (\alpha g - p, h) &= 0 \qquad \forall h \in \mathscr{L}^2(D), \\ \mathscr{B}'(p,v) + (u,v) &= (u_d,v) \qquad \forall v \in \mathscr{H}_0^1(D), \end{cases} \tag{7.11}$$

where $\mathscr{B}'$ is the adjoint bilinear form of $\mathscr{B}$, $\mathscr{B}'(p,v) = \mathscr{B}(v,p)$. As a consequence of Theorem 7.1.2, it has been proven in [99, 91] that there exists an adjoint variable $p^* \in \mathscr{H}_0^1(D)$ associated to the optimal solution (u^*, g^*) such that (u^*, g^*, p^*) is a solution of the stochastic optimal system (7.11). In the following, we will show that (u^*, g^*, p^*) is the unique solution to system (7.11); moreover, (u^*, g^*) is also the unique solution of the stochastic optimal control problem (7.8).

7.2.2 Saddle point formulation

We adopt the same approach of the saddle point formulation as that used in chapter 6 and obtain the following saddle point problem: find $(\underline{u}, p) \in \mathscr{U} \otimes \mathscr{H}_0^1(D)$ such that

$$\begin{cases} \mathscr{A}(\underline{u}, \underline{v}) + \mathscr{B}(\underline{v}, p) &= (\underline{u}_d, \underline{v}) \qquad \forall \underline{v} \in \mathscr{U}, \\ \mathscr{B}(\underline{u}, q) &= \mathscr{F}(q) \qquad \forall q \in \mathscr{H}_0^1(D), \end{cases} \tag{7.12}$$

where the bilinear forms $\mathscr{A}$ and $\mathscr{B}$ are defined as

$$\mathscr{A}(\underline{u}, \underline{v}) := (u, v) + \alpha(g, h) \quad \forall \underline{u}, \underline{v} \in \mathscr{U}, \tag{7.13}$$

being $\underline{u}, \underline{v}$ given by $\underline{u} = (u, g) \in \mathscr{U}$ and $\underline{v} = (v, h) \in \mathscr{U}$; and

$$\mathscr{B}(\underline{u}, q) := \mathscr{B}(u, q) - \mathscr{G}(q) \quad \forall \underline{u} \in \mathscr{U}, q \in \mathscr{H}_0^1(D). \tag{7.14}$$

We can show that this problem is equivalent to the optimal control problem (7.8) and the optimality system (7.11), and admit a unique stochastic optimal solution.

The semi-weak formulation of the saddle point problem is given by: $\forall y \in \Gamma$, find $(\underline{u}(y), p(y)) \in U \otimes V$ (where $U := H_0^1(D) \otimes L^2(D)$ and $V := H_0^1(D)$) such that

$$\begin{cases} A(\underline{u}(y), \underline{v}) + B(\underline{v}, p(y); y) &= (\underline{u}_d, \underline{v}) \qquad \forall \underline{v} \in U, \\ B(\underline{u}(y), q; y) &= F(q; y) \qquad \forall q \in V, \end{cases} \tag{7.15}$$

where the semi-weak bilinear forms A and B and linear functional F are the deterministic counterparts (without taking stochastic integral $\int_\Gamma \cdot \rho(y) dy$) of $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{F}$ defined in (7.13), (7.14) and (7.6), respectively. Note that B depends on y also through the random coefficient $a(y)$ and F through the random force $f(y)$. From the semi-weak formulation, we can derive the analytic regularity of the solution under certain smoothness assumption of the random input data. For the sake of simplicity, we omit the proof of the above results that is similar to the previous chapter. Details are reported in [43].

7.3 Numerical approximation

Thanks to the equivalence of the stochastic optimal control problem (7.8) and its saddle point formulation (7.12), it is sufficient to consider numerical approximation of (7.12) to solve (7.8), which involves both deterministic approximation of the optimal solution in the physical domain D and stochastic approximation in the probability domain Γ. In this section, we present a finite element method with suitable preconditioning techniques [185, 161] for deterministic approximation and use a stochastic collocation method introduced in chapter 1 and chapter 5 for stochastic approximation. In order to alleviate the global computational cost, we propose the model-order reduction strategy empowered by a weighted reduced basis method [49].

7.3.1 Finite element method

Recall the definition of the finite element space X_h^k defined in (6.57) of the previous chapter. Given any $y \in \Gamma$, by applying Galerkin projection of the solution $(\underline{u}(y), p(y))$ in the finite element space $U_h^k \otimes X_h^k \subset U \otimes H_0^1(D)$, where $U_h^k := X_h^k \otimes X_h^k$, we obtain the semi-weak saddle point problem (7.15) in finite element formulation as: find $(\underline{u}_h(y), p_h(y)) \in U_h^k \otimes X_h^k$

$$\begin{cases} A(\underline{u}_h(y), \underline{v}_h) + B(\underline{v}_h, p_h(y); y) &= (\underline{u}_d, \underline{v}_h) \qquad \forall \underline{v}_h \in U_h^k, \\ B(\underline{u}_h(y), q_h; y) &= F(q_h; y) \qquad \forall q_h \in X_h^k. \end{cases} \tag{7.16}$$

The finite element solution of (7.16) is written as

$$u_h(x,y) = \sum_{i=1}^{N_h} u_i(y)\phi_i(x), g_h(x,y) = \sum_{i=1}^{N_h} g_i(y)\phi_i(x), p_h(x,y) = \sum_{i=1}^{N_h} p_i(y)\phi_i(x), \tag{7.17}$$

where $\phi_i, 1 \le i \le N_h$ are the finite element bases in X_h^k, N_h is the number of degrees-of-freedom (d.o.f). The algebraic formulation of (7.16) reads

$$\begin{pmatrix} A_h & B_h^T(y) \\ B_h(y) & 0 \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}}_h(y) \\ \mathbf{p}_h(y) \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{u}}_{dh} \\ \mathbf{f}_h(y) \end{pmatrix}, \tag{7.18}$$

which can be written in a more explicit formulation corresponding to the optimality system (7.11) in the deterministic setting as

$$\begin{pmatrix} M_h & 0 & C_h^T(y) \\ 0 & \alpha M_h & -M_h^T \\ C_h(y) & -M_h & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h(y) \\ \mathbf{g}_h(y) \\ \mathbf{p}_h(y) \end{pmatrix} = \begin{pmatrix} \mathbf{u}_{dh} \\ 0 \\ \mathbf{f}_h(y) \end{pmatrix}. \tag{7.19}$$

Notations have the following meanings: $A_h = (M_h, 0_{N_h \times N_h}; 0_{N_h \times N_h}, \alpha M_h)$ with

$$(M_h)_{i,j} = (\phi_j, \phi_i), \quad (0_{N_h \times N_h})_{i,j} = 0, \quad 1 \le i, j \le N_h, \tag{7.20}$$

$B_h(y) = (C_h(y); -M_h)$ with

$$(C_h(y))_{i,j} = (a(y)\nabla\phi_j, \nabla\phi_i), \quad 1 \le i, j \le N_h, \tag{7.21}$$

the finite element optimal solution is $\underline{\mathbf{u}}_h(y) = (\mathbf{u}_h(y); \mathbf{g}_h(y))$, with

$$\mathbf{u}_h(y) = (u_1(y), \ldots, u_{N_h}(y))^T, \quad \mathbf{g}_h(y) = (g_1(y), \ldots, g_{N_h}(y))^T, \tag{7.22}$$

the adjoint variable $\mathbf{p}_h(y)$ is defined as

$$\mathbf{p}_h(y) = (p_1(y), \ldots, p_{N_h}(y))^T, \tag{7.23}$$

the right hand side $\underline{\mathbf{u}}_{dh} = (\mathbf{u}_{dh}; \mathbf{0}_{N_h}), \mathbf{u}_{dh}, \mathbf{f}_h(y)$ as

$$(\mathbf{u}_{dh})_i = (u_d, \phi_i), \quad (\mathbf{0}_{N_h})_i = 0, \quad \mathbf{f}_h(y) = (f(y), \phi_i), \quad 1 \le i \le N_h. \tag{7.24}$$

When $\alpha \ll 1$ and $N_h \gg 1$, the matrix of the linear system (7.19) is ill-conditioned with very large condition number, leading to computational challenge for direct solve of (7.19). We prefer using GMRES iterations with the following (optimal) preconditioner [185, 164]

$$P = \begin{pmatrix} \hat{M}_h & 0 & 0 \\ 0 & \alpha\hat{M}_h & 0 \\ 0 & 0 & \hat{C}_h(\bar{y}) M_h^{-1} \hat{C}_h^T(\bar{y}) \end{pmatrix}, \tag{7.25}$$

where $\hat{M}_h$ is approximated by symmetric Gauss-Seidel method and $\hat{C}_h$ represents an algebraic multigrid V-cycles approximation for C_h at a reference value $\bar{y} \in \Gamma$ [168].

7.3.2 Weighted reduced basis method

Solving the optimization problem (7.19) is rather expensive when N_h becomes very large and only a few tens or hundreds of complete solves of the system (7.19) may become affordable in practice. Then the stochastic collocation method (even with (anisotropic) sparse-grid structure [149, 148]) can hardly be employed because the number of collocation nodes easily overpasses this computational constraint, especially for high-dimensional problems. The approach that we propose relies on a weighted reduced basis method that has been developed in chapter 2. We tailor this method for the stochastic optimal control problems with the following ingredients corresponding to those in chapter 2. Similar settings for deterministic optimal control problems can be found in [142] and [179] for related applications.

Reduced basis method

For any given choice of $y \in \Gamma$, e.g., the collocation points used by stochastic collocation method, we seek a reduced basis solution $(\underline{u}_r(y), p_r(y)) \in U_{N_r} \otimes X^p_{N_r}$ such that

$$\begin{cases} A(\underline{u}_r(y), \underline{v}_r) + B(\underline{v}_r, p_r(y); y) &= (\underline{u}_d, \underline{v}_r) \qquad \forall \underline{v}_r \in U_{N_r}, \\ B(\underline{u}_r(y), q_r; y) &= F(q_r; y) \qquad \forall q_r \in X^p_{N_r}, \end{cases} \tag{7.26}$$

where the reduced basis space $U_{N_r} = X^e_{N_r} \otimes X^g_{N_r}$ and $X^p_{N_r}$ are constructed from "snapshots" - solutions of (7.19) at some selected samples $y^n, 1 \le n \le N_r$, i.e.,

$$\begin{aligned} X^u_{N_r} &= \text{span}\{u_h(y^n), 1 \le n \le N_r\}, \\ X^g_{N_r} &= \text{span}\{g_h(y^n), 1 \le n \le N_r\}, \\ X^p_{N_r} &= \text{span}\{p_h(y^n), 1 \le n \le N_r\}. \end{aligned} \tag{7.27}$$

Note that in order to guarantee the inf-sup condition for system (7.26), we use an enriched reduced basis space $X^e_{N_r}$ as union of $X^u_{N_r}$ and $X^p_{N_r}$ [144], i.e.,

$$X^e_{N_r} = X^u_{N_r} \cup X^p_{N_r} = \text{span}\{u_h(y^n), p_h(y^n), 1 \le n \le N_r\}. \tag{7.28}$$

For the sake of algebraic stability in assembling the reduced basis matrices and performing Galerking projection [178], we orthonormalize the snapshots in the reduced basis space $X^e_{N_r}$ and $X^g_{N_r}$ by Gram-Schmidt process with respect to the inner-products $(a(\bar{y})\nabla\cdot, \nabla\cdot)$ ($\bar{y}$ being a reference value, e.g., the center of Γ) and $(\cdot,\cdot)$, yielding

$$X^e_{N_r} = \{\zeta^e_n, 1 \le n \le 2N_r\} \text{ and } X^g_{N_r} = \{\zeta^g_n, 1 \le n \le N_r\}. \tag{7.29}$$

Let the reduced basis solution at $y \in \Gamma$ be written as

$$u_r(y) = \sum_{n=1}^{2N_r} u_n(y)\zeta^e_n, \quad g_r(y) = \sum_{n=1}^{N_r} g_n(y)\zeta^g_n, \quad p_r(y) = \sum_{n=1}^{2N_r} p_n(y)\zeta^e_n, \tag{7.30}$$

and the solution coefficient vector at $y \in \Gamma$ as $\mathbf{u}_r(y) = (u_1(y), \ldots, u_{2N_r}(y))^T$, $\mathbf{g}_r(y) = (g_1(y), \ldots, g_{N_r}(y))^T$, $\mathbf{p}_r(y) = (p_1(y), \ldots, p_{2N_r}(y))^T$, we obtain the reduced algebraic optimality system corresponding to the

full algebraic optimality system (7.19) as

$$\begin{pmatrix} M_r & 0 & C_r^T(y) \\ 0 & \alpha D_r & -E_r^T \\ C_r(y) & -E_r & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_r(y) \\ \mathbf{g}_r(y) \\ \mathbf{p}_r(y) \end{pmatrix} = \begin{pmatrix} \mathbf{u}_{dr} \\ 0 \\ \mathbf{f}_r(y) \end{pmatrix}, \tag{7.31}$$

being a $5N_r \times 5N_r$ dense system, where the reduced optimality matrix is defined as

$$\begin{pmatrix} M_r & 0 & C_r^T(y) \\ 0 & \alpha D_r & -E_r^T \\ C_r(y) & -E_r & 0 \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_e^T M_h \mathcal{Z}_e & 0 & \mathcal{Z}_e^T C_h^T(y) \mathcal{Z}_e \\ 0 & \alpha \mathcal{Z}_g^T M_h \mathcal{Z}_g & -\mathcal{Z}_g^T M_h^T \mathcal{Z}_e \\ \mathcal{Z}_e^T C_h(y) \mathcal{Z}_e & -\mathcal{Z}_e^T M_h \mathcal{Z}_g & 0 \end{pmatrix}, \tag{7.32}$$

and the reduced optimal solution and the right hand side are given by

$$\begin{pmatrix} \mathbf{u}_r(y) \\ \mathbf{g}_r(y) \\ \mathbf{p}_r(y) \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_e^T \mathbf{u}_h(y) \\ \mathcal{Z}_g^T \mathbf{g}_h(y) \\ \mathcal{Z}_e^T \mathbf{p}_h(y) \end{pmatrix} \text{ and } \begin{pmatrix} \mathbf{u}_{dr} \\ 0 \\ \mathbf{f}_r(y) \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_e^T \mathbf{u}_{dh} \\ 0 \\ \mathcal{Z}_e^T \mathbf{f}_h(y) \end{pmatrix}, \tag{7.33}$$

where $\mathcal{Z}_e = (\zeta_1^e, \ldots, \zeta_{2N_r}^e)$ and $\mathcal{Z}_g = (\zeta_1^g, \ldots, \zeta_{N_r}^g)$ are column vector matrices.

A weighted greedy algorithm

The efficiency of reduced basis method depends critically on the choice of the samples $y^n, 1 \le n \le N_r$, for which we turn to a weighted greedy algorithm [49]. To start, we randomly choose a realization $y^1 \in \Gamma$ (or use the reference value $\bar{y}$), and solve the full optimality system (7.19) to get the solution $(u_h(y^1), g_h(y^1), p_h(y^1))$. By Gram-Schmidt process, we construct the first reduced basis space X_1^e and X_1^g. For $N_r = 2, \ldots, N_{max}$ (where N_{max} is a prescribed maximum number of reduced bases), we solve the following weighted $L^\infty(\Gamma; X_\rho)$ optimization problem

$$y^{N_r} = \arg\sup_{y \in \Gamma} ||(u_h(y), g_h(y), p_h(y)) - (u_r(y), g_r(y), p_r(y))||_{X_\rho}, \tag{7.34}$$

where X_ρ is a weighted Hilbert space (with weight ρ) equipped with the norm

$$||(v(y), h(y), q(y))||_{X_\rho} = \left\{ \left(||v(y)||_{\bar{X}}^2 + \alpha ||h(y)||_{L^2(D)}^2 + ||q(y)||_{\bar{X}}^2 \right) \rho(y) \right\}^{1/2}, \tag{7.35}$$

where α is the regularization parameter given in the cost functional (7.9), $\rho(y)$ is taken as the joint probability density function evaluated at $y \in \Gamma$ and $||v(y)||_{\bar{X}}^2 = (a(\bar{y})\nabla v(y), \nabla v(y))$ at a reference value $\bar{y} \in \Gamma$. However, solving accurately the infinite-dimensional optimization problem (7.34) is computationally impossible. Instead, we replace Γ by a training set Ξ_{train}, e.g., the collocation nodes used in the stochastic collocation method. Moreover, instead of using the "truth" error defined in (7.34), we consider a cheap, sharp and reliable error bound $\triangle_{N_r}^\rho$ such that

$$||(u_h(y), g_h(y), p_h(y)) - (u_r(y), g_r(y), p_r(y))||_{X_\rho} \le \triangle_{N_r}^\rho (u_r(y), g_r(y), p_r(y)). \tag{7.36}$$

Upon replacement of Γ and the truth error, we have the weighted greedy algorithm

$$y^{N_r} = \arg \sup_{y \in \Xi_{train}} \triangle_{N_r}^\rho (u_r(y), g_r(y), p_r(y)), \quad N_r = 2, \ldots, N_{max}, \tag{7.37}$$

based on which, we can hierarchically build the reduced basis spaces $X_{N_r}^e$ and $X_{N_r}^g$.

Remark 7.3.1 *The weighted norm $||\cdot||_{X_\rho}$ is defined with the joint probability density function ρ in order*

to assign associated weight or importance in choosing the samples for construction of the reduced basis space. In this way the a posteriori error bound and the truth error are kept small, resulting in a more accurate solution, when the probability at the sample is large. This numerical scheme aims to balance the accuracy and importance of the stochastic solution, achieving higher accuracy of statistical moments of interest [49]; see illustration in Section 7.4.1.

A weighted a posteriori error bound

For the purpose of computing a weighted a posteriori error bound $\triangle_{N_r}^{\rho}$, we reformulate the saddle point problem (7.15) as a weakly coercive problem at first [23, 209].

For every $y \in \Gamma$, let $\mathrm{u}(y) := (u(y), g(y), p(y)) \in \mathrm{U} := \bar{X} \otimes L^2(D) \otimes \bar{X} \simeq X_\rho$ and $\mathrm{v} := (v, h, q) \in \mathrm{U}$, we define the bilinear form $\mathrm{B} : \mathrm{U} \otimes \mathrm{U} \to \mathbb{R}$ as

$$\mathrm{B}(\mathrm{u}(y), \mathrm{v}; y) := A(\underline{u}(y), \underline{v}) + B(\underline{v}, p(y); y) + B(\underline{u}(y), q; y), \tag{7.38}$$

and the linear functional $\mathrm{F} : \mathrm{U} \to \mathbb{R}$ as

$$\mathrm{F}(\mathrm{v}; y) := (\underline{u}_d, \underline{v}) + F(q; y). \tag{7.39}$$

Then the saddle point problem (7.15) is equivalent to the following problem: given $y \in \Gamma$, find $\mathrm{u} \in \mathrm{U}$ such that

$$\mathrm{B}(\mathrm{u}(y), \mathrm{v}; y) = \mathrm{F}(\mathrm{v}; y) \quad \mathrm{v} \in \mathrm{U}. \tag{7.40}$$

It can be shown [209] that the bilinear form B is continuous and weakly coercive, i.e.,

$$\gamma(y) := \sup_{\mathrm{v}\in\mathrm{U}} \sup_{\mathrm{u}(y)\in\mathrm{U}} \frac{\mathrm{B}(\mathrm{u}(y), \mathrm{v}; y)}{||\mathrm{u}(\mathrm{y})||_{\mathrm{U}} ||\mathrm{v}||_{\mathrm{U}}} < \infty \text{ and } \beta(y) := \inf_{\mathrm{v}\in\mathrm{U}} \sup_{\mathrm{u}(y)\in\mathrm{U}} \frac{\mathrm{B}(\mathrm{u}(y), \mathrm{v}; y)}{||\mathrm{u}(\mathrm{y})||_{\mathrm{U}} ||\mathrm{v}||_{\mathrm{U}}} > 0, \tag{7.41}$$

where $||\mathrm{v}||_{\mathrm{U}} := ||v||_{\bar{X}} + \sqrt{\alpha}||h||_{L^2(D)} + ||q||_{\bar{X}}$ corresponding to (7.35). Moreover, there exists a unique solution $\mathrm{u}(y) \in \mathrm{U}$ of problem (7.40) satisfying the stability estimate

$$||\mathrm{u}(y)||_{\mathrm{U}} \leq \frac{1}{\beta(y)} ||\mathrm{F}(y)||_{\mathrm{U}'}. \tag{7.42}$$

Consequently, we have similar results (7.41) and (7.42) for the finite element solution $\mathrm{u}_h(y)$ of problem (7.19) with constants $\gamma_h(y), \beta_h(y)$. Let the residual be defined as

$$\mathrm{R}(\mathrm{v}_h; y) = \mathrm{F}(\mathrm{v}_h; y) - \mathrm{B}(\mathrm{u}_r(y), \mathrm{v}_h; y) \quad \forall \mathrm{v}_h \in \mathrm{U}_h := X_h^k \otimes X_h^k \otimes X_h^k, \tag{7.43}$$

then we have that the error between the finite element solution and the reduced basis solution $\mathrm{e}(y) = (u_h(y), g_h(y), p_h(y)) - (u_r(y), g_r(y), p_r(y))$ satisfies

$$\mathrm{B}(\mathrm{e}(y), \mathrm{v}_h; y) = \mathrm{R}(\mathrm{v}_h; y) \quad \mathrm{v}_h \in \mathrm{U}_h. \tag{7.44}$$

which yields, by the stability estimate (7.42), that

$$||\mathrm{e}(y)||_{\mathrm{U}_h} \leq \frac{1}{\beta_h(y)} ||\mathrm{R}(\mathrm{v}_h; y)||_{\mathrm{U}_h'}. \tag{7.45}$$

Therefore, we can define the a weighted posteriori error bound as: for $\forall\, y \in \Xi_{train}$

$$\triangle_{N_r}^{\rho}(\mathrm{u}_r(y)) := \frac{\sqrt{\rho(y)}}{\beta_{LB}(y)} ||\mathrm{R}(y)||_{\mathrm{U}'} \geq \sqrt{\rho(y)} ||\mathrm{e}(y)||_{\mathrm{U}} = ||\mathrm{e}(y)||_{X_\rho}, \tag{7.46}$$

where a lower bound $\beta_{LB}(y) \le \beta_h(y), \forall\, y \in \Xi_{train}$ can be evaluated by a cheap successive constraint method properly extended to a Babuška inf-sup condition [178, 177]. As for evaluation of the weighted residual norm $||\mathrm{R}(y)||_{X_\rho}$, we turn to an efficient offline-online decomposition procedure.

Remark 7.3.2 *In the definition of the compound Hilbert space* U*, we use the Hilbert space $\bar{X}$ equipped with norm $||\cdot||_{\bar{X}} = (a(\bar{y})\nabla\cdot, \nabla\cdot)$ for both the state variable $u(y)$ and the adjoint variable $p(y)$ in order to obtain good stability of the inf-sup constant $\beta_h(y), \forall\, y \in \Gamma$. In fact, when $a(y)$ is not far from the reference value $a(\bar{y})$, the inf-sup constant $\beta_h(y)$ is also close to $\beta_h(\bar{y})$, which enables us to use a uniformly lower bound $\beta_{LB} \le \beta_h(y), \forall\, y \in \Gamma$, for the sake of computational efficiency [49].*

Offline-online decomposition

The offline-online decomposition procedure decomposes the reduced basis method into the expensive offline construction stage and cheap online evaluation stage. More explicitly, we build the reduced basis space $X^e_{N_r}$ and $X^g_{N_r}$, assemble and store all matrices in (7.32) and the right hand side vector (7.33) in an offline stage. In particular, the quantities in (7.32) and (7.33) that depend on the random variable $y \in \Gamma$ are assembled as

$$\mathcal{Z}_e^T C_h(y) \mathcal{Z}_e = \sum_{k=0}^{K} y_k \mathcal{Z}_e^T C_h^k \mathcal{Z}_e \text{ and } \mathcal{Z}_e^T \mathbf{f}_h(y) = \sum_{k=0}^{K} y_k \mathcal{Z}_e^T \mathbf{f}_h^k, \tag{7.47}$$

where $\mathcal{Z}_e^T C_h^k \mathcal{Z}_e$ and $\mathcal{Z}_e^T \mathbf{f}_h^k, 0 \le k \le K$ are assembled offline with the matrices $(C_h^k)_{i,j} = (a_k \nabla\phi_j, \nabla\phi_i), 0 \le k \le K, 1 \le i, j \le N_h$ and the vectors $\mathbf{f}_h^k = (f_k, \phi_i), 0 \le k \le K, 1 \le i \le N_h$. Recall that $y_0 = 1$ and $a_k, f_k, 0 \le k \le K$ are defined in (7.4). For a more compact notation, we define

$$\mathrm{B}_r^0 = \begin{pmatrix} \mathcal{Z}_e^T M_h \mathcal{Z}_e & 0 & \mathcal{Z}_e^T (C_h^0)^T \mathcal{Z}_e \\ 0 & \alpha \mathcal{Z}_g^T M_h \mathcal{Z}_g & -\mathcal{Z}_g^T M_h^T \mathcal{Z}_e \\ \mathcal{Z}_e^T C_h^0 \mathcal{Z}_e & -\mathcal{Z}_e^T M_h \mathcal{Z}_g & 0 \end{pmatrix}, \quad \mathrm{F}_r^0 = \begin{pmatrix} \mathcal{Z}_e^T \mathbf{u}_{dh} \\ 0 \\ \mathcal{Z}_e^T \mathbf{f}_h^0 \end{pmatrix}, \tag{7.48}$$

and

$$\mathrm{B}_r^k = \begin{pmatrix} 0 & 0 & \mathcal{Z}_e^T (C_h^k)^T \mathcal{Z}_e \\ 0 & 0 & 0 \\ \mathcal{Z}_e^T C_h^k \mathcal{Z}_e & 0 & 0 \end{pmatrix}, \quad \mathrm{F}_r^k = \begin{pmatrix} 0 \\ 0 \\ \mathcal{Z}_e^T \mathbf{f}_h^k \end{pmatrix}, \quad 1 \le k \le K. \tag{7.49}$$

Then the reduced algebraic optimality system (7.31) can be written as

$$\sum_{k=0}^{K} y_k \mathrm{B}_r^k \mathbf{u}_r^c(y) = \sum_{k=0}^{K} y_k \mathrm{F}_r^k, \tag{7.50}$$

where $\mathbf{u}_r^c(y) = (\mathbf{u}_r(y); \mathbf{g}_r(y); \mathbf{p}_r(y))$ is the coefficient of reduced basis solution at $y \in \Gamma$. A direct solver, e.g., by Gauss elimination, can be applied to solve the reduced basis optimality system (7.50) with complexity $O((5N_r)^3)$, since $N_r \ll N_h$ in practice.

From the definition of the residual (7.43), we have by Riesz representation theorem [165] that there exists a unique element $\hat{\mathrm{e}}(y) \in \mathrm{U}_h$ such that

$$(\hat{\mathrm{e}}(y), \mathrm{v}_h)_{\mathrm{U}_h} = \mathrm{R}(\mathrm{v}_h; y) \quad \forall \mathrm{v}_h \in \mathrm{U}_h. \tag{7.51}$$

Therefore, we have $||\mathrm{R}(y)||_{\mathrm{U}_h'} = ||\hat{\mathrm{e}}(y)||_{\mathrm{U}_h}$, to evaluate which we make the following definition of bilinear form and linear function corresponding to (7.38) and (7.39):

$$\mathrm{B}^0(\mathrm{u}(y), \mathrm{v}) = A(\underline{u}(y), \underline{v}) + B^0(\underline{v}, p(y)) + B^0(\underline{u}(y), q), \mathrm{F}^0(\mathrm{v}) = (\underline{u}_d, \underline{v}) + F^0(q), \tag{7.52}$$

and

$$\mathrm{B}^k(\mathrm{u}(y),\mathrm{v}) = B^k(\underline{v},p(y)) + B^k(\underline{u}(y),q), \mathrm{F}^k(\mathrm{v}) = F^k(q), \quad 1 \le k \le K, \tag{7.53}$$

where $\forall(\underline{v},q) \in \mathrm{U}$, we have $B^0(\underline{v},q) = (a_0\nabla v, \nabla q) - (h,q)$, $F^0(q) = (f_0,q)$, and $B^k(\underline{v},q) = (a_k\nabla v,\nabla q)$, $F^k(q) = (f_k,q), 1 \le k \le K$. By the above definition, we obtain by Riesz representation theorem that there exist $\mathrm{f}_n, \mathrm{b}_k^n$ such that $(\mathrm{f}_k,\mathrm{v}) = \mathrm{F}^k(\mathrm{v})$ and $(\mathrm{b}_k^n,\mathrm{v}) = -\mathrm{B}^k(\zeta_n^c,\mathrm{v})$ for $\forall \mathrm{v} \in \mathrm{U}, 0 \le k \le K, 1 \le n \le 5N_r$, where $\zeta_n^c = (\zeta_n^e,0,0), 1 \le n \le 2N_r$, $\zeta_n^c = (0,\zeta_{n-2N_r}^g,0), 2N_r+1 \le n \le 3N_r$, $\zeta_n^c = (0,0,\zeta_{n-3N_r}^e), 3N_r+1 \le n \le 5N_r$ and the compound reduced basis space $\mathrm{U}_r = X_{N_r}^e \otimes X_{N_r}^g \otimes X_{N_r}^e$. To this end, we have by the definition of the residual (7.43)

$$\begin{aligned} ||\hat{\mathrm{e}}(y)||_{\mathrm{U}_h}^2 = &\sum_{k=0}^{K}\sum_{k'=0}^{K} y_k(\mathrm{f}_k,\mathrm{f}_{k'})y_{k'} + 2\sum_{k=0}^{K}\sum_{k'=0}^{N}\sum_{n=1}^{5N_r} y_k(\mathrm{f}_k,\mathrm{b}_{k'}^n)(\mathbf{u}_r)_k y_{k'} \\ &+ \sum_{k=0}^{K}\sum_{k'=0}^{K}\sum_{n=1}^{5N_r}\sum_{n'=1}^{5N_r} y_k(\mathbf{u}_r)_n(\mathrm{b}_k^n,\mathrm{b}_{k'}^{n'})(\mathbf{u}_r)_{n'}y_{k'}, \end{aligned} \tag{7.54}$$

where all the quantities of inner-product are computed and stored in the offline stage, and only $O((K+1)^2 \times (5N_r)^2)$ operations, being K and $N_r \ll N_h$ very small, are needed for online evaluation of the a posterior error bound $\triangle_{N_r}^{\rho}$ defined in (7.46).

7.4 Numerical tests

In this section, we carry out several numerical tests to illustrate the computational efficiency and numerical accuracy of the weighted reduced basis method compared to the non-weighted reduced basis method and stochastic collocation method with tensor product grid, isotropic and anisotropic sparse grid. Theoretical error estimates obtained in the last section are verified by three examples with different dimensions, ranging from one dimension to moderate dimension $(1-10)$ and to high dimension $(10-100)$, and with different probability distributions.

7.4.1 One-dimensional problems

The first example focuses on the demonstration of the convergence property of the weighted reduced basis method compared to other methods with probability density functions of distinct shape. The physical domain is specified as $D = (0,1)^2$ with a uniform mesh of 712 vertices, over which we construct finite element space for spatial discretization by continuous piecewise linear polynomials. We set $f = 1$ and the coefficient a of problem (7.1) as

$$a(x,y) = \frac{1}{10}(1.1 + \sin(2\pi x_1)y), \tag{7.55}$$

where $x = (x_1,x_2) \in D$ and the random variable $y \sim \mathrm{Beta}(\mu_1,\mu_2)$ obeys beta distribution supported on $\Gamma = [-1,1]$ with two shape parameters $\mu_1,\mu_2 \in \mathbb{N}_+$. The probability density function of y is displayed in Fig. 7.1 when (μ_1,μ_2) take values of $(1,1)$, $(10,10)$ and $(100,100)$, featuring very different shapes with distinct weight. The observation data u_d is set as the solution of (7.1) at the reference value $\bar{y} = 0$ and control $g(x_1,x_2) = sin(\pi x_1)sin(\pi x_2)$. We define the worst case scenario error as

$$\max_{1\le m\le M_{test}} (||\underline{u}(y^m) - \underline{u}_N(y^m)||_U + ||p(y^m) - p_N(y^m)||_{\bar{X}}), \tag{7.56}$$

where $y^m, 1 \le m \le M_{test}$ are testing samples randomly drawn according to its probability density function, $(\underline{u},p)$ is the finite element solution and $(\underline{u}_N,p_N)$ is the solution by (weighted) reduced basis method or stochastic collocation method with N bases or collocation nodes. The expectation error is

defined in a posteriori way as

$$\left| ||E_l[\underline{u}]||_U^2 - ||E_L[\underline{u}]||_U^2 \right| + \left| ||E_l[p]||_{\tilde{X}}^2 - E_L[p]||_{\tilde{X}}^2 \right| \tag{7.57}$$

for ease of computation, where $l, 1 \leq l \leq L-1$ represents the level of approximation by quadrature formula. We apply the weighted reduced basis method and reduced basis method with M_{train} training samples drawn according to the probability distribution, and also stochastic collocation method based on Gauss-Jacobi quadrature nodes to solve the stochastic optimal control problem (7.8) with regularization parameter $\alpha = 1$. The convergence results are shown in the following few figures Fig. 7.1 - 7.4, for which we have used $M_{train} = 100$ training samples and $M_{test} = 100$ test samples.

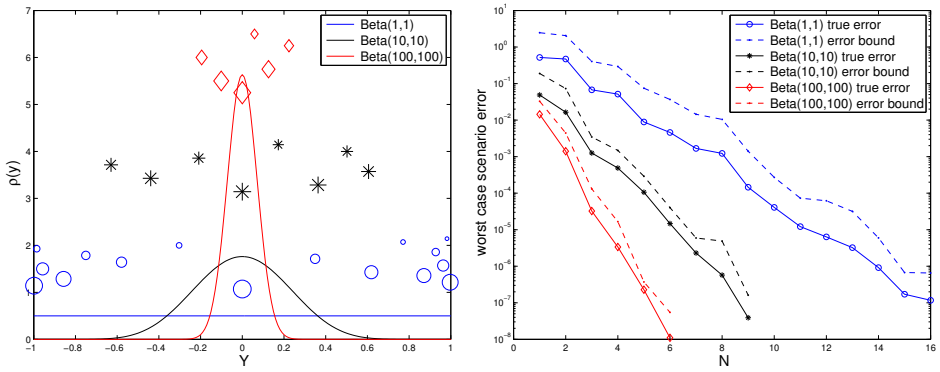

Figure 7.1: Left: Probability density function of Beta(μ_1, μ_2) distribution with different (μ_1, μ_2) and samples selected by weighted reduced basis approximation in order, the bigger the size the earlier it has been selected; Right: convergence result of the true error and error bound by wRBM.

On the left of Figure 7.1, the samples selected by weighted reduced basis method are plotted in sequential order, where the larger the markers are, the earlier the samples have been selected. The right of Figure 7.1 shows the convergence of the true error (error between approximation and true value) and the error bound $\triangle_N$ defined in (7.46) in three different settings. From Figure 7.1 we can see that the most important samples (or samples with large probability) can be efficiently selected by the weighted reduced basis method, leading to less samples (thus less bases in the reduced basis space) for the more concentrated probability distribution.

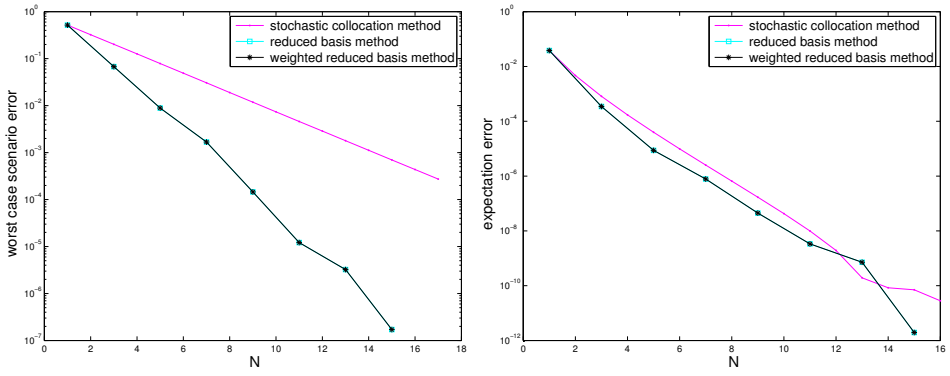

Figure 7.2: Comparison of worst case scenario error (left) and expectation error (right) by (weighted) reduced basis method and stochastic collocation method with $(\mu_1, \mu_2) = (1, 1)$.

When $(\mu_1, \mu_2) = (1, 1)$, the beta distribution becomes a uniform distribution with probability density

function $\rho = 1/2$, in which case the weighted reduced basis method is the same as reduced basis method, as we can see from their convergence results in Figure 7.3, from which we can also observe that the reduced basis method converges faster than stochastic collocation method for worst case scenario error, while for the expectation error they display quite close convergence rates.

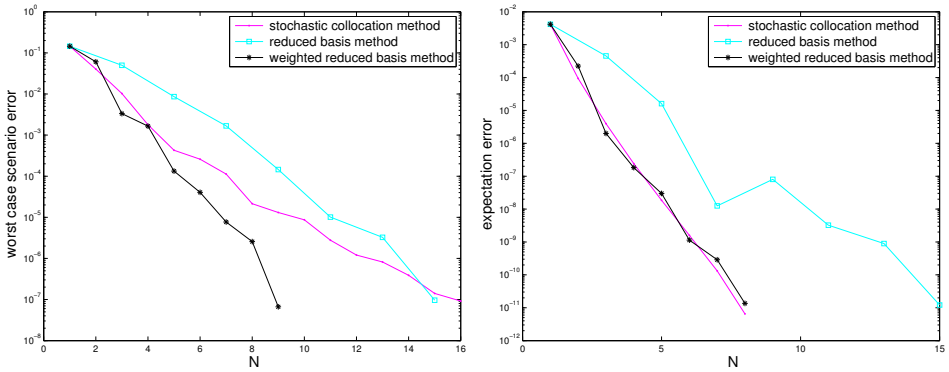

Figure 7.3: Comparison of worst case scenario error (left) and expectation error (right) by (weighted) reduced basis method and stochastic collocation method with $(\mu_1, \mu_2) = (10, 10)$.

For $(\mu_1, \mu_2) = (10, 10)$, the weighted reduced basis method performs evidently better than the reduced basis method measured in both errors, and converges faster than stochastic collocation method as for worst case scenario error and comparable in expectation error (note that here the Gauss-Jacobi quadrature formula is optimal for evaluation of expectation), which demonstrates that the weighted reduced basis method works efficiently for evaluation of statistical moments of the solution. This conclusion has been further illustrated by the convergence results displayed in Figure 7.4 for the test with $(\mu_1, \mu_2) = (100, 100)$. However, we remark that the computation for both offline construction and online evaluation by the (weighted) reduced basis method is more expensive than that by stochastic collocation method in one-dimensional problems; see [50] for more detailed comparison of computational cost between reduced basis method and stochastic collocation method.

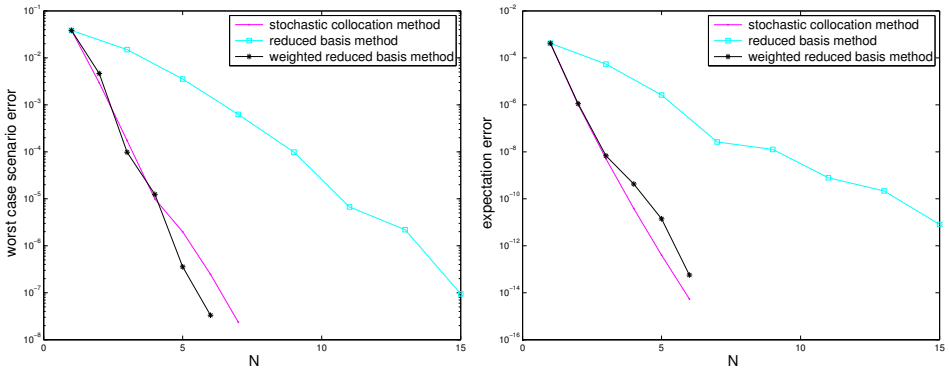

Figure 7.4: Comparison of worst case scenario error (left) and expectation error (right) by (weighted) reduced basis method and stochastic collocation method with $(\mu_1, \mu_2) = (100, 100)$.

7.4.2 Moderate-dimensional problems

The example presented in this section is devoted to demonstrate the computational efficiency and numerical accuracy of the weighted reduced basis method. We define a general random field g as a truncation of Karhunen–Loève expansion (see (27) of the preliminary chapter) of a Gaussian random field with correction length L [149]

$$g(x_i, y) = \mathbb{E}[g] + \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1 + \sum_{k=1}^{K} \sqrt{\lambda_k}\left(\sin(k\pi x_i) y_{2k} + \cos(k\pi x_i) y_{2k+1}\right), \tag{7.58}$$

where the random variables $y_k, 1 \le k \le 2K+1$ follow standard normal distribution, the eigenvalues $\lambda_1 = 0.4782, \lambda_2 = 0.0752, \lambda_3 = 0.0034$, accounting for around 99.5% uncertainties of the random field truncated with 7 random variables. In order to guarantee assumption (7.2), we cut off the random variables $|y_k| \le 3, 1 \le k \le K$ (with tail probability less than 0.5%) and set $\mathbb{E}[g] = 8$. For simplicity, we do not consider the cut-off error and the truncation error. We set $a = g(x_1, y)/10$, $f = g(x_2, y)$, $\alpha = 1$ and the observation data u_d as the solution of (7.1) at the reference value $\bar{y}_k = 0, 1 \le k \le 2K+1$ and control function $g(x, y) = sin(\pi x_1) sin(\pi x_2)$.

To test the finite element error, we set $y = \bar{y}$, $h = 1/4, 1/8, 1/16, 1/32, 1/64$ and use the optimal solution at $h = 1/64$ as the "true" value. Figure 7.5 (left) displays the linear and quadratic decay of finite element error $\mathscr{E}_h(\bar{y})$ with $\mathbb{P}_1$ and $\mathbb{P}_2$ elements. The right of Figure 7.5 depicts the reduced basis error $\mathscr{E}_r$ and the error bound $\triangle_{N_r}$ of the optimal solution (in fact, we take the worst case scenario error at 100 test samples), from which we can see that the cheap error bound is rather sharp and accurate, decaying exponentially fast with respect to the number of reduced bases. We remark that the error bound depends on the lower bound of the inf-sup constant $\beta_{LB}(y), y \in \Gamma$ in (7.46), which falls inside [0.5, 1] in the training set $y \in \Xi_{train}$ with 1000 samples. For the sake of computational efficiency, we can take a uniform lower bound $\beta_{LB} = 0.5$ for any new $y \in \Gamma$.

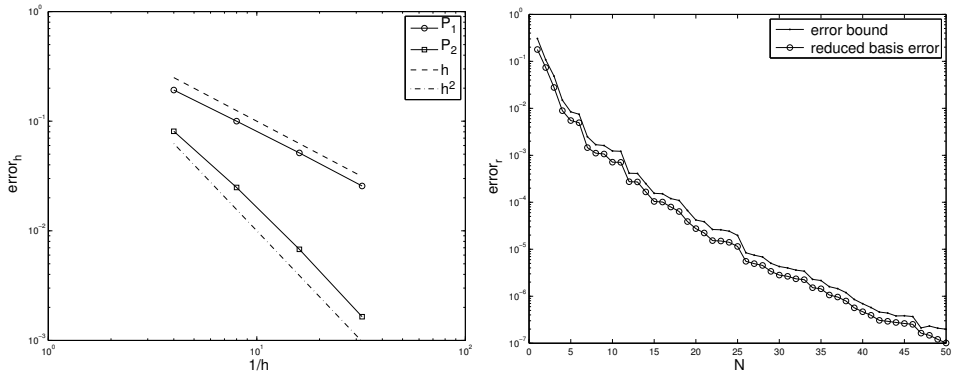

Figure 7.5: Left: finite element error of $\mathbb{P}_1$ and $\mathbb{P}_2$; right: reduced basis error and error bound.

Figure 7.6 reports the comparison of stochastic approximation errors between the weighted reduced basis method (RBM) and the stochastic collocation method (SCM) with Gauss-Hermite collocation nodes in both tensor-product and sparse-grid settings. The convergence comparison measured by worst case scenario error is depicted on the left of Figure 7.6, which shows that the reduced basis approximation converges much faster than the stochastic collocation approximation, with error reaching 10^{-7} with only 50 bases (thus 50 solve of the full optimality system (7.19)), while it requires $78079 \approx 1562 \times 50$ collocation nodes (thus 78079 solve) for sparse-grid setting to attain the same error although it converges faster than the tensor-product setting.

Alternative to the necessity of cut-off, we may assume log-normal structure [149] of the random field and apply weighted empirical interpolation method [48] to obtain an affine decomposition (31).

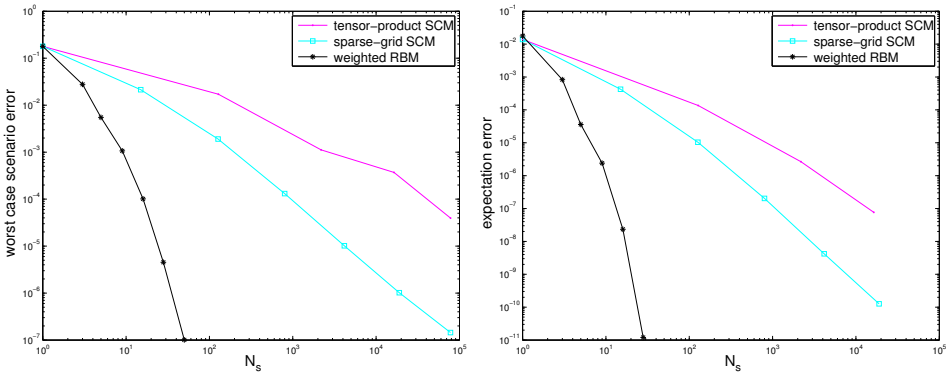

Figure 7.6: Comparison between weighted reduced basis method and stochastic collocation method (tensor-product and sparse-grid) for worst case scenario error (left) and error of expectation (right).

For evaluation of the expectation error, we use the expectation of the optimal solution by the sparse-grid stochastic collocation method at the deepest level ($q - K = 6$) as a "true" value. The weighted reduced basis expectation is evaluated via quadrature formula at the deepest level of sparse-grid with the optimal solution computed by online reduced basis procedure at all the collocation nodes. We can see from the right of Figure 7.6 that only 28 bases or solve are needed for weighted reduced basis method to obtain a more accurate expectation than the stochastic collocation method (with 18943 solve in sparse-grid setting and 16384 by tensor-product setting). Thanks to the cheap online evaluation, the weighted reduced basis method is much more efficient than the stochastic collocation method to evaluate the statistics of the solution, especially when a solve of the full optimality system is very expensive.

7.4.3 High-dimensional problems

In this section, we show that the weighted reduced basis method (wRBM) can be effectively applied to solve high-dimensional problems and its combination with the adaptive sparse grid stochastic collocation method (aSCM) developed in chapter 5 provides an efficient way to evaluate statistical moments of the solution.

We assume that the random coefficient $a = g(x_1, y)/10$ with g defined in (7.58) where $L = 1/128$, which features a slow decay of the eigenvalues ($\lambda_1 = 0.0138, \lambda_{50} = 0.0095$). Moreover, we assume that the random variables $y_k, 1 \le k \le 2K+1$ follow uniform distribution with zero mean and unit variance, supported on $[-\sqrt{3}, \sqrt{3}]$. We set $f = 10$ and $\mathbb{E}[a] = 20$ that satisfy Assumption 7.1 and u_d as for moderate-dimensional problems in the last section with $g(x_1, x_2) = sin(\pi x_1)sin(\pi x_2)$. We apply a dimensional-adaptive algorithm (see [79] for details) with maximum number of collocation nodes specified as $10^1, 10^2, 10^3, 10^4, 10^5$ to construct the adaptive sparse-grid stochastic collocation approximation. The weighted reduced basis approximation is constructed with 1000 training samples and tested with 100 test samples. The convergence results are depicted in Figure 7.7 for 11, 31 and 101 dimensional problems. In the reduced basis construction, only 30 bases have been used to achieve more accurate approximation (measured in worst case scenario approximation error) than the adaptive sparse grid stochastic collocation method with 10^5 collocation nodes, requiring 10^5 full solve of the optimality systems.

As for evaluation of the expectation by weighted reduced basis method, we first compute the reduced

Instead of using a fixed number of training samples, we can choose adaptively the collocation nodes on the sparse grid as the training samples or use an adaptive greedy algorithm [212].

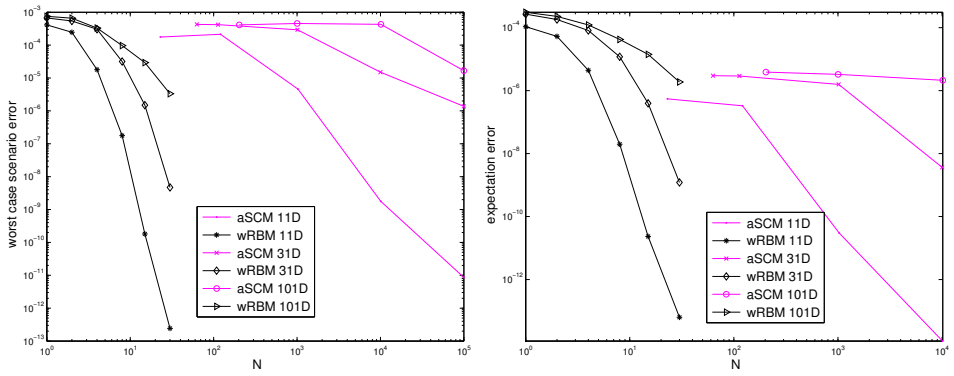

Figure 7.7: Worst case scenario approximation error (left) and expectation error (right) for different stochastic dimensions $N = 11, 31, 101$ of adaptive SCM and weighted RBM.

basis solution at the collocation nodes in the adaptive sparse-grid in the deepest level (with 10^5 collocation nodes) and then compute the expectation by Clenshaw–Curtis quadrature formula [149] on the sparse grid. From the right of Figure 7.7, we can see that 30 reduced bases are sufficient for the weighted reduced basis method to achieve comparable accuracy as the stochastic collocation method.

7.5 Summary

In this chapter, we studied stochastic optimal control problems with elliptic PDE constraint and developed and analyzed an efficient computational method to solve them. An analysis of existence, uniqueness and stochastic regularity of the optimal solution was carried out by virtue of a saddle point formulation of the optimal control problems. In numerical approximation of stochastic optimality system, we applied finite element method with proper preconditioning techniques in the deterministic space and stochastic collocation method in the stochastic space. In order to alleviate the computational effort, we proposed a model order reduction approach based on a weighted reduced basis method. A global error analysis of our computation method was conducted thanks to the stochastic regularity result of the optimal solution. Numerical tests have illustrated the efficiency and accuracy of the computational method proposed in this chapter.

Promising research opportunities arise from several computational challenges: firstly, when the dimension of the stochastic space becomes very high (in the order of 1000 or even more) with many effective dimensions, efficient (quasi, multilevel, etc.) Monte-Carlo methods should be adopted together with the model order reduction approach in order to harness the computational burden; secondly, when the cost functional is more general, e.g., accounting for high moments or probability distribution of measurements, the optimality system becomes nonlinear and coupled in stochastic space, for which appropriate extension of our computation method is needed; at third, generalization and application of the method in stochastic optimal control problems with more complex constraints, e.g., time-dependent and nonlinear problems, are ongoing.

8 Stochastic optimal control problem constrained by Stokes equations

In this chapter, we study a stochastic optimal control problem constrained by Stokes equations with random inputs and distributed control function, which shares all the computational challenges of the stochastic elliptic optimal control problem in the last chapter, but features the additional difficulty arising from the saddle point structure of the underlying Stokes model [161, 22]. To tackle these challenges, we develop a multilevel and weighted reduced basis method, using multilevel greedy algorithm and weighted a posteriori error estimate. More in detail, (anisotropic) sparse grid stochastic collocation method is applied for stochastic approximation of the optimal solution in the probability space and finite element method with (optimal) preconditioning techniques is used for deterministic approximation in physical space, leading to a large number of finite element optimality systems to solve. Then we project the finite element optimality system into an adaptively constructed reduced basis space, leading to a reduced optimality system that can be solved with very cheap computational cost. For the construction of the reduced basis space, we design a multilevel greedy algorithm and propose a weighted a posteriori error bound, which produces quasi-optimal “snapshots" space that well approximate the low-dimensional manifold of the quantities of interest. A global error analysis is carried out for the complete numerical approximation based on the regularity of the optimal solution, in particular the stochastic regularity obtained for the specific Stokes control problem. Numerical experiments with stochastic dimensions ranging from 10 to 100 are performed to verify the error convergence results and demonstrate the efficiency and accuracy of our computational method for large scale and high-dimensional PDEs-constrained optimization problems. The main contribution of this work is the development of efficient model order reduction techniques to solve stochastic optimal control problems with PDEs (Stokes equations) constraints. For a deterministic setting of Stokes optimal control problems in a “double" inner-outer saddle point formulation, see [143] and [142]. Another contribution of this chapter is the detailed analysis of the stochastic regularity of the optimal solution with respect to input random variables and the associated error convergence analysis for fluid control problems with Stokes constraint. Our numerical experiments demonstrate that the proposed method achieves considerable computational saving: for large-scale and “reducible" problems, it is definitely cheaper than both the stochastic collocation method [172] and Galerkin projection method [99] that have been recently developed for solving stochastic optimal control problems.

This chapter is organized as follows. The stochastic optimal control problem with Stokes constraint is presented in section 8.1 with certainty assumptions on the random input data; section 8.2 is devoted to prove the well-posedness of the stochastic optimal solution, followed by section 8.3 for the study

---

Reference for this chapter:

P. Chen, A. Quarteroni and G. Rozza. Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations. Submitted, 2013.

of stochastic regularity; detailed numerical approximation of the problem is presented in section 8.4, which provides the basis for the development of the multilevel and weighted reduced basis method in section 8.5; in section 8.6, global error estimates are carried out and verified by numerical experiments in section 8.7; summary is provided in the last section 8.8.

8.1 Problem statement

8.1.1 Stochastic Stokes equations

Recall the definitions in the preliminary chapter. We consider the following stochastic Stokes equations: given a random variable $\nu : \Omega \to \mathbb{R}_+$, two random vector fields $\mathbf{f} : D \times \Omega \to \mathbb{R}^d$ and $\mathbf{h} : \partial D_N \times \Omega \to \mathbb{R}^d$, find a solution $\{\mathbf{u}, p\} : D \times \Omega \to \mathbb{R}^d \times \mathbb{R}$ such that the following equations hold almost surely (for almost every $\omega \in \Omega$)

$$\begin{cases} -\nu(\omega)\triangle\mathbf{u}(\cdot,\omega) + \nabla p(\cdot,\omega) = \mathbf{f}(\cdot,\omega) & \text{in } D, \\ \nabla\cdot\mathbf{u}(\cdot,\omega) = 0 & \text{in } D, \\ \mathbf{u}(\cdot,\omega) = \mathbf{0} & \text{on } \partial D_D, \\ \nu(\omega)\nabla u(\cdot,\omega)\cdot\mathbf{n} - p(\cdot,\omega)\mathbf{n} = \mathbf{h}(\cdot,\omega) & \text{on } \partial D_N, \end{cases} \tag{8.1}$$

where ∂D_D and ∂D_N represent the Dirichlet and Neumann boundaries such that $\partial D_D \cup \partial D_N = \partial D$ and $\partial D_D \cap \partial D_N = \emptyset$. In particular, we consider a homogeneous Dirichlet boundary condition and a nonhomogeneous Neumann boundary condition.

At any realization $\omega \in \Omega$, the Stokes equations (8.1) are commonly used to quantify the velocity $\mathbf{u}$ and pressure p of fluid flow where advective inertial forces are negligible compared to viscous forces measured via the kinematic viscosity parameter ν. This occurs, e.g., for low speed channel flows, the flow of viscous polymers or micro-organisms [5]. In practice, the viscosity ν may vary in a large extent rather than stay as a fixed constant for many fluids depending on the temperature, the multicomponent property of the fluid and some other factors [75]. Quantification of the body force $\mathbf{f}$ and boundary condition $\mathbf{h}$, for instance by experimental measurements, may also be faced with various noises or uncertainties. Incorporation of these different uncertainties leads to the study of stochastic Stokes equations.

In order to solve (8.1) in the distribution sense, we write its weak formulation as: find $\{\mathbf{u}, p\} \in \mathcal{V} \times \mathcal{Q}$ such that

$$\begin{cases} a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = (\mathbf{f},\mathbf{v}) + (\mathbf{h},\mathbf{v})_{\partial D_N} & \forall \mathbf{v} \in \mathcal{V}, \\ b(\mathbf{u},q) = 0 & \forall q \in \mathcal{Q}, \end{cases} \tag{8.2}$$

where $\mathcal{V} := \left\{\mathbf{v} \in \mathcal{H}^{1,d}(D) : \mathbf{v} = \mathbf{0} \text{ on } \partial D_D\right\}$, $\mathcal{Q} := L^2(\Omega) \otimes Q(D)$, with $Q(D)$ defined as

$$Q(D) := \left\{ q \in L^2(D) : \int_D q dx = 0 \right\}. \tag{8.3}$$

The bilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is defined as

$$a(\mathbf{w},\mathbf{v}) := \int_\Omega \int_D \nu \nabla\mathbf{w} \otimes \nabla\mathbf{v} dx dP(\omega) = \sum_{i,j=1}^{d} \int_\Omega \int_D \nu \frac{\partial w_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} dx dP(\Omega) \tag{8.4}$$

and the bilinear form $b(\cdot,\cdot) : \mathcal{V} \times \mathcal{Q} \to \mathbb{R}$ reads

$$b(\mathbf{v},q) = -\int_\Omega \int_D \nabla\cdot\mathbf{v} q dx dP(\omega) = -\sum_{i=1}^{d} \int_\Omega \int_D \frac{\partial v_i}{\partial x_i} q dx dP(\omega). \tag{8.5}$$

The stochastic inner product $(\mathbf{f},\mathbf{v})$ and $(\mathbf{h},\mathbf{v})_{\partial D_N}$ are defined on the domain D and Neumann boundary

∂D_N, respectively. Similar to the elliptic case in Assumption 7.1, we make the following assumption for the random input data in order to guarantee the well-posedness of saddle problem (8.2).

Assumption 8.1 *The random viscosity ν is positive and uniformly bounded from below and from above, i.e., there exist two constants $0 < \nu_{\min} \le \nu_{max} < \infty$ such that*

$$P(\omega : \nu_{min} \le \nu(\omega) \le \nu_{max}) = 1. \tag{8.6}$$

The random force field $\mathbf{f}$ and Neumann boundary field $\mathbf{h}$ satisfy

$$||\mathbf{f}||_{\mathscr{L}} < \infty \text{ and } ||\mathbf{h}||_{\mathscr{H}} < \infty, \tag{8.7}$$

where we denote $\mathscr{L} = \mathscr{L}^{2,d}(D)$ and $\mathscr{H} = \mathscr{L}^{2,d}(\partial D_N)$ for simplicity.

The well-posedness of the stochastic Stokes problem (8.2) can be obtained by the following theorem, whose proof follows the same lines of the Brezzi theorem (see Theorem 6.1.5) for the deterministic setting and will thus be omitted here; see [27, 161, 165] for details.

Theorem 8.1.1 *Under Assumption 8.1, there exists a unique solution to the stochastic Stokes problem (8.2). Moreover, the following stability estimate holds*

$$||\mathbf{u}||_{\mathscr{V}} \le \frac{1}{\alpha_a}\left(C_P||\mathbf{f}||_{\mathscr{L}} + \frac{\alpha_a+\gamma_a}{\beta_b}C_T||\mathbf{h}||_{\mathscr{H}}\right), \tag{8.8}$$

and

$$||p||_{\mathscr{Q}} \le \frac{1}{\beta_b}\left(\left(1+\frac{\gamma_a}{\alpha_a}\right)C_P||\mathbf{f}||_{\mathscr{L}} + \frac{\gamma_a(\alpha_a+\gamma_a)}{\alpha_a\beta_b}C_T||\mathbf{h}||_{\mathscr{H}}\right), \tag{8.9}$$

where the positive constants $\alpha_a, \gamma_a, \beta_b, \gamma_b$ are defined such that

$$a(\mathbf{w},\mathbf{v}) \le \gamma_a||\mathbf{w}||_{\mathscr{V}}||\mathbf{v}||_{\mathscr{V}} \quad \forall \mathbf{w},\mathbf{v} \in \mathscr{V} \tag{8.10}$$

and

$$a(\mathbf{v},\mathbf{v}) \ge \alpha_a||\mathbf{v}||^2_{\mathscr{V}} \quad \forall \mathbf{v} \in \mathscr{V}_0, \tag{8.11}$$

being $\mathscr{V}_0 := \{\mathbf{v} \in \mathscr{V} : b(\mathbf{v},q) = 0, \forall q \in \mathscr{Q}\}$ the kernel of b, and

$$\inf_{q\in\mathscr{Q}}\sup_{\mathbf{v}\in\mathscr{V}}\frac{b(\mathbf{v},q)}{||\mathbf{v}||_{\mathscr{V}}||q||_{\mathscr{Q}}} \ge \beta_b, \tag{8.12}$$

where β_b is called an inf-sup constant or compatibility constant, and the continuity

$$b(\mathbf{v},q) \le \gamma_b||\mathbf{v}||_{\mathscr{V}}||q||_{\mathscr{Q}} \quad \forall \mathbf{v} \in \mathscr{V}, \forall q \in \mathscr{Q}. \tag{8.13}$$

The constants C_P and C_T are those of the Poincaré inequality and trace theorem [161],

$$||\mathbf{v}||_{\mathscr{L}} \le C_P||\mathbf{v}||_{\mathscr{V}} \text{ and } ||\mathbf{v}||_{\mathscr{H}} \le C_T||\mathbf{v}||_{\mathscr{V}} \quad \forall \mathbf{v} \in \mathscr{V}. \tag{8.14}$$

8.1.2 Finite dimensional assumption

We employ the finite dimensional noise assumption 0.1 made in the preliminary chapter for the random input data. More explicitly, we provide the following examples for ν and $\mathbf{h}$.

Example 1. For a multicomponent fluid flow, the viscosity is propositional to the contribution of each

component [106], which can be described by

$$\nu(Y(\omega)) = \sum_{n=1}^{N} \nu_n Y_n(\omega) + \nu_0 \left(1 - \sum_{n=1}^{N} Y_n(\omega)\right) = \nu_0 + \sum_{n=1}^{N} (\nu_n - \nu_0) Y_n(\omega), \tag{8.15}$$

where $Y_n, 1 \le n \le N$ are uniformly distributed in $[0, 1/N]$ and $\nu_n > 0, 0 \le n \le N$.

Example 2. The random vector field $\mathbf{h}$ is given by the truncated Karhunen–Loève expansion (recall the KL expansion in (34)):

$$\mathbf{h}(x, Y(\omega)) = \mathbb{E}[\mathbf{h}](x) + \sum_{n=1}^{N} \sqrt{\lambda_n} \mathbf{h}_n(x) Y_n(\omega) \quad x \in \partial D_N, \tag{8.16}$$

where $(\lambda_n, \mathbf{h}_n)$ are the eigenpairs of a continuous and bounded covariance function.

Under this assumption, the stochastic Stokes equations (8.1) can be viewed as a set of parameterized equations defined in a tensor product of the spatial domain and the parameter space $D \times \Gamma$. We remark that the Hilbert space $L^2(\Omega)$ is equivalent to $L^2_\rho(\Gamma)$ and we use the same notation $\mathscr{L}, \mathscr{H}, \mathscr{V}, \mathscr{Q}$ for the stochastic Hilbert spaces. Moreover, Theorem 8.1.1 holds under this assumption.

8.1.3 Constrained optimal control problem

We study a distributed optimal control problem constrained by the stochastic Stokes equations. Let us define the cost functional

$$\begin{aligned} \mathscr{J}(\mathbf{u}, p, \mathbf{f}) &= \frac{1}{2} ||\mathbf{u} - \mathbf{u}_d||^2_{\mathscr{L}} + \frac{1}{2} ||p - p_d||^2_{\mathscr{L}^2(D)} + \frac{\alpha}{2} ||\mathbf{f}||^2_{\mathscr{G}} \\ &= \mathbb{E}\left[\frac{1}{2} \int_D (\mathbf{u} - \mathbf{u}_d)^2 dx + \frac{1}{2} \int_D (p - p_d)^2 dx + \frac{\alpha}{2} \int_D \mathbf{f}^2 dx \right], \end{aligned} \tag{8.17}$$

where the first two terms measure the discrepancy between the solution $\{\mathbf{u}, p\} \in \mathscr{V} \times \mathscr{Q}$ of the stochastic Stokes problem (8.2) and the observational data $\{\mathbf{u}_d, p_d\} \in L^{2,d}(D) \times Q(D)$ that represent the mean of measurements. The admissible control space $\mathscr{G}$ in the last term is a non-empty, closed, bounded and convex subset of $\mathscr{L}^{2,d}(D)$. This term is used to regularize in mathematical sense the control function $\mathbf{f}$ with a regularization parameter $\alpha > 0$, which can also be viewed as a penalization of the control energy. The optimal control problem constrained by the stochastic Stokes problem (8.2) can be formulated as: find an optimal solution $\{\mathbf{u}^*, p^*, \mathbf{f}^*\}$ such that

$$\mathscr{J}(\mathbf{u}^*, p^*, \mathbf{f}^*) = \min \left\{ \mathscr{J}(\mathbf{u}, p, \mathbf{f}) : \{\mathbf{u}, p, \mathbf{f}\} \in \mathscr{V} \times \mathscr{Q} \times \mathscr{G} \text{ and solve (8.2)} \right\}. \tag{8.18}$$

It is easy to see that the cost functional $\mathscr{J}$ is weakly lower semicontinuous in $\mathscr{G}$, i.e.

$$\lim_{n \to \infty} \inf \mathscr{J}(\mathbf{u}, p, \mathbf{f}_n) \ge \mathscr{J}(\mathbf{u}, p, \mathbf{f}) \quad \forall \{\mathbf{u}, p\} \in \mathscr{V} \times \mathscr{Q} \tag{8.19}$$

for any sequence $\{\mathbf{f}_n\}_{n=1}^{\infty} \in \mathscr{G}$ such that $\mathbf{f}_n \rightharpoonup \mathbf{f}$ as $n \to \infty$. Then, we have the following result by Lions' argument [123]:

Theorem 8.1.2 *Under Assumption 8.1 and the finite dimensional noise assumption, there exists an optimal solution $\{\mathbf{u}^*, p^*, \mathbf{f}^*\} \in \mathscr{V} \times \mathscr{Q} \times \mathscr{G}$ to the stochastic optimal control problem (8.18).*

Remark 8.1.1 *In the cost functional, we have used the $\mathscr{L}^{2,d}(D)$ norm for measuring the discrepancy between the velocity field and its mean value of measurements. Extension to the case with $\mathscr{V}$ norm is straightforward by requiring that the data $\{\mathbf{u}_d, p_d\}$ possess higher regularity in the spatial domain.*

Another extension to stochastic data $\{\mathbf{u}_d, p_d\}$ can be handled in the same way as in this work, provided they depend explicitly on a finite dimensional random vector, i.e., $\{\mathbf{u}_d, p_d\}(\cdot,\omega) = \{\mathbf{u}_d, p_d\}(\cdot, Y(\omega))$.

Remark 8.1.2 *When the higher moments of the observational data $\{\mathbf{u}_d, p_d\}$ or the control function $\mathbf{f}$, e.g., variance, skewness, etc., or the probability distribution of $\{\mathbf{u}_d, p_d\}$ are incorporated into the cost functional in more general settings [197], we face essentially nonlinear and fully coupled problems, which will be addressed in [44].*

8.2 Saddle point formulation

In order to prove the uniqueness of the optimal solution of the constrained optimal control problem (8.18), we turn to a saddle point formulation and establish its equivalence to the optimality system obtained by Lagrangian variational approach in solving (8.18).

8.2.1 Optimality system

We first employ the variational approach [200] to derive an optimality system (known as Karush–Kuhn–Tucker (KKT) conditions) in solving the constrained optimal control problem (8.18). Define a compound bilinear form $\mathscr{B} : (\mathscr{V}\times\mathscr{Q}\times\mathscr{G})\times(\mathscr{V}\times\mathscr{Q}) \to \mathbb{R}$ to represent the weak formulation of the stochastic Stokes equations (8.2) as

$$\mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\}) = a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) + b(\mathbf{u}, q) - (\mathbf{f}, \mathbf{v}). \tag{8.20}$$

Associated with this bilinear form, we define the Lagrangian functional

$$\mathscr{L}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{u}^a, p^a\}) = \mathscr{J}(\mathbf{u}, p, \mathbf{f}) + \mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{u}^a, p^a\}) - (\mathbf{h}, \mathbf{u}^a)_{\partial D_N}, \tag{8.21}$$

where $\{\mathbf{u}^a, p^a\} \in \mathscr{V}\times\mathscr{Q}$ are the adjoint (or dual) variables of the Stokes equations (8.2) corresponding to the state (or primal) variables $\{\mathbf{u}, p\}$. The Lagrangian functional (8.21) is Gâteaux differentiable with respect to $\{\mathbf{u}, p, \mathbf{f}, \mathbf{u}^a, p^a\}$ [200], so that we can take Gâteaux derivative of (8.21) with respect to the state variable $\{\mathbf{u}, p\}$ in test directions $\{\mathbf{v}^a, q^a\}$, control variable $\mathbf{f}$ in $\mathbf{g}$, and adjoint variable $\{\mathbf{u}^a, p^a\}$ in $\{\mathbf{v}, q\}$, respectively, obtaining the first order optimality system as

$$\begin{cases} (\{\mathbf{u}, p\}, \{\mathbf{v}^a, q^a\}) & + & \mathscr{B}(\{\mathbf{v}^a, q^a, \mathbf{0}\}, \{\mathbf{u}^a, p^a\}) & & \\ & & = (\mathbf{u}_d, \mathbf{v}^a) + (p_d, p^a) & \forall \{\mathbf{v}^a, q^a\} \in \mathscr{V}\times\mathscr{Q}, \\ \alpha(\mathbf{f}, \mathbf{g}) - (\mathbf{u}^a, \mathbf{g}) & & = 0 & \forall \mathbf{g} \in \mathscr{G}, \\ \mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\}) & & = (\mathbf{h}, \mathbf{v})_{\partial D_N} & \forall \{\mathbf{v}, q\} \in \mathscr{V}\times\mathscr{Q}, \end{cases} \tag{8.22}$$

where we can identify the third equation with the state equation (8.2), the first one with the adjoint equation and the second one with the optimality equation. More explicitly, the optimality system can be rewritten as

$$\left\{\begin{array}{cccc:cccc} (\mathbf{u}, \mathbf{v}^a) & & & & +a(\mathbf{u}^a, \mathbf{v}^a) & +b(\mathbf{v}^a, p^a) & = (\mathbf{u}_d, \mathbf{v}^a) & \forall \mathbf{v}^a \in \mathscr{V}, \\ & (p, q^a) & & & +b(\mathbf{u}^a, \mathbf{q}^a) & & = (p_d, q^a) & \forall q^a \in \mathscr{Q}, \\ & & \alpha(\mathbf{f}, \mathbf{g}) & & -(\mathbf{u}^a, \mathbf{g}) & & = 0 & \forall \mathbf{g} \in \mathscr{G}, \\ \hdashline a(\mathbf{u}, \mathbf{v}) & +b(\mathbf{v}, p) & -(\mathbf{f}, \mathbf{v}) & & & & = (\mathbf{h}, \mathbf{v})_{\partial D_N} & \forall \mathbf{v} \in \mathscr{V}, \\ b(\mathbf{u}, q) & & & & & & = 0 & \forall q \in \mathscr{Q}, \end{array}\right. \tag{8.23}$$

whose saddle point structure can be more easily appreciated. For the sake of numerical approximation, let us introduce the following operators corresponding to system (8.23):

$$\begin{array}{lllllll}
M_v: & \mathcal{V}\to\mathcal{V} & \text{such that} & (M_v\mathbf{u},\mathbf{v}) & = & (\mathbf{u},\mathbf{v}) & \forall\mathbf{u},\mathbf{v}\in\mathcal{V},\\
M_q: & \mathcal{Q}\to\mathcal{Q} & \text{such that} & (M_q p,q) & = & (p,q) & \forall p,q\in\mathcal{Q},\\
M_g: & \mathcal{G}\to\mathcal{G} & \text{such that} & (M_g\mathbf{f},\mathbf{g}) & = & (\mathbf{v},\mathbf{g}) & \forall\mathbf{f}\in\mathcal{G},\mathbf{g}\in\mathcal{G},\\
M_c: & \mathcal{G}\to\mathcal{V} & \text{such that} & (M_c\mathbf{g},\mathbf{v}) & = & (\mathbf{g},\mathbf{v}) & \forall\mathbf{g}\in\mathcal{G},\mathbf{v}\in\mathcal{V},\\
M_n: & \mathcal{H}\to\mathcal{V} & \text{such that} & (M_n\mathbf{h},\mathbf{v}) & = & (\mathbf{h},\mathbf{v})_{\partial D_N} & \forall\mathbf{h}\in\mathcal{H},\mathbf{v}\in\mathcal{V},\\
A: & \mathcal{V}\to\mathcal{V} & \text{such that} & a(\mathbf{u},\mathbf{v}) & = & (\nu\nabla\mathbf{u},\mathbf{v}) & \forall\mathbf{u},\mathbf{v}\in\mathcal{V},\\
B: & \mathcal{V}\to\mathcal{Q} & \text{such that} & b(\mathbf{v},q) & = & -(\nabla\cdot\mathbf{v},q) & \forall\mathbf{v}\in\mathcal{V},q\in\mathcal{Q},
\end{array} \tag{8.24}$$

from which we obtain the following saddle point linear optimality system as

$$\left(\begin{array}{ccc:cc}
M_v & 0 & 0 & A & B^T\\
0 & M_p & 0 & B & 0\\
0 & 0 & \alpha M_g & -M_c^T & 0\\
\hdashline
A & B^T & -M_c & 0 & 0\\
B & 0 & 0 & 0 & 0
\end{array}\right)\left(\begin{array}{c}\mathbf{u}\\ p\\ \mathbf{f}\\ \mathbf{u}^a\\ p^a\end{array}\right)=\left(\begin{array}{c}M_v\mathbf{u}_d\\ M_p p_d\\ 0\\ M_n\mathbf{h}\\ 0\end{array}\right). \tag{8.25}$$

Remark 8.2.1 *The optimality system (8.25) can be regarded as the first order necessary condition to guarantee existence of a solution to the optimal control problem (8.18). However, the uniqueness of the optimal solution is not an immediate result.*

8.2.2 Saddle point formulation

In order to obtain the uniqueness and study the stochastic regularity (Sec. 8.3) of the optimal solution, we introduce a compound saddle point formulation of the constrained optimal control problem (8.18).

Let $\mathscr{A}:(\mathcal{V}\times\mathcal{Q}\times\mathcal{G})\times(\mathcal{V}\times\mathcal{Q}\times\mathcal{G})\to\mathbb{R}$ be a compound bilinear form defined as

$$\mathscr{A}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{v},q,\mathbf{g}\})=(\mathbf{u},\mathbf{v})+(p,q)+\alpha(\mathbf{f},\mathbf{g}), \tag{8.26}$$

then we have that the cost functional (8.17) can be expressed as

$$\mathscr{J}(\mathbf{u},p,\mathbf{f})=\frac{1}{2}\mathscr{A}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{u},p,\mathbf{f}\})-\mathscr{A}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{u}_d,p_d,\mathbf{0}\})+C, \tag{8.27}$$

where C is the constant $\mathscr{A}(\{\mathbf{u}_d,p_d,\mathbf{0}\},\{\mathbf{u}_d,p_d,\mathbf{0}\})/2$. Then the following proposition establishes the equivalence between the constrained optimal control problem (8.18) and the saddle point problem (8.33), whose proof follows the one in the deterministic setting; see [27, 23] for details.

Proposition 8.2.1 *Suppose that the bilinear form $\mathscr{A}$ is symmetric, non-negative and continuous, i.e., there exists a constant $\gamma>0$ such that $\forall\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{v},q,\mathbf{g}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}$, we have*

$$|\mathscr{A}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{v},q,\mathbf{g}\})|\le\gamma||\{\mathbf{u},p,\mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}||\{\mathbf{v},q,\mathbf{g}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}. \tag{8.28}$$

Moreover, suppose that $\mathscr{A}$ is strongly coercive in the kernel space of $\mathscr{B}$, defined as

$$\mathcal{K}:=\big\{\{\mathbf{u},p,\mathbf{f}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}:\quad\mathscr{B}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{v},q\})=0\quad\forall\{\mathbf{v},q\}\in\mathcal{V}\times\mathcal{Q}\big\}, \tag{8.29}$$

i.e., there exists a constant $\epsilon>0$ such that $\forall\{\mathbf{v},q,\mathbf{g}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}$, we have

$$\mathscr{A}(\{\mathbf{v},q,\mathbf{g}\},\{\mathbf{v},q,\mathbf{g}\})\ge\epsilon||\{\mathbf{v},q,\mathbf{g}\}||^2_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}. \tag{8.30}$$

Suppose that $\mathscr{B}$ is continuous, i.e., there exists a constant $\delta > 0$ such that $\forall \{\mathbf{u}, p, \mathbf{f}\} \in \mathcal{V} \times \mathcal{Q} \times \mathcal{G}, \{\mathbf{v}, q\} \in \mathcal{V} \times \mathcal{Q}$, we have

$$|\mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\})| \leq \delta ||\{\mathbf{u}, p, \mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} ||\{\mathbf{v}, q\}||_{\mathcal{V}\times\mathcal{Q}}. \tag{8.31}$$

Furthermore, suppose that $\mathscr{B}$ satisfies the inf-sup condition, i.e., there exists a constant $\beta > 0$ such that

$$\inf_{\{\mathbf{v},q\}\in\mathcal{V}\times\mathcal{Q}} \sup_{\{\mathbf{u},p,\mathbf{f}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} \frac{\mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\})}{||\{\mathbf{u}, p, \mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} ||\{\mathbf{v}, q\}||_{\mathcal{V}\times\mathcal{Q}}} \geq \beta. \tag{8.32}$$

Then the constrained optimal control problem (8.18) is equivalent to the following saddle point problem: find $\{\mathbf{u}, p, \mathbf{f}\} \in \mathcal{V} \times \mathcal{Q} \times \mathcal{G}$ and $\{\mathbf{u}^a, p^a\} \in \mathcal{V} \times \mathcal{Q}$ such that

$$\begin{cases} \mathscr{A}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}^a, q^a, \mathbf{g}\}) + \mathscr{B}(\{\mathbf{v}^a, q^a, \mathbf{g}\}, \{\mathbf{u}^a, p^a\}) \\ = (\{\mathbf{u}_d, p_d, \mathbf{0}\}, \{\mathbf{v}^a, q^a, \mathbf{g}\}) \quad \forall \{\mathbf{v}^a, q^a, \mathbf{g}\} \in \mathcal{V} \times \mathcal{Q} \times \mathcal{G}, \\ \mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\}) = (\mathbf{h}, \mathbf{v})_{\partial D_N} \quad \forall \{\mathbf{v}, q\} \in \mathcal{V} \times \mathcal{Q}. \end{cases} \tag{8.33}$$

Remark 8.2.2 *By establishing the equivalence between the optimality system (8.22) and the saddle point system (8.33), it can be shown that the variables $\{\mathbf{u}^a, p^a\}$ (and $\{\mathbf{v}^a, q^a, \mathbf{g}\}$) used in the saddle point formulation (8.33) are coincident with the adjoint variables $\{\mathbf{u}^a, p^a\}$ (and test variables $\{\mathbf{v}^a, q^a, \mathbf{g}\}$) as introduced in the Lagrangian functional (8.21). Moreover, we highlight that some mathematical properties such as stochastic regularity (Sec. 8.3) of the two systems hold the same.*

8.2.3 Equivalence, uniqueness and stability estimates

Lemma 8.2.2 *The constrained optimal control problem (8.18), the saddle point problem (8.33) and the first order optimality system (8.22) are equivalent problems.*

Proof To prove the equivalence between the first two problems, we only need to verify the assumptions in Proposition 8.2.1. By the definition (8.26), it is easy to check that $\mathscr{A}$ is symmetric and non-negative; $\mathscr{A}$ is also continuous

$$\begin{aligned} |\mathscr{A}(\{\mathbf{u}, p, \mathbf{f}\},\{\mathbf{v}, q, \mathbf{g}\})| &\leq ||\mathbf{u}||_{\mathcal{V}} ||\mathbf{v}||_{\mathcal{V}} + ||p||_{\mathcal{Q}} ||q||_{\mathcal{Q}} + \alpha ||\mathbf{f}||_{\mathcal{G}} ||\mathbf{g}||_{\mathcal{G}} \\ &\leq \gamma ||\{\mathbf{u}, p, \mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} ||\{\mathbf{v}, q, \mathbf{g}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}, \forall \{\mathbf{v}, q, \mathbf{g}\} \in \mathcal{V} \times \mathcal{Q} \times \mathcal{G}, \end{aligned} \tag{8.34}$$

where the continuity constant is $\gamma = 1$ and $||\{\mathbf{v}, q, \mathbf{g}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} := ||\mathbf{v}||_{\mathcal{V}} + ||q||_{\mathcal{Q}} + \sqrt{\alpha}||\mathbf{g}||_{\mathcal{G}}$. For any $\{\mathbf{v}, q, \mathbf{g}\} \in \mathcal{K}$, the kernel of $\mathscr{B}$ defined in (8.29), we have by Theorem 8.1.1 that $||\mathbf{v}||_{\mathcal{V}} \leq C_P ||\mathbf{g}||_{\mathcal{G}} / \alpha_a$, which yields

$$\begin{aligned} \mathscr{A}(\{\mathbf{v}, q, \mathbf{g}\}, \{\mathbf{v}, q, \mathbf{g}\}) &= ||\mathbf{v}||^2_{\mathcal{L}} + ||q||^2_{\mathcal{Q}} + \alpha ||\mathbf{g}||^2_{\mathcal{G}} \\ &\geq \frac{\alpha_a^2 \alpha}{2C_P^2} ||\mathbf{v}||^2_{\mathcal{V}} + ||q||^2_{\mathcal{Q}} + \frac{\alpha}{2} ||\mathbf{g}||^2_{\mathcal{G}} \\ &\geq \frac{1}{3} \min\left\{ \frac{\alpha_a^2 \alpha}{2C_P^2}, \frac{1}{2} \right\} ||\{\mathbf{v}, q, \mathbf{g}\}||^2_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}, \end{aligned} \tag{8.35}$$

from which we can infer that $\mathscr{A}$ is coercive on $\mathcal{K}$ with a coercivity constant $\epsilon = (1/3)\min\{\alpha_a^2 \alpha/(2C_P^2), 1/2\}$. As for the continuity of the bilinear form $\mathscr{B}$ defined in (8.20), by Assumption 8.1 and Theorem 8.1.1 we have for any $\{\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\}\} \in (\mathcal{V} \times \mathcal{Q} \times \mathcal{G}) \times (\mathcal{V} \times \mathcal{Q})$,

$$\begin{aligned} |\mathscr{B}(\{\mathbf{u}, p, \mathbf{f}\}, \{\mathbf{v}, q\})| &\leq \nu_{max} ||\mathbf{u}||_{\mathcal{V}} ||\mathbf{v}||_{\mathcal{V}} + \gamma_b ||\mathbf{v}||_{\mathcal{V}} ||p||_{\mathcal{Q}} + \gamma_b ||\mathbf{u}||_{\mathcal{V}} ||q||_{\mathcal{Q}} + ||\mathbf{f}||_{\mathcal{G}} ||\mathbf{v}||_{\mathcal{V}} \\ &\leq \max\{\nu_{max}, \gamma_b, 1/\sqrt{\alpha}\} ||\{\mathbf{u}, p, \mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} ||\{\mathbf{v}, q\}||_{\mathcal{V}\times\mathcal{Q}}, \end{aligned} \tag{8.36}$$

the continuity constant being $\delta = \max\{\nu_{max}, \gamma_b, 1/\sqrt{\alpha}\}$. Finally, $\mathscr{B}$ satisfies the inf-sup condition, as

$$\begin{aligned} &\inf_{\{\mathbf{v},q\}\in\mathcal{V}\times\mathcal{Q}} \sup_{\{\mathbf{u},p,\mathbf{f}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} \frac{\mathscr{B}(\{\mathbf{u},p,\mathbf{f}\},\{\mathbf{v},q\})}{||\{\mathbf{u},p,\mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}||\{\mathbf{v},q\}||_{\mathcal{V}\times\mathcal{Q}}} \\ \geq &\inf_{\{\mathbf{v},q\}\in\mathcal{V}\times\mathcal{Q}} \sup_{\{\mathbf{u},p,\mathbf{0}\}\in\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} \frac{\mathscr{B}(\{\mathbf{u},p,\mathbf{0}\},\{\mathbf{v},q\})}{||\{\mathbf{u},p,\mathbf{0}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}}||\{\mathbf{v},q\}||_{\mathcal{V}\times\mathcal{Q}}} \geq \beta. \end{aligned} \tag{8.37}$$

The inf-sup constant $\beta > 0$ depends on $\alpha_a, \gamma_a, \beta_b$ as follows (see [209]):

$$\beta = \frac{1}{k_{12} + \max\{k_{11}, k_{22}\}}, \tag{8.38}$$

where

$$k_{11} = \alpha_a^{-2}(1 + \beta_b^{-2}\gamma_a^2), \quad k_{22} = \beta_b^{-2}\gamma_a^2 k_{11} + \beta_b^{-2} \text{ and } k_{12} = \beta_b^{-1}\gamma_a k_{11}. \tag{8.39}$$

We conclude that solving the constrained optimal control problem (8.18) is equivalent to solve the saddle point problem (8.33), thanks to Proposition 8.2.1. The equivalence between the optimality system (8.22) and the saddle point system (8.33) can be observed by noticing that by adding the second equation (optimal equation) of (8.22) to its first one (adjoint equation), we obtain the first equation of (8.33). □

Thanks to Lemma 8.2.2 and using Theorem 8.1.1, we can conclude that the optimal solution is unique and satisfies the a priori (boundedness) estimate:

Theorem 8.2.3 *There exists a unique optimal solution to the constrained optimal control problem (8.18). Moreover, the optimal solution* $\{\mathbf{u}, p, \mathbf{f}\}$ *and the adjoint variables* $\{\mathbf{u}^a, p^a\}$ *satisfy the following stability estimates:*

$$||\{\mathbf{u}, p, \mathbf{f}\}||_{\mathcal{V}\times\mathcal{Q}\times\mathcal{G}} \leq \alpha_1||\{\mathbf{u}_d, p_d\}||_{\mathcal{L}\times\mathcal{Q}} + \beta_1||\mathbf{h}||_{\mathcal{H}} \tag{8.40}$$

and

$$||\{\mathbf{u}^a, p^a\}||_{\mathcal{V}\times\mathcal{Q}} \leq \alpha_2||\{\mathbf{u}_d, p_d\}||_{\mathcal{L}\times\mathcal{Q}} + \beta_2||\mathbf{h}||_{\mathcal{H}} \tag{8.41}$$

where the constants $\alpha_1, \beta_1, \alpha_2, \beta_2$ *are defined as*

$$\alpha_1 = \frac{1}{\epsilon}\max\{C_P, 1\}, \quad \beta_1 = \frac{1}{\epsilon}\frac{\epsilon+\gamma}{\beta}C_T, \tag{8.42}$$

and

$$\alpha_2 = \frac{1}{\beta}\left(1 + \frac{\gamma}{\epsilon}\right)\max\{C_P, 1\}, \quad \beta_2 = \frac{1}{\beta}\frac{\gamma(\epsilon+\gamma)}{\epsilon\beta}C_T, \tag{8.43}$$

with the constants ϵ, γ, β *defined in Proposition 8.2.1.*

8.3 Stochastic regularity

In this section, we show that under suitable assumptions for the regularity of the viscosity $\nu : \Gamma \to \mathbb{R}_+$ and boundary data $\mathbf{h} : \Gamma \to H$ in the stochastic space Γ, the solution $\{\mathbf{u}, p, \mathbf{f}, \mathbf{u}^a, p^a\} : \Gamma \to V \times Q \times G \times V \times Q$ can be analytically extended to a complex region that covers the stochastic space Γ. (Here and in the following, we denote L, V, Q, G, H as the deterministic Hilbert space corresponding to their stochastic counterparts $\mathscr{L}, \mathcal{V}, \mathcal{Q}, \mathcal{G}, \mathcal{H}$, e.g., $H = L^{2,d}(\partial D_N)$.)

Let $\mathbf{k} = (k_1, \ldots, k_N) \in \mathbb{N}_0^N$ be a N-dimensional multi-index of non-negative integers, and denote $\mathbf{k}! = \prod_{i_1}^{k_1} i_1 \cdots \prod_{i_N}^{k_N} i_N$, $|\mathbf{k}| = \sum_{n=1}^N k_n$, and $|\mathbf{k}|! = \prod_{i=1}^{|\mathbf{k}|} i$; let $\partial_y^{\mathbf{k}}\{\cdot\} = \partial_{y_1}^{k_1}\partial_{y_2}^{k_2}\cdots\partial_{y_N}^{k_N}\{\cdot\}$ represent the $\mathbf{k}$-th order partial derivative with respect to the parameter $y = (y_1, \ldots, y_N)$. Let us also define the following

constants for ease of notation

$$C_\alpha = \alpha_1 + \alpha_2, C_\beta = \beta_1 + \beta_2, C_{\alpha,\beta} = \max\{\alpha_1 + \alpha_2, \beta_1 + \beta_2\}, \tag{8.44}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are the stability constants defined in the Brezzi theorem (8.2.3).

We make the following assumption of stochastic regularity on the input data:

Assumption 8.2 *For every $y \in \Gamma$, there exists a N-dimensional positive rate vector $\mathbf{r} = (r_1, \dots, r_N) \in \mathbb{R}_+^N$ such that the $\mathbf{k}$-th order derivative of the viscosity $\nu : \Gamma \to \mathbb{R}_+$ and the boundary condition $\mathbf{h} : \Gamma \to H$ satisfy*

$$C_{\alpha,\beta} \frac{|\partial_y^{\mathbf{k}} \nu(y)|}{\nu(\bar{y})} \leq \mathbf{r}^{\mathbf{k}} \text{ and } \frac{C_\beta ||\partial_y^{\mathbf{k}} \mathbf{h}(y)||_H}{C_\alpha ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + C_\beta ||\mathbf{h}(y)||_H} \leq |\mathbf{k}|! \mathbf{r}^{\mathbf{k}}. \tag{8.45}$$

Theorem 8.3.1 *Under assumption 8.2, we have the following a priori estimate for the $\mathbf{k}$-th order derivative of the solution $\{\mathbf{u}, p, \mathbf{f}, \mathbf{u}^a, p^a\} : \Gamma \to V \times Q \times G \times V \times Q$*

$$\begin{aligned} &||\partial_y^{\mathbf{k}} \{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} + ||\partial_y^{\mathbf{k}} \{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \\ &\leq C(C_\alpha ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + C_\beta ||\mathbf{h}(y)||_H) |\mathbf{k}|! (r\mathbf{r})^{\mathbf{k}}, \end{aligned} \tag{8.46}$$

where $r\mathbf{r} = (r r_1, r r_2, \dots, r r_N)$ with the constant rate $r > 1/\log(2)$, and the constant C is independent of $\mathbf{k}$, which will be provided explicitly in the proof.

Proof The semi-weak formulation of the saddle point problem (8.33) reads: find $\{\mathbf{u}(y), p(y), \mathbf{f}(y)\} \in V \times Q \times G$ and $\{\mathbf{u}^a(y), p^a(y)\} \in V \times Q$ such that

$$\begin{cases} \mathscr{A}\left(\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}, \{\mathbf{v}^a, q^a, \mathbf{g}\}\right) + \mathscr{B}(\{\mathbf{v}^a, q^a, \mathbf{g}\}, \{\mathbf{u}^a(y), p^a(y)\}; y) \\ \qquad = (\mathbf{u}_d, \mathbf{v}^a) + (p_d, q^a) \quad \forall \{\mathbf{v}^a, q^a, \mathbf{g}\} \in V \times Q \times G, \\ \mathscr{B}(\{\mathbf{u}(y), p(y), \mathbf{g}(y)\}, \{\mathbf{v}, q\}; y) = (\mathbf{h}(y), \mathbf{v})_{\partial D_N} \quad \forall \{\mathbf{v}, q\} \in V \times Q, \end{cases} \tag{8.47}$$

where we have used the same bilinear forms $\mathscr{A}$ and $\mathscr{B}$ for ease of notation, which can be identified in the semi-weak sense by their explicit dependence on the parameter y. Taking $\mathbf{k}$-th ($|\mathbf{k}| > 0$) order partial derivative of problem (8.47) with respect to the parameter y, we obtain the following problem thanks to the general Leibniz rule: find $\partial_y^{\mathbf{k}} \{\mathbf{u}(y), p(y), \mathbf{f}(y)\} \in V \times Q \times G$ and $\partial_y^{\mathbf{k}} \{\mathbf{u}^a(y), p^a(y)\} \in V \times Q$ such that

$$\begin{cases} \mathscr{A}\left(\partial_y^{\mathbf{k}} \{\mathbf{u}(y), p(y), \mathbf{f}(y)\}, \{\mathbf{v}^a, q^a, \mathbf{g}\}\right) + \mathscr{B}(\{\mathbf{v}^a, q^a, \mathbf{g}\}, \partial_y^{\mathbf{k}} \{\mathbf{u}^a(y), p^a(y)\}; y) \\ \qquad = -\sum_{\mathbf{k}' \in \Lambda(\mathbf{k})} (\partial_y^{\mathbf{k}-\mathbf{k}'} \nu(y) \nabla \partial_y^{\mathbf{k}'} \mathbf{u}^a(y), \nabla \mathbf{v}^a) \quad \forall \{\mathbf{v}^a, q^a, \mathbf{g}\} \in V \times Q \times G, \\ \mathscr{B}(\partial_y^{\mathbf{k}} \{\mathbf{u}(y), p(y), \mathbf{g}(y)\}, \{\mathbf{v}, q\}; y) = (\partial_y^{\mathbf{k}} \mathbf{h}(y), \mathbf{v})_{\partial D_N} \\ \qquad - \sum_{\mathbf{k}' \in \Lambda(\mathbf{k})} (\partial_y^{\mathbf{k}-\mathbf{k}'} \nu(y) \nabla \partial_y^{\mathbf{k}'} \mathbf{u}(y), \nabla \mathbf{v}) \quad \forall \{\mathbf{v}, q\} \in V \times Q, \end{cases} \tag{8.48}$$

where the multivariate index set $\Lambda(\mathbf{k})$ is defined as

$$\Lambda(\mathbf{k}) = \left\{\mathbf{k}' \in \mathbb{N}_0^N : k_n' \leq k_n, \forall 1 \leq n \leq N, \text{ and } \mathbf{k}' \neq \mathbf{k}\right\}. \tag{8.49}$$

By the Brezzi theorem 8.2.3, the solution of problem (8.48) admits the following estimate

$$\begin{cases} ||\partial_y^{\mathbf{k}}\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} \leq \alpha_1 \sum\limits_{\mathbf{k}'\in\Lambda(\mathbf{k})} \dfrac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})} ||\partial_y^{\mathbf{k}'}\mathbf{u}^a(y)||_V \\ \qquad\qquad +\beta_1\left(||\partial_y^{\mathbf{k}}\mathbf{h}(y)||_H + \sum\limits_{\mathbf{k}'\in\Lambda(\mathbf{k})} \dfrac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})} ||\partial_y^{\mathbf{k}'}\mathbf{u}(y)||_V\right), \\ ||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \leq \alpha_2 \sum\limits_{\mathbf{k}'\in\Lambda(\mathbf{k})} \dfrac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})} ||\partial_y^{\mathbf{k}'}\mathbf{u}^a(y)||_V \\ \qquad\qquad +\beta_2\left(||\partial_y^{\mathbf{k}}\mathbf{h}(y)||_H + \sum\limits_{\mathbf{k}'\in\Lambda(\mathbf{k})} \dfrac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})} ||\partial_y^{\mathbf{k}'}\mathbf{u}(y)||_V\right), \end{cases} \tag{8.50}$$

where the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ are given in (8.42) and (8.43). Adding the second inequality of (8.46) to the first one and noting that $\forall \mathbf{k}' \in \Lambda(\mathbf{k})$,

$$||\partial_y^{\mathbf{k}'}\mathbf{u}^a(y)||_V \leq ||\partial_y^{\mathbf{k}'}\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \tag{8.51}$$

and

$$||\partial_y^{\mathbf{k}'}\mathbf{u}(y)||_V \leq ||\partial_y^{\mathbf{k}'}\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G}, \tag{8.52}$$

thus

$$\begin{aligned} &||\partial_y^{\mathbf{k}}\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} + ||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \leq C_\beta ||\partial_y^{\mathbf{k}}\mathbf{h}(y)||_H + C_{\alpha,\beta} \\ &\sum_{\mathbf{k}'\in\Lambda(\mathbf{k})} \frac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})} \left(||\partial_y^{\mathbf{k}'}\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} + ||\partial_y^{\mathbf{k}'}\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q}\right), \end{aligned} \tag{8.53}$$

where the constants C_β and $C_{\alpha,\beta}$ are defined in (8.44).

To prove the estimate (8.46) for a general $\mathbf{k} \in \mathbb{N}_0^N$, we adopt an induction argument based on the recursive result (8.53). To start, we consider the case when $|\mathbf{k}| = 0$. Applying the Brezzi theorem to the semi-weak problem (8.47), we have

$$\begin{cases} ||\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} \leq \alpha_1 ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + \beta_1 ||\mathbf{h}(y)||_H, \\ ||\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \leq \alpha_2 ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + \beta_2 ||\mathbf{h}(y)||_H. \end{cases} \tag{8.54}$$

Adding the second inequality of (8.54) to the first one, we find

$$\begin{aligned} &||\{\mathbf{u}(y), p(y), \mathbf{f}(y)\}||_{V\times Q\times G} + ||\{\mathbf{u}^a(y), p^a(y)\}||_{V\times Q} \\ &\leq C_\alpha ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + C_\beta ||\mathbf{h}(y)||_H \\ &= (C_\alpha ||\{\mathbf{u}_d, p_d\}||_{L\times Q} + C_\beta ||\mathbf{h}(y)||_H)|\mathbf{k}|!\mathbf{r}^{\mathbf{k}}, \end{aligned} \tag{8.55}$$

which verifies the estimate (8.46) for $|\mathbf{k}| = 0$ by noting that $r > 1$ and $C = 1$.

When $|\mathbf{k}| = 1$, i.e., there exists $n, 1 \leq n \leq N$ such that $k_n = 1$ and $k_{n^*} = 0$ for all $n^* \neq n, 1 \leq n^* \leq N$, we

have by the estimates (8.53) and (8.55) and Assumption 8.2

$$\begin{aligned}
&||\partial_y^{\mathbf{k}}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}||_{V\times Q\times G}+||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y),p^a(y)\}||_{V\times Q}\\
&=||\partial_y^{k_n}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}||_{V\times Q\times G}+||\partial_y^{k_n}\{\mathbf{u}^a(y),p^a(y)\}||_{V\times Q}\\
&\le C_\beta||\partial_y^{k_n}\mathbf{h}(y)||_H+C_{\alpha,\beta}\frac{|\partial_y^{k_n}\nu(y)|}{\nu(\bar{y})}\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)\\
&\le 2\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)r_n\\
&=2\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)|\mathbf{k}|!\mathbf{r}^{\mathbf{k}},
\end{aligned}\tag{8.56}$$

which yields the estimate (8.46) for $|\mathbf{k}|=1$ by noting that $r>1$ and $C=2$.

As for more general $\mathbf{k}$ with $|\mathbf{k}|>1$, we first prove the following auxiliary estimate

$$\begin{aligned}
&||\partial_y^{\mathbf{k}}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}||_{V\times Q\times G}+||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y),p^a(y)\}||_{V\times Q}\\
&\le\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)s(|\mathbf{k}|)\mathbf{r}^{\mathbf{k}},
\end{aligned}\tag{8.57}$$

where $s(|\mathbf{k}|)$ depends only on $|\mathbf{k}|$ according to the following recursive formula,

$$s(0)=1,s(1)=2,s(|\mathbf{k}|)=1+\sum_{|\mathbf{k}'|=0}^{|\mathbf{k}|-1}\left(\begin{array}{c}|\mathbf{k}|\\|\mathbf{k}'|\end{array}\right)s(|\mathbf{k}'|).\tag{8.58}$$

In fact, (8.57) holds for $|\mathbf{k}|=0$ and $|\mathbf{k}|=1$ due to (8.55) and (8.56). By induction, we assume that the stability estimate (8.57) holds for every $\mathbf{k}'\in\Lambda(\mathbf{k})$, so that (8.53) implies

$$\begin{aligned}
&||\partial_y^{\mathbf{k}}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}||_{V\times Q\times G}+||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y),p^a(y)\}||_{V\times Q}\le C_\beta||\partial_y^{\mathbf{k}}\mathbf{h}(y)||_H\\
&+C_{\alpha,\beta}\sum_{\mathbf{k}'\in\Lambda(\mathbf{k})}\frac{|\partial_y^{\mathbf{k}-\mathbf{k}'}\nu(y)|}{\nu(\bar{y})}\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)s(|\mathbf{k}'|)\mathbf{r}^{\mathbf{k}'}\\
&\le\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)\left(|\mathbf{k}|!\mathbf{r}^{\mathbf{k}}+\sum_{\mathbf{k}'\in\Lambda(\mathbf{k})}\mathbf{r}^{\mathbf{k}-\mathbf{k}'}s(|\mathbf{k}|')\mathbf{r}^{\mathbf{k}'}\right)\\
&=\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)\left(|\mathbf{k}|!+\sum_{|\mathbf{k}'|=0}^{|\mathbf{k}|-1}\left(\begin{array}{c}|\mathbf{k}|\\|\mathbf{k}'|\end{array}\right)s(|\mathbf{k}'|)\right)\mathbf{r}^{\mathbf{k}}\\
&=\left(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H\right)s(|\mathbf{k}|)\mathbf{r}^{\mathbf{k}},
\end{aligned}\tag{8.59}$$

where we have used the assumption 8.2 for the second inequality, the fact that $\mathbf{r}^{\mathbf{k}}=\mathbf{r}^{\mathbf{k}-\mathbf{k}'}\mathbf{r}^{\mathbf{k}'}$ for any $\mathbf{k}'\in\Lambda(\mathbf{k})$, and the following relation by summation reordering

$$\sum_{\mathbf{k}'\in\Lambda(\mathbf{k})}s(|\mathbf{k}|')=\sum_{|\mathbf{k}'|=0}^{|\mathbf{k}|-1}\left(\begin{array}{c}|\mathbf{k}|\\|\mathbf{k}'|\end{array}\right)s(|\mathbf{k}'|),\tag{8.60}$$

thanks to the definition of $\Lambda(\mathbf{k})$ in (8.49). By this end, it is left to establish a suitable bound for $s(|\mathbf{k}|)$ in order to prove the estimate (8.46) from the estimate (8.57). Let us denote $k=|\mathbf{k}|$, $k'=|\mathbf{k}'|$, and define $t(k)=s(k)/k!$, so that from (8.58) we have

$$t(k)=\frac{1}{k!}\left(k!+\sum_{k'=0}^{k-1}\frac{k!}{(k-k')!}\frac{s(k')}{k'!}\right)=1+\sum_{k'=0}^{k-1}\frac{t(k')}{(k-k')!}.\tag{8.61}$$

Suppose that for all k, $t(k)\le c_r r^k$ for some positive constants c_r,r to be determined, so that (8.61) yields

$$t(k)-1=\sum_{k'=0}^{k-1}\frac{t(k')}{(k-k')!}=\sum_{k'=1}^{k}\frac{t(k-k')}{k'!}\le c_r r^k\sum_{k'=1}^{k}\frac{r^{-k'}}{k'!}\le c_r r^k\left(e^{\frac{1}{r}}-1\right). \tag{8.62}$$

On the other hand, $t(k)-1\le c_r r^k-1$ from our assumption. Hence, we only require that $c_r r^k\left(e^{1/r}-1\right)\le c_r r^k-1$, which can be satisfied when $r>1/\log(2)$ and $c_r\ge 1/(2-e^{1/r})$. Therefore, $s(k)=t(k)k!\le c_r r^k k!$, implying that

$$s(|\mathbf{k}|)\le c_r r^{|\mathbf{k}|}|\mathbf{k}|!=c_r\mathbf{r}_r^{\mathbf{k}}|\mathbf{k}|!, \tag{8.63}$$

where the N-dimensional constant rate vector $\mathbf{r}_r=(r,\ldots,r)$. The proof is concluded by substituting (8.63) into (8.57), noting $\mathbf{r}_r^{\mathbf{k}}\mathbf{r}^{\mathbf{k}}=(r\mathbf{r})^{\mathbf{k}}$, and setting $C=c_r$ in (8.46). □

Let us define a complex region associated with the stability estimate (8.46) as

$$\Sigma:=\left\{z\in\mathbb{C}:\exists y\in\Gamma \text{ such that } \sum_{n=1}^{N} r r_n|z_n-y_n|<1\right\}. \tag{8.64}$$

Then we have that the solution does not only have bounded partial derivative but can be analytically extended to the complex region Σ, as stated in the following theorem:

Theorem 8.3.2 *Under assumption 8.2, the solution of the semi-weak saddle point problem (8.47) admits an analytical extension to the region Σ defined in (8.64).*

Proof Given any $y\in\Gamma$, the Taylor expansion of the solution of problem (8.47) $\{\mathbf{u},p,\mathbf{f}\}:\Gamma\to V\times Q\times G$ and $\{\mathbf{u}^a,p^a\}:\Gamma\to V\times Q$ about y reads

$$\{\mathbf{u}(z),p(z),\mathbf{f}(z)\}=\sum_{\mathbf{k}\in\mathbb{N}_0^N}\frac{\partial_y^{\mathbf{k}}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}}{\mathbf{k}!}(z-y)^{\mathbf{k}} \tag{8.65}$$

and

$$\{\mathbf{u}^a(z),p^a(z)\}=\sum_{\mathbf{k}\in\mathbb{N}_0^N}\frac{\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y),p^a(y)\}}{\mathbf{k}!}(z-y)^{\mathbf{k}}, \tag{8.66}$$

where $(z-y)^{\mathbf{k}}=\prod_{n=1}^{N}(z_n-y_n)^{k_n}$. By Theorem 8.3.1, we have

$$\begin{aligned}
&||\{\mathbf{u}(z),p(z),\mathbf{f}(z)\}||_{V\times Q\times G}+||\{\mathbf{u}^a(z),p^a(z)\}||_{V\times Q}\\
&\le\sum_{\mathbf{k}\in\mathbb{N}_0^N}\left(||\partial_y^{\mathbf{k}}\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}||_{V\times Q\times G}+||\partial_y^{\mathbf{k}}\{\mathbf{u}^a(y),p^a(y)\}||_{V\times Q}\right)\frac{|z-y|^{\mathbf{k}}}{\mathbf{k}!}\\
&\le C(C_\alpha||\{\mathbf{u}_d,p_d\}||_{L\times Q}+C_\beta||\mathbf{h}(y)||_H)\sum_{\mathbf{k}\in\mathbb{N}_0^N}|\mathbf{k}|!(r\mathbf{r})^{\mathbf{k}}\frac{|z-y|^{\mathbf{k}}}{\mathbf{k}!}.
\end{aligned} \tag{8.67}$$

Upon reordering, we have

$$\sum_{\mathbf{k}\in\mathbb{N}_0^N}|\mathbf{k}|!(r\mathbf{r})^{\mathbf{k}}\frac{|z-y|^{\mathbf{k}}}{\mathbf{k}!}=\sum_{k=0}^{\infty}\sum_{|\mathbf{k}|=k}\frac{k!}{\mathbf{k}!}\prod_{n=1}^{N}(r r_n|z_n-y_n|)^{k_n}. \tag{8.68}$$

By Newton's generalized binomial formula, we have

$$\sum_{k=0}^{\infty}\sum_{|\mathbf{k}|=k}\frac{k!}{\mathbf{k}!}\prod_{n=1}^{N}(r r_n|z_n-y_n|)^{k_n}=\sum_{k=0}^{\infty}\left(\sum_{n=1}^{N} r r_n|z_n-y_n|\right)^k, \tag{8.69}$$

which converges when $\sum_{n=1}^{N} r r_n|z_n-y_n|<1$. This concludes our proof. □

8.4 Numerical approximation

In order to solve the constrained optimization problem (8.18), we need to solve the equivalent saddle point problem (8.33) (or, equivalently (8.25)) thanks to Lemma 8.2.2. Hereafter, we present a numerical approximation of system (8.25) in the physical domain D by a finite element method. As for the numerical approximation in the probability space, we use the stochastic collocation method as introduced in chapter 1.

Recall the definition of the finite element space X_h^k in (6.57). We define $V_h^k := (X_h^k)^d \cap V$, $Q_h^m := X_h^m \cap Q$, and $G_h^l := (X_h^l)^d \cap G$ with $k, m, l \geq 1$ as finite element approximation spaces corresponding to the Hilbert spaces V, Q and G, respectively, defined in Section 8.3. The semi-weak finite element approximation of the saddle point problem (8.33) reads: for any $y \in \Gamma$, find $\{\mathbf{u}_h(y), p_h(y), \mathbf{f}_h(y)\} \in V_h^k \times Q_h^m \times G_h^l$ and $\{\mathbf{u}_h^a(y), p_h^a(y)\} \in V_h^k \times Q_h^m$ such that

$$\begin{cases} \mathscr{A}\left(\{\mathbf{u}_h(y), p_h(y), \mathbf{f}_h(y)\}, \{\mathbf{v}_h^a, q_h^a, \mathbf{g}_h\}\right) + \mathscr{B}(\{\mathbf{v}_h^a, q_h^a, \mathbf{g}_h\}, \{\mathbf{u}_h^a(y), p_h^a(y)\}; y) \\ \qquad = (\mathbf{u}_d, \mathbf{v}_h^a) + (p_d, q_h^a) \quad \forall \{\mathbf{v}_h^a, q_h^a, \mathbf{g}_h\} \in V_h^k \times Q_h^m \times G_h^l, \\ \mathscr{B}(\{\mathbf{u}_h(y), p_h(y), \mathbf{g}_h(y)\}, \{\mathbf{v}_h, q_h\}; y) = (\mathbf{h}(y), \mathbf{v}_h)_{\partial D_N} \quad \forall \{\mathbf{v}_h, q_h\} \in V_h^k \times Q_h^m. \end{cases} \tag{8.70}$$

The well-posedness of problem (8.70) can be guaranteed by fulfilling the same conditions in Proposition 8.2.1 in finite element spaces. In particular, the compatibility condition (8.32) is satisfied in the finite element spaces V_h^k, Q_h^m, G_h^l as a consequence (as can be observed from the proof (8.37)) of the compatibility condition (8.12) in V_h^k, Q_h^m, for which we may use, e.g., the Taylor-Hood elements ($m = k-1, k \geq 2$) among many feasible choices [161], leading to stable finite element approximation featuring optimal convergence rate. We set $l = k$ for the control function space G_h^l.

Let the finite element solution of the saddle point problem (8.70) be written as

$$\mathbf{u}_h(y) = \sum_{n=1}^{N_v} u_n(y)\boldsymbol{\psi}_n, p_h(y) = \sum_{n=1}^{N_p} p_n(y)\varphi_n, \mathbf{f}_h(y) = \sum_{n=1}^{N_v} f_n(y)\boldsymbol{\psi}_n, \tag{8.71}$$

and

$$\mathbf{u}_h^a(y) = \sum_{n=1}^{N_v} u_n^a(y)\boldsymbol{\psi}_n, p_h^a(y) = \sum_{n=1}^{N_p} p_n^a(y)\varphi_n, \tag{8.72}$$

where $\boldsymbol{\psi}_n, 1 \leq n \leq N_v$ and $\varphi_n, 1 \leq n \leq N_p$ are the bases of the finite element spaces V_h^k and Q_h^k, respectively. Note that the bases of V_h^k and G_h^l are the same when $k = l$. Corresponding to the matrix operators defined in (8.24), the finite element mass matrices $M_{v,h}$ (note that $M_{g,h} = M_{c,h} = M_{v,h}$ when $k = l$) and $M_{p,h}$ are obtained as

$$(M_{v,h})_{mn} = (\boldsymbol{\psi}_n, \boldsymbol{\psi}_m), 1 \leq m, n \leq N_v; (M_{p,h})_{mn} = (\varphi_n, \varphi_m), 1 \leq m, n \leq N_p, \tag{8.73}$$

and the mass matrix for Neumann boundary condition is given by

$$(M_{n,h})_{mn} = (\boldsymbol{\psi}_m, \boldsymbol{\psi}_n)_{\partial D_N}, 1 \le m, n \le N_v. \tag{8.74}$$

The stiffness matrix A_h^y is obtained as

$$(A_h^y)_{mn} = a(\boldsymbol{\psi}_n, \boldsymbol{\psi}_m; y), 1 \le m, n \le N_v, \tag{8.75}$$

and the matrix B_h^y corresponding to the compatibility condition is written as

$$(B_h)_{mn} = b(\boldsymbol{\psi}_m, \varphi_n), 1 \le m \le N_v, 1 \le n \le N_p. \tag{8.76}$$

Let $U_h(y) = (u_1(y), \ldots, u_{N_v}(y))^T$ represent the coefficient vector for the finite element function $\mathbf{u}_h(y)$, and $P_h(y)$, $F_h(y)$, $U_h^a(y)$, $P_h^a(y)$ the coefficient vectors for the functions $p_h(y)$, $\mathbf{f}_h(y)$, $\mathbf{u}_h^a(y)$, and $U_{d,h}$, $P_{d,h}$, $H_h(y)$ the values of $\mathbf{u}_d$, p_d, $\mathbf{h}(y)$ at the finite element nodes. To this end, the algebraic formulation of problem (8.70) can be written via the optimality operator system (8.25) as

$$\left(\begin{array}{ccc:cc} M_{v,h} & 0 & 0 & A_h^y & B_h^T \\ 0 & M_{p,h} & 0 & B_h & 0 \\ 0 & 0 & \alpha M_{g,h} & -M_{c,h}^T & 0 \\ \hdashline A_h^y & B_h^T & -M_{c,h} & 0 & 0 \\ B_h & 0 & 0 & 0 & 0 \end{array}\right) \left(\begin{array}{c} U_h(y) \\ P_h(y) \\ F_h(y) \\ U_h^a(y) \\ P_h^a(y) \end{array}\right) = \left(\begin{array}{c} M_{v,h}U_{d,h} \\ M_{p,h}P_{d,h} \\ 0 \\ M_{n,h}H_h(y) \\ 0 \end{array}\right). \tag{8.77}$$

The matrix of the linear system (8.77) becomes ill-conditioned with large condition number when h or α is very small, which makes it unsuitable for direct solve. Alternatively, we seek the solution by MINRES iteration with the help of the following block diagonal preconditioner [185, 169] (which share similar structure as (7.25) in chapter 7),

$$P(y) = \left(\begin{array}{ccc} \hat{M}_{s,h} & 0 & 0 \\ 0 & \alpha\hat{M}_{g,h} & 0 \\ 0 & 0 & \hat{K}_{s,h}^y M_{s,h}^{-1} (\hat{K}_{s,h}^y)^T \end{array}\right). \tag{8.78}$$

The mass matrix $M_{s,h}$ and the saddle point matrix $K_{s,h}^y$ corresponding to the Stokes problem (8.2) in deterministic setting are defined as

$$M_{s,h} = \left(\begin{array}{cc} M_{v,h} & 0 \\ 0 & M_{p,h} \end{array}\right) \text{ and } K_{s,h}^y = \left(\begin{array}{cc} A_h^y & B_h^T \\ B_h & 0 \end{array}\right), \tag{8.79}$$

where the matrices $\hat{M}_{s,h}$, $\hat{M}_{g,h}$ and $\hat{K}_{s,h}^y$ can be regarded as convenient approximations of $M_{s,h}$, $M_{g,h}$ and $K_{s,h}^y$ obtained by using suitable iteration methods [144, 169], e.g., symmetric Gauss-Seidel iteration for $\hat{M}_{s,h}$ and $\hat{M}_{g,h}$, and inexact Uzawa iteration for $\hat{K}_{s,h}^y$.

8.5 Multilevel and weighted reduced basis method

To solve a full system (8.77) at one sample $y \in \Gamma$ is very expensive when the number of degrees of freedom of the finite element approximation is large. The task becomes prohibitive when the dimension of the probability space Γ is so high that a large number of samples are necessary to be used in order to obtain accurate statistics of interest. To circumvent this computational obstacle, we adopt a reduced basis method [49, 66, 142, 143] and propose a new algorithm featuring multilevel greedy algorithm and weighted a posteriori error bound. The crucial consideration is that the optimal solution of the constrained optimization problem (8.18) resides in a low-dimensional manifold, despite the fact that the random inputs live in high-dimensional probability space.

8.5.1 Reduced basis approximation

The idea behind reduced basis approximation is to take “snapshots" - that is high fidelity solutions of the underlying PDE model - as bases and then approximate the solution at a new sample by Galerkin projection on the pre-selected snapshots [178, 49]. Specific to the finite element problem (8.77), the associated reduced basis problem can be formulated as: for any $y \in \Gamma$, find $\{\mathbf{u}_r(y), p_r(y), \mathbf{f}_r(y)\} \in V_{N_r} \times Q_{N_r} \times G_{N_r}$ and $\{\mathbf{u}_r^a(y), p_r^a(y)\} \in V_{N_r} \times Q_{N_r}$ such that

$$\begin{cases} \mathscr{A}\left(\{\mathbf{u}_r(y), p_r(y), \mathbf{f}_r(y)\}, \{\mathbf{v}_r^a, q_r^a, \mathbf{g}_r\}\right) + \mathscr{B}(\{\mathbf{v}_r^a, q_r^a, \mathbf{g}_r\}, \{\mathbf{u}_r^a(y), p_r^a(y)\}; y) \\ \qquad = (\mathbf{u}_d, \mathbf{v}_r^a) + (p_d, q_r^a) \quad \forall \{\mathbf{v}_r^a, q_r^a, \mathbf{g}_r\} \in V_{N_r} \times Q_{N_r} \times G_{N_r}, \\ \mathscr{B}(\{\mathbf{u}_r(y), p_r(y), \mathbf{g}_r(y)\}, \{\mathbf{v}_r, q_r\}; y) = (\mathbf{h}(y), \mathbf{v}_r)_{\partial D_N} \quad \forall \{\mathbf{v}_r, q_r\} \in V_{N_r} \times Q_{N_r}, \end{cases} \tag{8.80}$$

where $V_{N_r}, Q_{N_r}, G_{N_r}$ are reduced basis spaces constructed from the snapshots at the pre-selected samples $y^1, \ldots, y^{N_r}$. More in detail, G_{N_r} is constructed by

$$G_{N_r} = \text{span}\{\mathbf{f}_h(y^n), 1 \le n \le N_r\}. \tag{8.81}$$

As for Q_{N_r}, we take the union of the state and adjoint snapshots of pressure in order to guarantee the approximate stability in the reduced basis space [144, 143], written as

$$Q_{N_r} = Q_{N_r}^s \cup Q_{N_r}^a = \text{span}\{p_h(y^n), p_h^a(y^n), 1 \le n \le N_r\}. \tag{8.82}$$

As for V_{N_r}, a simple union of the state and adjoint snapshots of velocity is not sufficient to satisfy the compatibility condition (8.32) in the reduced basis spaces. To overcome this difficulty, the reduced basis velocity space can be enriched by introducing the supremizer operator $T : Q_h^m \to V_h^k$ [180, 177],

$$(Tq_h, \mathbf{v}_h)_A = b(\mathbf{v}_h, q_h) \quad \forall \mathbf{v} \in V_h^k, \tag{8.83}$$

where the A-scalar product is defined as

$$(\mathbf{u}, \mathbf{v})_A = a(\mathbf{u}, \mathbf{v}; \bar{y}) \quad \forall \mathbf{u}, \mathbf{v} \in V, \tag{8.84}$$

being $\bar{y} \in \Gamma$ a reference value, for instance, the center of Γ. Then, we construct the reduced basis velocity space V_{N_r} as the union of state and adjoint velocity snapshots enriched by pressure supremizers

$$V_{N_r} = V_{N_r}^s \cup V_{N_r}^a = \text{span}\{\mathbf{u}_h(y^n), Tp_h(y^n), \mathbf{u}_h^a(y^n), Tp_h^a(y^n), 1 \le n \le N_r\}. \tag{8.85}$$

It can be proven [177] that the compatibility condition (8.12) is satisfied in $V_{N_r}^s$ and $V_{N_r}^a$ with $\beta_b^{N_r} \ge \beta_b^h$, being $\beta_b^{N_r}$ and β_b^h the compatibility constants of the bilinear form b of (8.12) in the reduced basis space and finite element space, respectively. Consequently, the compatibility condition (8.32) is satisfied in V_{N_r} following the proof of (8.37), with the compatibility constants $\beta^{N_r} \ge \beta^h$ corresponding to that in (8.38). Following the argument in the proof of Lemma 8.2.2, it is straightforward to check that the other conditions in Proposition 8.2.1 are also satisfied in the reduced basis space $V_{N_r} \times Q_{N_r} \times G_{N_r}$. Hence, there exists a unique reduced basis solution to problem (8.80).

For the sake of algebraic stability, we perform Gram-Schmidt orthonormalization [177] on the reduced basis spaces V_{N_r}, Q_{N_r} and G_{N_r}, obtaining the orthonormal bases such that $V_{N_r} = \text{span}\{\boldsymbol{\zeta}_n^v, 1 \le n \le 4N_r\}$, $Q_{N_r} = \text{span}\{\zeta_n^p, 1 \le n \le 2N_r\}$ and $G_{N_r} = \text{span}\{\boldsymbol{\zeta}_n^g, 1 \le n \le N_r\}$. Finally, at any $y \in \Gamma$, we project the finite element solution $\{\mathbf{u}_h(y), p_h(y), \mathbf{f}_h(y)\} \in V_h^k \times Q_h^m \times G_h^l$ into the reduced basis space $V_{N_r} \times Q_{N_r} \times G_{N_r}$ as

$$\mathbf{u}_h(y) = \sum_{n=1}^{4N_r} u_n(y)\boldsymbol{\zeta}_n^v, p_h(y) = \sum_{n=1}^{2N_r} p_n(y)\zeta_n^p, \mathbf{f}_h(y) = \sum_{n=1}^{N_r} f_n(y)\boldsymbol{\zeta}_n^g, \tag{8.86}$$

and the adjoint variables $\{\mathbf{u}_h^a(y), p_h^a(y)\} \in V_h^k \times Q_h^m$ into $V_{N_r} \times Q_{N_r}$ as

$$\mathbf{u}_h^a(y) = \sum_{n=1}^{4N_r} u_n^a(y)\boldsymbol{\zeta}_n^v, p_h^a(y) = \sum_{n=1}^{2N_r} p_n^a(y)\zeta_n^p. \tag{8.87}$$

Let $U_r(y) = (u_1(y), \ldots, u_{4N_r}(y))$ denote the coefficient vector of the reduced basis approximation, and define $P_r(y), F_r(y), U_r^a(y)$ and $P_r^a(y)$ similarly, corresponding to those of the finite element approximation. Let $\mathcal{Z}_{N_r}^v = (\boldsymbol{\zeta}_1^v, \ldots, \boldsymbol{\zeta}_{4N_r}^v)^T$, $\mathcal{Z}_{N_r}^p = (\zeta_1^p, \ldots, \zeta_{2N_r}^p)^T$ and $\mathcal{Z}_{N_r}^g = (\boldsymbol{\zeta}_1^g, \ldots, \boldsymbol{\zeta}_{N_r}^g)^T$, by which we define the reduced basis mass matrices as follows: $M_{v,r} = (\mathcal{Z}_{N_r}^v)^T M_{v,h} \mathcal{Z}_{N_r}^v$, $M_{p,r} = (\mathcal{Z}_{N_r}^p)^T M_{p,h} \mathcal{Z}_{N_r}^p$, $M_{g,r} = (\mathcal{Z}_{N_r}^g)^T M_{g,h} \mathcal{Z}_{N_r}^g$, $M_{c,r} = (\mathcal{Z}_{N_r}^v)^T M_{c,h} \mathcal{Z}_{N_r}^g$, $M_{n,r} = (\mathcal{Z}_{N_r}^v)^T M_{n,h} \mathcal{Z}_{N_r}^v$, and the Stokes matrices A_r^y and B_r as $A_r^y = (\mathcal{Z}_{N_r}^v)^T A_h^y \mathcal{Z}_{N_r}^v$, and $B_r = (\mathcal{Z}_{N_r}^p)^T B_h \mathcal{Z}_{N_r}^v$. The reduced basis data vector $U_{d,r}, P_{d,r}, H_r(y)$ are defined as $U_{d,r} = (\mathcal{Z}_{N_r}^v)^T U_{d,h}, P_{d,r} = (\mathcal{Z}_{N_r}^p)^T P_{d,h}, H_r(y) = (\mathcal{Z}_{N_r}^v)^T H_h(y)$. By projecting the finite element system (8.77) into the reduced basis spaces, we obtain the algebraic formulation of the reduced basis problem corresponding to the finite element algebraic system (8.77) as

$$\left(\begin{array}{ccc:cc} M_{v,r} & 0 & 0 & A_r^y & B_r^T \\ 0 & M_{p,r} & 0 & B_r & 0 \\ 0 & 0 & \alpha M_{g,r} & -M_{c,r}^T & 0 \\ \hdashline A_r^y & B_r^T & -M_{c,r} & 0 & 0 \\ B_r & 0 & 0 & 0 & 0 \end{array}\right) \left(\begin{array}{c} U_r(y) \\ P_r(y) \\ F_r(y) \\ U_r^a(y) \\ P_r^a(y) \end{array}\right) = \left(\begin{array}{c} M_{v,r} U_{d,r} \\ M_{p,r} P_{d,r} \\ 0 \\ M_{n,r} H_r(y) \\ 0 \end{array}\right), \tag{8.88}$$

which is a $13N_r \times 13N_r$ linear system, whose numerical solution costs far less computational effort than solving the finite element system (8.77) thanks to the fact that N_r is much smaller than the number of degrees of freedom of the finite element discretization.

8.5.2 A multilevel greedy algorithm

The efficiency of the reduced basis approximation depends critically on the choice of reduced bases, and thus on the samples $y^1, \ldots, y^{N_r}$ selected in the construction of the reduced basis spaces $V_{N_r}, Q_{N_r}, G_{N_r}$. In order to choose the most representative samples, we propose a multilevel greedy algorithm based on the sparse grid construction for stochastic collocation method and reduce the computational cost of the construction of the reduced basis spaces.

The multilevel greedy algorithm for construction of the reduced basis space is presented in Algorithm 10.

To begin, we choose the first sample from the zeroth level of the sparse grid, i.e., $y^1 \in H(q,N)$ (or $H_{\boldsymbol{\alpha}}(q,N)$ for anisotropic sparse grid) with $q - N = 0$, where only one collocation node is available. We solve the finite element problem (8.77) at y^1 and construct the reduced basis space V_1, Q_1, G_1 according to (8.81), (8.82) and (8.85).

Let $\mathscr{E}_r$ denote the reduced basis approximation error defined as

$$\mathscr{E}_r(y) := ||\mathrm{u}_h - \mathrm{u}_r||_{\mathrm{V}}, \tag{8.89}$$

where we denote the Hilbert space $\mathrm{V} = V \times Q \times G \times V \times Q$, the solution $\mathrm{u}(y) := \{\mathbf{u}(y), p(y), \mathbf{f}(y), \mathbf{u}^a(y), p^a(y)\}$ with finite element approximation u_h and reduced basis approximation u_r. At each of the level $q - N = l, l = 1, 2, \ldots, L$ with prescribed $L \in \mathbb{N}_+$, we first construct the set of collocation nodes $H(q,N)$ of the sparse grid and then choose the "most representative" sample y^{N_r+1} by minimizing $\mathscr{E}_r(y)$ over the new collocation nodes in the current level of the sparse grid, i.e.

$$y^{N_r+1} = \arg \max_{y \in H(q,N) \setminus H(q-1,N)} \mathscr{E}_r(y). \tag{8.90}$$

Note that in the hierarchical sparse grid with nested collocation nodes, we have $H(q-1,N) \subset H(q,N), q \geq N+1$, which provides further computational efficiency since there is no need to evaluate the error at the collocation nodes in the previous level. After updating the reduced basis spaces V_{N_r}, Q_{N_r} and G_{N_r} by the finite element solution of problem (8.77) at y^{N_r+1}, we set $N_r+1 \rightarrow N_r$ and proceed to choose the next sample until the error $\mathscr{E}_r(y^{N_r+1})$ is smaller than a prescribed tolerance ϵ_{tol}. Then we move to the next level $q-N=l+1$. However, in order to compute the reduced basis approximation error (8.89), we have to solve the full finite element system (8.77), which is out of reach. Instead of computing an exact reduced basis approximation error $\mathscr{E}_r(y)$, we seek to evaluate a cheap, sharp and reliable error bound $\triangle_r(y)$ depending on $\{\mathbf{u}_r(y), p_r(y), \mathbf{f}_r(y), \mathbf{u}_r^a(y), p_r^a(y)\}$ at $y \in \Gamma$ such that

$$c\triangle_r(y) \leq \mathscr{E}_r(y) \leq \triangle_r(y) \tag{8.91}$$

with the constant c as close to 1 as possible.

Algorithm 10 A multilevel greedy algorithm

procedure INITIALIZATION
  Set maximum sparse grid level L, tolerance ϵ_{tol}, $q=N$, take $y^1 \in H(q,N)$;
  Solve (8.77), construct the initial reduced basis spaces V_1, Q_1, G_1, set $N_r=1$.
end procedure

procedure CONSTRUCTION
  for $q=N+1,\ldots,N+L$ **do**
    Construct the set of collocation nodes $H(q,N)$, take $H(q,N)\setminus H(q-1,N)$;
    Solve (8.88) to obtain $y^{N_r+1}=\arg\max_{y\in H(q,N)\setminus H(q-1,N)} \triangle_r(y)$;
    while $\triangle_r(y^{N_r+1}) \geq \epsilon_{tol}$ **do**
      Set $N_r \leftarrow N_r+1$;
      Solve (8.77) at y^{N_r}, update the reduced basis spaces $V_{N_r}, Q_{N_r}, G_{N_r}$;
      Solve (8.88) to obtain $y^{N_r+1}=\arg\max_{y\in H(q,N)\setminus H(q-1,N)} \triangle_r(y)$.
    end while
  end for
end procedure

8.5.3 A weighted a posteriori error bound

In order to efficiently evaluate a sharp and reliable bound for the reduced basis approximation error, we carry out a residual-based a posteriori error estimate. At first, we reformulate the semi-weak saddle point problem (8.47) as an elliptic problem: for any $y \in \Gamma$, find $\mathrm{u}(y) \in \mathrm{V}$

$$\mathrm{B}(\mathrm{u}(y),\mathrm{v};y) = \mathrm{F}(\mathrm{v};y) \quad \forall \mathrm{v} \in \mathrm{V}, \tag{8.92}$$

where the bilinear form $\mathrm{B}(\cdot,\cdot;y): \mathrm{V}\times\mathrm{V} \rightarrow \mathbb{R}$ is given by

$$\begin{aligned} \mathrm{B}(\mathrm{u}(y),\mathrm{v};y) &= \mathscr{A}\left(\{\mathbf{u}(y),p(y),\mathbf{f}(y)\},\{\mathbf{v}^a,q^a,\mathbf{g}\}\right) \\ &+ \mathscr{B}(\{\mathbf{v}^a,q^a,\mathbf{g}\},\{\mathbf{u}^a(y),p^a(y)\};y) + \mathscr{B}(\{\mathbf{u}(y),p(y),\mathbf{g}(y)\},\{\mathbf{v},q\};y), \end{aligned} \tag{8.93}$$

and the linear functional

$$\mathrm{F}(\mathrm{v};y) = (\mathbf{u}_d,\mathbf{v}^a) + (p_d,q^a) + (\mathbf{h}(y),\mathbf{v})_{\partial D_N}. \tag{8.94}$$

The bilinear form $\mathrm{B}(\cdot,\cdot;y): \mathrm{V}\times\mathrm{V} \rightarrow \mathbb{R}$ can be proven to be continuous and weakly coercive [209] since the bilinear forms $\mathscr{A},\mathscr{B}$ satisfy the conditions in Proposition 8.2.1, yielding the continuous and weak

coercivity constants $\gamma_c(y)$ and $\beta_c(y)$ defined as

$$\gamma_c(y) := \frac{\mathrm{B}(\mathrm{u},\mathrm{v};y)}{||\mathrm{u}||_{\mathrm{V}}||\mathrm{v}||_{\mathrm{V}}} < \infty \text{ and } \beta_c(y) := \inf_{\mathrm{v}\in\mathrm{V}} \sup_{\mathrm{u}\in\mathrm{V}} \frac{\mathrm{B}(\mathrm{u},\mathrm{v};y)}{||\mathrm{u}||_{\mathrm{V}}||\mathrm{v}||_{\mathrm{V}}} > 0. \tag{8.95}$$

Therefore, by Babuška theorem [209] we have the following stability estimate

$$||\mathrm{u}(y)||_{\mathrm{V}} \leq \frac{||\mathrm{F}(y)||_{\mathrm{V}'}}{\beta_c(y)}, \tag{8.96}$$

where V' is the dual space of V. By the construction of the finite element approximation in section 8.4 and the reduced basis approximation in section 8.5.1, we have that the relation (8.95) holds in both finite element space $\mathrm{V}_h = V_h \times Q_h \times G_h \times V_h \times Q_h$ and reduced basis space $\mathrm{V}_r = V_r \times Q_r \times G_r \times V_r \times Q_r$ with constants $\gamma_c^{N_r}(y) \leq \gamma_c^h(y) \leq \gamma_c(y)$ and $\beta_c^{N_r}(y) \geq \beta_c^h(y) \geq \beta_c(y)$. Moreover, the stability estimate (8.96) holds for the finite element solution and the reduced basis solution with the constant $\beta_c^h(y)$ and $\beta_c^{N_r}(y)$, respectively. Let the reduced basis approximation error be defined as $\mathrm{e}(y) = \mathrm{u}_h(y) - \mathrm{u}_r(y)$. To seek an error bound for $\mathrm{e}(y)$, we consider the residual

$$\mathrm{R}(\mathrm{v}_h;y) := \mathrm{F}(\mathrm{v}_h;y) - \mathrm{B}(\mathrm{u}_r(y),\mathrm{v}_h;y) \quad \mathrm{v}_h \in \mathrm{V}_h. \tag{8.97}$$

Noting that $\mathrm{F}(\mathrm{v}_h;y) = \mathrm{B}(\mathrm{u}_h,\mathrm{v}_h;y), \forall \mathrm{v}_h \in \mathrm{V}_h$, we have from (8.97)

$$\mathrm{B}(\mathrm{e}(y),\mathrm{v}_h;y) = \mathrm{R}(\mathrm{v}_h;y) \quad \mathrm{v}_h \in \mathrm{V}_h. \tag{8.98}$$

By the stability estimate (8.96) in the finite element space, we obtain

$$||\mathrm{e}(y)||_{\mathrm{V}_h} \leq \frac{||\mathrm{R}(y)||_{\mathrm{V}_h'}}{\beta_c^h(y)} =: \triangle_r(y). \tag{8.99}$$

Taking the probability density function $\rho : \Gamma \to \mathbb{R}_+$ into account, we replace $\mathscr{E}_r(y)$ in (8.90) by a weighted a posteriori error bound [49] $\triangle_r^\rho(y) = \sqrt{\rho(y)}\triangle_r(y)$. The error bound $\triangle_r^\rho(y)$ assigns high importance at the sample with big probability density, leading to more efficient (using less bases to achieve the same accuracy) evaluation of statistical moments of interest; see [49] for proof and illustrative examples. In order to evaluate the error bound (8.99), we need to compute both the constant $\beta_c^h(y)$ and the norm of the residual $||\mathrm{R}||_{\mathrm{V}_h'}$. For the former, we may apply successive constraint method [101] to compute a lower bound $\beta_c^{LB}(y) \leq \beta_c^h(y)$ (or a surrogate lower bound [143]) with cheap computational cost, or simply use a uniform lower bound $\beta_c^{LB} \leq \beta_c^h(y)$ evaluated at the minimum random viscosity ν_{min} provided that the random coefficient $\nu(y)$ varies in a relatively small range. As for the latter, we turn to an offline-online decomposition procedure in order to reduce computational effort in the many-query context.

8.5.4 Offline-online decomposition

The offline-online decomposition takes advantage of the affine structure of the data, as given in examples (8.15) and (8.16). If the data are provided in a non-affine structure, e.g., log-normal Karhunen–Loève expansion [149], we may apply a weighted empirical interpolation method to obtain an affine decomposition of the data function at first; see [48] for details and error analysis. Let us assume that the random viscosity and the Neumann boundary condition undergoes, after possibly performing empirical interpolation [11, 48], the following affine structure

$$\nu(y) = \sum_{n=1}^{N_\nu} \nu_n \theta_n^\nu(y) \text{ and } \mathbf{h}(x,y) = \sum_{n=1}^{N_h} \mathbf{h}_n(x)\theta_n^h(y) \quad \forall (x,y) \in \partial D_N \times \Gamma, \tag{8.100}$$

where $\theta_n^v, 1 \le n \le N_v$ and $\theta_n^h, 1 \le n \le N_h$ are functions of the random vector $y \in \Gamma$. Let the matrix A_r^y and vector $H_r(y)$ in (8.88) be assembled as

$$A_r^y = \sum_{n=1}^{N_v} A_r^n \theta_n^v(y) \text{ and } H_r(y) = \sum_{n=1}^{N_h} H_r^n \theta_n^h(y), \tag{8.101}$$

where the deterministic reduced basis matrices $A_r^n, 1 \le n \le N_v$ are defined as

$$A_r^n = (\mathcal{Z}_{N_r}^v)^T A_h^n \mathcal{Z}_{N_r}^v \text{ with } (A_h^n)_{ij} = (\nu_n \nabla \boldsymbol{\psi}_i, \nabla \boldsymbol{\psi}_j), 1 \le i, j \le N_v, \tag{8.102}$$

and the deterministic reduced basis vectors $H_r^n, 1 \le n \le N_h$ are defined as

$$H_r^n = (\mathcal{Z}_{N_r}^v)^T H_h^n \text{ with } (H_h^n)_i = (\mathbf{h}_n, \boldsymbol{\psi}_i)_{\partial D_N}, 1 \le i \le N_v. \tag{8.103}$$

Accordingly, we decompose the global matrix of the linear system (8.88) as

$$\mathrm{B}_r^0 = \left(\begin{array}{ccc:c:c} M_{v,r} & 0 & 0 & 0 & B_r^T \\ 0 & M_{p,r} & 0 & B_r & 0 \\ 0 & 0 & \alpha M_{g,r} & -M_{c,r}^T & 0 \\ \hdashline 0 & B_r^T & -M_{c,r} & 0 & 0 \\ \hdashline B_r & 0 & 0 & 0 & 0 \end{array} \right) \tag{8.104}$$

and $\mathrm{B}_r^n, 1 \le n \le N_v$ with only the blocks $(4,1),(1,4)$ as A_r^n the other blocks zero. Similarly, we decompose the vector on the right hand side of the linear system (8.88) as $\mathrm{F}_r^0 = (M_{v,r}U_{d,r}, M_{p,r}P_{d,r}, 0, 0, 0)^T$ and $\mathrm{F}_r^n = (0,0,0,M_{n,r}H_r^n,0)^T, 1 \le n \le N_h$. Thus, the algebraic formulation of the problem (8.92) can be written as: for any $y \in \Gamma$, find $\mathrm{U}_r(y) := (U_r(y), P_r(y), F_r(y), U_r^a(y), P_r^a(y))^T \in \mathbb{R}^{13N_r}$ such that

$$\left(\sum_{n=0}^{N_v} \theta_n^v(y) \mathrm{B}_r^n \right) \mathrm{U}_r(y) = \sum_{n=0}^{N_h} \theta_n^h(y) \mathrm{F}_r^n. \tag{8.105}$$

Since $\mathrm{B}_r^n, 1 \le n \le N_v$ and $\mathrm{F}_r^n, 1 \le n \le N_h$ are independent of y, we can assemble them in offline stage. Given any $y \in \Gamma$, the reduced basis solution can be obtained by solving the linear system (8.105) with at most $O(N_v + N_h)$ operations for assembling and $O((13N_r)^3)$ operations for solve.

As for the evaluation of the residual norm $||\mathrm{R}(y)||_{\mathrm{V}_h'}$, we first seek the Riesz representation [165] of $\mathrm{R}(y)$ as $\hat{\mathrm{e}}(y) \in \mathrm{V}_h$ such that

$$(\hat{\mathrm{e}}(y), \mathrm{v}_h)_{\mathrm{V}_h} = \mathrm{R}(\mathrm{v}_h; y) \quad \forall \mathrm{v}_h \in \mathrm{V}_h, \tag{8.106}$$

so that we have $||\mathrm{R}(y)||_{\mathrm{V}_h'} = ||\hat{\mathrm{e}}(y)||_{\mathrm{V}_h}$. Let $\mathrm{B}_n : \mathrm{V}_h \times \mathrm{V}_h \to \mathbb{R}$ denote the bilinear form defined in the finite element space corresponding to the matrix $\mathrm{B}_r^n, 0 \le n \le N_v$ and $\mathrm{F}_n : \mathrm{V}_h \to \mathbb{R}$ the linear functional corresponding to the vector $\mathrm{F}_r^n, 0 \le n \le N_h$, then the residual defined in (8.97) can be decomposed as

$$\mathrm{R}(\mathrm{v}_h; y) = \sum_{n=0}^{N_h} \theta_n^h(y) \mathrm{F}_n(\mathrm{v}_h) - \sum_{n=0}^{N_v} \theta_n^v(y) \mathrm{B}_n(\mathrm{u}_r, \mathrm{v}_h) \quad \forall \mathrm{v}_h \in \mathrm{V}_h. \tag{8.107}$$

By Riesz representation theorem, we have that there exist $\mathrm{f}_n \in \mathrm{V}_h, 0 \le n \le N_h$ and $\mathrm{b}_n^k \in \mathrm{V}_h, 0 \le n \le N_v, 1 \le k \le 13N_r$ such that

$$(\mathrm{f}_n, \mathrm{v}_h)_{\mathrm{V}_h} = \mathrm{F}_n(\mathrm{v}_h) \text{ and } (\mathrm{b}_n^k, \mathrm{v}_h)_{\mathrm{V}_h} = -\mathrm{B}_n(\mathrm{u}_h^k, \mathrm{v}_h) \quad \forall \mathrm{v}_h \in \mathrm{V}_h, \tag{8.108}$$

where we have set the reduced basis solution as $\mathrm{u}_h^k = (\boldsymbol{\psi}_k^v, 0, 0, 0, 0), 1 \le k \le 4N_r$, $\mathrm{u}_h^k = (0, \varphi_{k-4N_r}^p, 0, 0, 0)$, $4N_r < k \le 6N_r$, $\mathrm{u}_h^k = (0, 0, \boldsymbol{\psi}_{k-6N_r}^g, 0, 0)$, $6N_r < k \le 7N_r$, $\mathrm{u}_h^k = (0, 0, 0, \boldsymbol{\psi}_{k-7N_r}^v, 0)$, $7N_r < k \le 11N_r$, $\mathrm{u}_h^k = (0, 0, 0, 0, \varphi_{k-11N_r}^p)$, $11N_r < k \le 13N_r$, being 0 the vector with length N_v, N_p, N_v, N_v, N_p at the first to

fifth argument. Finally, we obtain the norm $||\hat{e}(y)||_{\mathrm{V}_h}$ as

$$\begin{aligned}||\hat{e}(y)||^2_{\mathrm{V}_h} = &\sum_{n=1}^{N_h}\sum_{n'=1}^{N_h}\theta_n^h(y)(\mathrm{f}_n,\mathrm{f}_{n'})_{\mathrm{V}_h}\theta_{n'}^h(y)\\ &+2\sum_{n=1}^{N_h}\sum_{n'=1}^{N_v}\sum_{k=1}^{13N_r}\theta_n^h(y)(\mathrm{f}_n,\mathrm{b}_{n'}^k)_{\mathrm{V}_h}(\mathrm{u}_r)_k\theta_{n'}^v(y)\\ &+\sum_{n=1}^{N_h}\sum_{n'=1}^{N_v}\sum_{k=1}^{13N_r}\sum_{k'=1}^{13N_r}\theta_n^v(y)(\mathrm{u}_r)_k(\mathrm{b}_n^k,\mathrm{b}_{n'}^{k'})_{\mathrm{V}_h}(\mathrm{u}_r)_{k'}\theta_{n'}^v(y),\end{aligned} \tag{8.109}$$

where $(\mathrm{f}_n,\mathrm{f}_{n'})_{\mathrm{V}_h}, 1\le n,n'\le N_h$, $(\mathrm{f}_n,\mathrm{b}_{n'}^k)_{\mathrm{V}_h}, 1\le n\le N_h, 1\le n'\le N_v, 1\le k\le 13N_r$ and $(\mathrm{b}_n^k,\mathrm{b}_{n'}^{k'})_{\mathrm{V}_h}, 1\le n,n'\le N_v, 1\le k,k'\le 13N_r$ are independent of y and can be computed and stored in the offline stage, while in the online stage, we only need to assemble the formula (8.109) by $O((N_h+13N_rN_v)^2)$ operations. Recall that N_h and N_v are the number of affine terms of the random Neumann boundary condition and the viscosity, and N_r is the number of selected samples in the construction of reduced basis space, leading to fast evaluation of the error bound as they are small.

8.6 Error estimates

The global error of the numerical approximation presented in sections 8.4 and 8.5 comprises three components: the stochastic collocation approximation error [8, 149, 148], the finite element approximation error [165, 161], and the weighted reduced basis approximation error [20, 50, 49], which have been analyzed individually in different contexts. The global error estimate for the Stokes optimal control is similar to that for the elliptic optimal control presented in the last chapter. In the following, for simplicity we provide the finite element approximation error and a global error estimate in the context of the stochastic Stokes optimal control problem (8.18). The integration error $\mathscr{E}_s^e$ by the stochastic collocation method can be estimated the same as Lemmas 6.3.2 and 6.3.3 of chapter 6, and reduced basis error $\mathscr{E}_r$ can be bounded the same as in Proposition 1.4.5 of chapter 1 or in Theorem 2.2.5 of chapter 2 and thus omitted here.

8.6.1 Finite element approximation error

Recall that the bilinear forms $\mathscr{A}$ and $\mathscr{B}$ of the finite element problem (8.70) satisfy the conditions of Proposition 8.2.1 in the finite element space V_h^k, Q_h^m, G_h^l with the choice of Taylor-Hood elements. More explicitly, the finite element constants corresponding to those stated in the conditions of Proposition 8.2.1 can be bounded by

$$\gamma_h\le 1, \epsilon_h\ge\frac{1}{3}\min\left\{\frac{\alpha_a^2\alpha}{2C_P^2},\frac{1}{2}\right\}, \delta_h\le\max\{\nu_{max},\gamma_b,1/\sqrt{\alpha}\}, \beta_h\ge\beta, \tag{8.110}$$

being the constants $\alpha_a,\alpha,C_P,\nu_{max},\gamma_b,\beta$ presented in Lemma 8.2.2. Therefore, by Brezzi theorem [165, 161], we have the following estimate for the error $\mathscr{E}_h$ of the finite element approximation to

solution of the semi-weak saddle point problem (8.47):

$$\begin{aligned}
\mathscr{E}_h(y) &:= ||\mathrm{u}(y)-\mathrm{u}_h(y)||_{\mathrm{V}} \\
&\leq C_1^h \inf_{\{\mathbf{v}_h,q_h,\mathbf{g}_h\}\in V_h^k\times Q_h^m\times G_h^l} ||\{\mathbf{u}(y),p(y),\mathbf{f}(y)\}-\{\mathbf{v}_h,q_h,\mathbf{g}_h\}||_{V\times Q\times G} \\
&+ C_2^h \inf_{\{\mathbf{v}_h^a,q_h^a\}\in V_h^k\times Q_h^m} ||\{\mathbf{u}^a(y),p^a(y)\}-\{\mathbf{v}_h^a,q_h^a\}||_{V\times Q} \\
&= O(h^k)\left(C_1^h(||\mathbf{u}(y)||_{k+1}+||p(y)||_k+\sqrt{\alpha}||\mathbf{f}(y)||_{k+1})\right) \\
&+ O(h^k)\left(C_2^h(||\mathbf{u}^a(y)||_{k+1}+||p^a(y)||_k)\right),
\end{aligned} \tag{8.111}$$

where we have chosen $m=k-1$ and $l=k$; the constants C_1^h and C_2^h are given by

$$C_1^h=\left(1+\frac{\gamma_h}{\epsilon_h}\right)\left(1+\frac{\gamma_h}{\beta_h}\right)\left(1+\frac{\delta_h}{\beta_h}\right) \text{ and } C_2^h=1+\frac{\delta_h}{\epsilon_h}+\frac{\delta_h}{\beta_h}+\frac{\gamma_h\delta_h}{\epsilon_h\beta_h}. \tag{8.112}$$

Remark 8.6.1 *Equivalently, we may formulate the semi-weak saddle point finite element problem (8.70) as a weakly coercive elliptic problem and apply Babuška theorem to obtain similar finite element error estimate.*

8.6.2 Global error estimate

With the individual error estimate presented above, we obtain the global error estimate in the following theorem.

Theorem 8.6.1 *Under Assumptions 8.1, 8.2, for any given $y\in\Gamma$, by finite element approximation and reduced basis approximation we have*

$$||\mathrm{u}(y)-\mathrm{u}_r(y)||_{\mathrm{V}}\leq\mathscr{E}_h(y)+\mathscr{E}_r(y). \tag{8.113}$$

Moreover, the error for evaluation of the expectation using stochastic collocation method, finite element method and weighted reduced basis method can be bounded by

$$||\mathbb{E}[\mathrm{u}]-\mathbb{E}[\mathrm{u}_r]||_{\mathrm{V}}\leq\mathscr{E}_s^e+\max_{y\in H_{\boldsymbol{\alpha}}(q,N)}\mathscr{E}_h(y)+\max_{y\in H_{\boldsymbol{\alpha}}(q,N)}\mathscr{E}_r(y), \tag{8.114}$$

where $\boldsymbol{\alpha}=\mathbf{1}$ when using the isotropic sparse grid stochastic collocation method.

Proof The proof is straightforward by applying triangular inequality as follows:

$$||\mathrm{u}(y)-\mathrm{u}_r(y)||_{\mathrm{V}}\leq||\mathrm{u}(y)-\mathrm{u}_h(y)||_{\mathrm{V}}+||\mathrm{u}_h(y)-\mathrm{u}_r(y)||_{\mathrm{V}}\leq\mathscr{E}_h(y)+\mathscr{E}_r(y). \tag{8.115}$$

Similarly, we have the error estimate for the expectation of the optimal solution as

$$\begin{aligned}
||\mathbb{E}[\mathrm{u}]-\mathbb{E}[\mathrm{u}_r]||_{\mathrm{V}} &\leq ||\mathrm{u}-\mathrm{u}_r||_{L^2_\rho(\Gamma;\mathrm{V})} \\
&\leq ||\mathrm{u}-\mathrm{u}_s||_{L^2_\rho(\Gamma;\mathrm{V})}+||\mathrm{u}_s-\mathrm{u}_h||_{L^2_\rho(\Gamma;\mathrm{V})}+||\mathrm{u}_h-\mathrm{u}_r||_{L^2_\rho(\Gamma;\mathrm{V})} \\
&\leq \mathscr{E}_s^e+\max_{y\in H_{\boldsymbol{\alpha}}(q,N)}\mathscr{E}_h(y)+\max_{y\in H_{\boldsymbol{\alpha}}(q,N)}\mathscr{E}_r(y).
\end{aligned} \tag{8.116}$$

We remark that $\mathscr{E}_r(y)$ is bounded by $\triangle_r(y)$, explicitly computed at $y\in H_{\boldsymbol{\alpha}}(q,N)$. □

8.7 Numerical experiments

In this section, we perform two numerical experiments in testing reduced basis approximation error and stochastic collocation approximation error with sparse grid techniques in isotropic and anisotropic settings. The aim is to demonstrate the efficiency of the proposed reduced basis method in solving constrained optimization problem (8.18). Numerical examples for verifying finite element approximation error in a similar context can be found in [47].

We consider a two dimensional physical domain $D = (0,1)^2$. The observation data is set as in [92], $\mathbf{u}_d = (u_{d1}, u_{d2})$ and $p_d = 0$, where $u_{d1}(x) = \partial_{x_2}(\phi(x_1)\phi(x_2))/10$ and $u_{d2}(x) = -\partial_{x_1}(\phi(x_1)\phi(x_2))/10$ with $\phi(\xi) = (1-\cos(0.8\pi\xi))(1-\xi)^2, \xi \in [0,1]$. The random viscosity ν is given as in (8.15) which can be transformed as

$$\nu(y^\nu) = \frac{1}{2}\sum_{n=0}^{N_\nu} \nu_n + \frac{1}{2N_\nu}\sum_{n=1}^{N_\nu} (\nu_n - \nu_0) y_n^\nu, \tag{8.117}$$

where $y^\nu \in \Gamma_\nu = [-1,1]^{N_\nu}$ corresponding to N_ν uniformly distributed random variables. We set $\nu_0 = 0.01$, $\nu_n = \nu_0/2^n$ and use $N_\nu = 3$ for both the isotropic and anisotropic tests without loss of generality. Homogeneous Dirichlet boundary condition is imposed on the upper, lower and left edge. Random Neumann boundary condition is imposed on the right edge as given in (8.16) on the Neumann boundary, more explicitly, we set $\mathbf{h}(x, y^h) = (h_1(x_2, y^h), 0)$ with

$$h_1(x_2, y^h) = \frac{1}{10}\left(\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} y_1^h + \sum_{n=1}^{N_h} \sqrt{\lambda_n}\left(\sin(n\pi x_2) y_{2n}^h + \cos(n\pi x_2) y_{2n+1}^h\right)\right), \tag{8.118}$$

which comes from truncation of Karhunen–Loève expansion of a Gauss covariance field with correlation length $L = 1/16$ [149]; the eigenvalues $\lambda_n, 1 \le n \le N_h$ are given by

$$\lambda_n = \sqrt{\pi} L \exp\left(-(n\pi L)^2/4\right); \tag{8.119}$$

$y_n^h, 1 \le n \le 2N_h + 1$ are uncorrelated with zero mean and unit variance, which are independent of y^ν. Therefore, the random inputs are $y = (y^\nu, y^h)$, living in $N = N_\nu + 2N_h + 1$ dimensional probability space. As for the specification of the finite element approximation, we use P1 element for pressure space and P2 element for velocity and control space with 1342 elements in total.

8.7.1 Isotropic case

In the first experiment, we set $y_n^h, 1 \le n \le 2N_h + 1$ with $N_h = 3$ as independent standard normal random variables (thus the total stochastic dimension $N = 10$), and apply isotropic sparse grid stochastic collocation method with Gauss-Legendre abscissa for the collocation of y^ν and Gauss-Hermite abscissa for the collocation of y^h. In the multilevel greedy algorithm 10, we set the tolerance $\epsilon_{tol} = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and the interpolation level $q - N = 0, 1, 2, 3$ in the isotropic sparse grid Smolyak formula (1.15). A uniformly lower bound of the inf-sup constant $\beta_c^{LB} = 0.1436$ is used since the fluctuation or variance of ν is small compared to its mean value. The results for reduced basis construction is reported in Table 8.1. The number of collocation nodes in each level is shown in the second row; the number of selected samples as new bases in each level and the samples whose weighted error bound $\triangle_r^\rho$ is larger than the tolerance ϵ_{tol}, thus potential as new bases are shown in the 3rd-6th lines, from which we can see that the number of reduced bases is much less than that of collocation nodes. For example, with the smallest tolerance $\epsilon_{tol} = 10^{-5}$, we only need $1, 10, 22, 14$ new bases in each level, respectively, resulting in 47 bases in total out of 1581 collocation nodes. Since the number of samples as potential bases is also small (216 in total), the computational cost for sample selection in the construction of reduced basis space is negligible compared to the full solve of the finite element problem (8.77), especially for large scale problems featuring a small mesh size h.

Table 8.1: The number of samples selected by multilevel greedy algorithm 10 with different tolerance ϵ_{tol} in each of the sparse grid level; the value in $(\cdot)$ reports the number of samples potential as new bases.

tolerance \ level	$q-N=0$	$q-N=1$	$q-N=2$	$q-N=3$	in total
# nodes	1	21	221	1581	1581
$\epsilon_{tol}=10^{-1}$	1 (1)	6 (14)	1 (21)	0 (0)	8 (36)
$\epsilon_{tol}=10^{-2}$	1 (1)	8 (20)	7 (80)	4 (28)	20 (129)
$\epsilon_{tol}=10^{-3}$	1 (1)	9 (20)	13 (86)	5 (62)	28 (169)
$\epsilon_{tol}=10^{-4}$	1 (1)	9 (20)	18 (90)	9 (67)	37 (178)
$\epsilon_{tol}=10^{-5}$	1 (1)	10 (20)	22 (90)	14 (105)	47 (216)

Fig. 8.1 (left) displays the weighted error bound $\triangle_r^{\rho}$ and the true error of the reduced basis approximation in each level of the construction, from which we can see that the error bound is accurate and relatively sharp, providing good estimate of the true error with cheap computation. On the right of Fig. 8.1 we plot the expectation error (in $L^2_{\rho}(\Gamma;\mathrm{V})$ norm) of the reduced basis approximation using quadrature formula based on sparse grid of different levels, where the expectation error is defined as

$$\text{exp. error} = |||\mathrm{u}||_{L^2_{\rho}(\Gamma;\mathrm{V})} - ||\mathrm{u}_{s,r}||_{L^2_{\rho}(\Gamma;\mathrm{V})}| = |(\mathbb{E}[||\mathrm{u}||^2_{\mathrm{V}}])^{1/2} - (\mathbb{E}[||\mathrm{u}_{s,r}||^2_{\mathrm{V}}])^{1/2}|. \tag{8.120}$$

Note that the "true" value of $||\mathrm{u}||_{L^2_{\rho}(\Gamma;\mathrm{V})}$ is approximated by the finite element solution u_h computed at the deepest level $q-N=3$. From this figure, different accuracy with different ϵ_{tol} can be observed, implying that decreasing tolerance for the construction of the reduced basis space results in more accurate evaluation of statistics of the solution. How to balance the reduced basis approximation error (by choice of ϵ_{tol}) and the sparse grid quadrature error (by choice of $q-N$) is subject to further investigation.

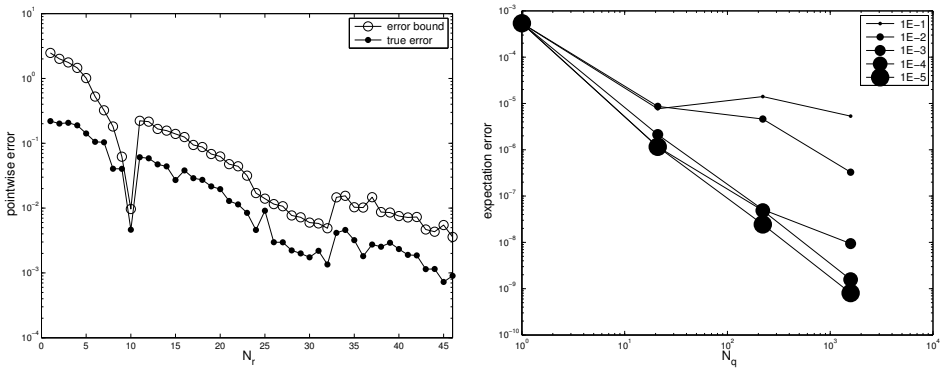

Figure 8.1: Left, weighted error bound $\triangle_r^{\rho}$ and true error of the reduced basis approximation at the selected samples; right, expectation error at different levels with different tolerance ϵ_{tol}.

8.7.2 Anisotropic case

In the second experiment, we solve the constrained optimization problem (8.18) in high-dimensional probability space by combination of the anisotropic sparse grid techniques and the multilevel weighted reduced basis method. We set $y_n^h, 1 \le n \le N_h$ in (8.118) with $N_h = 3,8,13,18,48$ as uniformly distributed

random variables, thus leading to $N = 10, 20, 30, 40, 100$ stochastic dimensions in total. The weight parameter $\boldsymbol{\alpha}$ is chosen a priori according to [148] in the following conservative way

$$\alpha_n = \frac{1}{2}\log\left(1 + \frac{2\tau_n}{|\Gamma_n|}\right), \text{ with } \tau_n = \frac{1}{4\sqrt{\lambda_n}}, \quad 1 \le n \le N_h. \tag{8.121}$$

We remark that for a more general random field where $\boldsymbol{\alpha}$ is difficult to obtain from a priori estimate, we may use a posteriori estimate by fitting an empirical convergence rate in each dimension [148], or use dimension-adaptive approach which determines the weight automatically [79]. The sparse grid level is chosen as $q - N = 0, 1, 2, 3, 4$. As for the tolerance for the construction of the reduced basis space, we use $\epsilon_{tol} = 10^{-5}$. The results for the construction of the reduced basis space with different dimension N and different sparse grid level $q - N$ (results for $q - N = 0$ are the same as in Table 8.1, thus omitted here) are presented in Table 8.2. Similar conclusion as for results in the isotropic case in Table 8.1 can be drawn for those in the anisotropic case in Table 8.2. For example, when $N = 40$, only 97 samples out of 40479 are used for the construction of the reduced basis space, thus resulting in only 97 full solve the finite element problem (8.77) instead of 40479, which considerably reduces the total computational cost. This observation holds even in the 100 dimensional case. Moreover, the number of nodes of sparse grid and the number of reduced bases increase as the dimension increase when N is small; see the change from 10 to 40. However, they stay almost the same when N becomes large; see the change from 40 to 100, which indicates that out of 100 random variables, the first 40 play the most important role on the impact of the stochastic optimal solution when we set sparse grid level at $q - N = 4$.

Table 8.2: The number of samples selected by multilevel greedy algorithm 10 in each of the sparse grid level with different dimensions; the value in $(\cdot)$ reports the number of samples potential as new bases.

dimension \ level	$q-N=1$	$q-N=2$	$q-N=3$	$q-N=4$	in total
$N=10$	5 (10)	13 (40)	19 (85)	10 (100)	48 (236)
# nodes	11	71	401	2141	2141
$N=20$	5 (10)	21 (60)	36 (205)	15 (204)	78 (480)
# nodes	11	91	1021	12121	12121
$N=40$	5 (10)	25 (92)	47 (397)	19 (432)	97 (932)
# nodes	11	123	2381	40769	40769
$N=100$	5 (10)	25 (92)	47 (397)	19 (436)	97 (936)
# nodes	11	123	2393	41349	41349

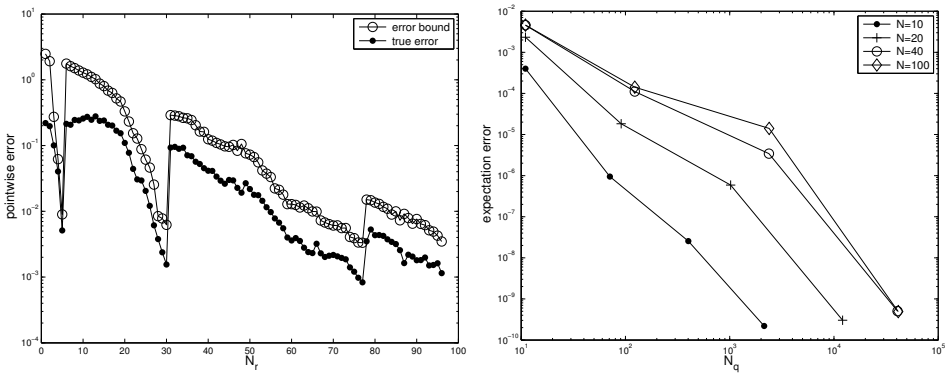

Figure 8.2: Weighted error bound $\triangle_r^{\rho}$ and true error of the reduced basis approximation at the selected samples with $N = 100$ (left); expectation error of different stochastic dimensions (right).

On the left of Fig. 8.2, we plot the weighted a posteriori error bound $\triangle_r^\rho$ and the true error of the reduced basis approximation at each sparse grid level with stochastic dimension $N = 100$. We can observe that the error bound is indeed accurate and sharp for the high-dimensional case, especially when the reduced basis space becomes large. The right of Fig. 8.2 depicts the expectation error at different sparse grid level. We show the expectation error with the “true" expectation for each stochastic dimension computed the same as in the isotropic sparse grid case, from which we can see that the expectation error converges with an algebraic rate that verifies the error estimate in section 8.6. Moreover, the error becomes very small at around 4×10^4 nodes for the 100 dimensional problem by anisotropic sparse grid technique, which would need around 7×10^7 nodes for isotropic sparse grid technique at the same sparse grid level $q - N = 4$. Furthermore, we can observe that no “plateau" (flattening) of expectation error appears as in Fig. 8.1, demonstrating that the multilevel reduced basis method is very efficient in producing the accurate statistics of the stochastic optimal solution even when the number of the reduced bases shown in Table 8.2 remains critically small (around 97 for high dimensions).

8.8 Summary

In this chapter we studied the mathematical properties of an optimal control problem constrained by stochastic Stokes equations and developed a computational strategy by using sparse grid techniques and the model order reduction approach. The existence and uniqueness of the stochastic optimal solution was proved by establishing the equivalence between the constrained optimization problem and the stochastic saddle point problem. Moreover, we obtained some stochastic regularity results of the optimal solution in the probability space under some mild assumptions on the random input data. In the fully discretized problem, we used finite element approximation in the deterministic space and stochastic collocation approximation in the probability space, and proposed a multilevel and weighted reduced basis method in order to reduce the computational effort in the many-query context, for which a global error estimate was carried out. This computational approach was proven to be very efficient by two numerical experiments, especially for high-dimensional and large-scale problems requiring a large number of samples and heavy computational cost for a full solve of the optimization problem at each sample. Further study on more general statistical cost functional, adaptive scheme to balance various computational errors and applications to practical flow control problems are ongoing [44].

Conclusions and Perspectives

In this thesis, we have developed and analyzed novel stochastic computational strategies and algorithms based on model order reduction techniques, specifically a reduced basis method, in order to overcome some common computational challenges arising in the solution of representative uncertainty quantification problems. To convey a comprehensive understanding of the thesis, we draw a few generic conclusions beyond the short summary in each chapter. Many other numerical challenges and open questions besides those we have dealt with were identified along the presentation of the thesis: some perspectives of further research and application possibilities are outlined in the following.

Conclusions

Throughout the thesis, we have performed detailed comparison of the reduced basis method and the stochastic collocation method with emphasis on their convergence properties and computational costs in different contexts of uncertainty quantification problems, e.g., statistical analysis, reliability analysis and stochastic optimal control problems. From the comparison results, we can conclude that the reduced basis approximation error converges much faster than the stochastic collocation approximation error under the same smoothness hypothesis. One major reason is that in the former method we use the solutions of the underlying problem as the bases, which can capture the main characteristics of the solution or the quantities of interest associated with the solution. On the other hand, the latter approximation method uses dictionary bases, such as Lagrange polynomials or piecewise splines, which are generic but blind to the underlying problems. Dictionary bases have also been employed for the other important stochastic computational method – the stochastic Galerkin projection method based on wavelets or generalized polynomial chaos, which have comparable or slightly faster convergence rates than the stochastic collocation method but would converge much slower than the reduced basis method. This major difference between problem-specific bases and dictionary bases leads to a significant contrast of the two approaches in terms of the number of approximation bases – the former needs only tens or hundreds of bases while the latter would require millions of bases or beyond for high-dimensional problems. Another reason is that the reduced basis output is computed by solving a reduced order model (by Galerkin projection) that retains the same structure of the underlying problem, while the collocation output is approximated by using some interpolation formula generic to different problems.

As for the computational costs, the reduced basis method is demonstrated to be much more efficient than the stochastic collocation method for large-scale and high-dimensional uncertainty quantification problems. One reason is that the full solution of a large-scale problem is computationally very expensive, and the former method is able to replace most of the full solutions by reduced basis solutions with much cheaper and affordable cost. Another reason is that a large number of samples are typically needed for high-dimensional problems under certain accuracy constraint, so that the latter method becomes too expensive or computationally prohibitive to be directly applied since it involves a full solution at each sample. The significant computational reduction by the reduced basis method is

however mitigated by a more complex implementation required by this method, as shown in our presentation related to different contexts of uncertainty quantification problems. More in detail, the collocation approach can take the underlying deterministic model as a black box and directly use its solver, while the implementation of the reduced basis approximation depends on the structure of the model for the evaluation of the a posteriori error bound and the assembling of the reduced basis system. Efficient implementation and computation of the reduced basis solution and the error bound for many complex problems, e.g., multiscale and multiphysics problems, may take a lot of workforce and are still onging research.

The a posteriori error bound plays a critical role in efficiently and accurately solving different uncertainty quantification problems by the reduced basis method. It is this ingredient that enables us to develop suitable algorithms for different types of problems. By incorporating the probability density function in the a posteriori error bound according to the model outputs, we developed the weighted algorithm that uses less reduced bases and achieves the same accuracy as the classical one for stochastic models with arbitrary probability measures. By taking the distance to the limit state surface of a failure domain into the a posteriori error indicator, we developed a goal-oriented adaptive algorithm that can construct a refined approximation near the limit sate surface with a limited number of reduced bases, which remarkably results in the same failure probability as obtained by solving the full model. By tailoring the stochastic optimality system to feature a convenient implementation of an a posteriori error bound for both the state and the adjoint variables, we obtained a systemic approximation and certification of the optimal solution of some stochastic optimal control problems. The a posteriori error bound does not only provide an appropriate criteria for efficiently constructing the reduced basis space but also offer a reliable certification for the accuracy of the quantities of interest. These two properties differ the reduced basis method from some other model order reduction techniques, such as that by proper orthogonal decomposition for which no accurate or reliable error bounds are provided except some error indicators (e.g., truncation tolerance). These properties of the reduced basis method may enable certified pointwise evaluation with remarkable applications in solving uncertainty quantification problems, such as accurate computation of failure probability.

Different from the application of the reduced basis method in the classical parametric models, its implementation in the stochastic models for uncertainty quantification problems appreciably benefits from the other stochastic computational methods for another key ingredient – the training set. As demonstrated in this thesis, various nodes for interpolation and integration problems used by the stochastic collocation method have been effectively employed as the training set for the reduced basis method. An immediate practice is the direct comparison of the approximation errors of the two methods by taking the collocation nodes as the training set in the first chapter, which also leads to combination of the two methods for efficiently evaluating the statistical moments of the model outputs by suitable quadrature/cubature formulas. Moreover, in solving high-dimensional uncertainty quantification problems, the hierarchical construction of a generalized sparse grid also makes the adaptive grid nodes available as the training samples for the reduced basis construction. The marriage of the quadrature/cubature nodes and the training samples yields an automatic, adaptive and efficient way for constructing and applying the reduced basis method for the solution of a large class of uncertainty quantification problems involving integration, such as the statistical moments, variance-based sensitivity analysis, stochastic optimal control problems with statistical observations.

Perspectives

We hope that the advantages of various computational strategies and algorithms demonstrated and advocated in this thesis can boost the development and the application of model order reduction techniques to solve more general uncertainty quantification problems. In order to achieve this objective, we outline some further computational challenges besides the ones identified in this thesis and provide some associated perspectives of promising research topics following two paths.

1. Development of model order reduction techniques:

    - As shown in this thesis, the a posteriori error bound is critically important depending on different goals of the problem for the efficiency and accuracy of the reduced order/basis approximation. Therefore, one should always try to make the goal of the problem, e.g., pointwise evaluation or integration, as clear as possible beforehand and construct a goal-oriented a posteriori error bound in order to enhance computational efficiency [32, 4, 39]. A further challenge of particular interest for practical engineering problems is to estimate the global approximation error including not only the reduced basis error but also spatial and temporal discretization errors [3]. Thus, how to accurately formulate and efficiently evaluate a global a posteriori error bound for balanced refinement in the probability space and the physical space remain an open problem. Moreover, it is crucial to design an effective strategy for its evaluation in massively parallel architectures for solving large-scale problems [111].
    - Model order reduction techniques have been well developed for linear and steady problems as mainly considered in this thesis, but they are far from mature for nonlinear and unsteady problems involving multiscale and multiphysical phenomenons. For instance, the unsteady Navier–Stokes equations with large Reynolds number or some turbulence flow models produce a large number of modes that can hardly be captured by a limited number of reduced bases; how to build a global reduced order model to approximate high-fidelity fluid and structure interaction models remains to be investigated; it is also significantly difficult to apply model order reduction techniques in approximating hyperbolic PDEs that feature locally supported traveling waves or shocks. One possible approach is to allow the reduced bases to evolve along time such that they can capture the time dependent main modes/characteristics of the dynamic underlying model [183, 184, 52, 53]. Careful separation and application of the reduced order approximation and the high-fidelity approximation in a hybrid way [195], as employed for risk analysis in this thesis, may provide another potential.

2. Application to uncertainty quantification problems:

    - High-dimensional problems have been partially addressed under smoothness hypothesis of the random input data and sparsity hypothesis of the outputs. However, when the outputs do not depend smoothly on the inputs either because of nonsmooth random inputs or due to high nonlinearity of the underlying model, or when the outputs are not sparse such that many different dimensions of the random inputs play equally important role or they have strong interaction with each other, the proposed computational framework in this thesis would not be efficient or even fail. Effective algorithms in detecting and resolving the low regularity points – discontinuous or singular points – should be incorporated in the adaptive construction of the reduced order models [128, 105]. Moreover, how to automatically choose and combine different sampling techniques, e.g., Latin hypercube sampling or generalized sparse grid sampling, in the reduced basis construction provides another important research area for accurate and efficient approximation of the quantities of interest.
    - One important branch of uncertainty quantification problems is to predict the risk of failure of a given system under uncertainties. The hybrid and goal-oriented adaptive reduced basis method was proved to work efficiently for this type of problem with relatively big failure probability. When the failure probability becomes critically small (e.g., smaller than 10^{-6}) and the consequence of the failure is catastrophic, known as high consequence rare events, the proposed algorithm can not be directly used since a huge number of standard Monte Carlo samples (e.g., more than 10^8) is needed to obtain a relatively accurate evaluation of the failure probability. However, almost all the samples except a few tens or hundreds in this example locate outside the failure domain, which are of less use but take

considerable evaluation cost even if the reduced basis online evaluation cost remains small. Moreover, a very accurate reduced basis approximation, thus relatively expensive online evaluation, should be constructed to guarantee that the adaptive a posteriori error indicator for the output becomes smaller than or at least comparable to the failure probability. This computational complexity becomes more evident in high-dimensional problems. Therefore, efficient combination of importance sampling techniques (e.g., using cross-entropy [120]) to reduce the total number of samples, and the hybrid and goal-oriented adaptive reduced basis models to facilitate accurate certification of the outputs is very promising albeit challenging.

- Stochastic inverse problems account for another important branch of uncertainty quantification problems, including optimal control/design/optimization, parameter estimation and data assimilation under various uncertainties. Variation approach based on Lagrange multiplier and a finite-element-stochastic-collocation-reduced-basis approximation has been employed to solve a simple type of stochastic optimal control problems – linear-quadratic optimal control. In more practical applications, the observation data may involve high statistical moments or probability distribution beyond the first order expectation as considered in this thesis, the underlying model may not be linear or steady, leading to strongly coupled [197] and highly nonlinear stochastic optimal control problems. In order to address these more general cases, the one-shot approach used in this thesis does not apply any more and we have to develop more efficient iterative algorithms and linearization strategies in combination with model order reduction for both state and adjoint systems. Additional computational complexity may arise from stochastic constraints for the control variables [200], which requires more careful formulation of the goal-oriented a posteriori error estimate for the construction of the reduced order models. These challenges and the reduction opportunities also apply for other types of stochastic inverse problems, such as Bayesian inversion by Markov chain Monte Carlo methods for parameter estimation [122], Kalman filter and its extensions for data assimilation [17]. A common feature of these stochastic inverse problems is that the many-query requirement comes not only from the stochastic emulation but also from the iterative emulation of the underlying models, which makes the reduced order models even more appealing for substantial reduction of computation.

Bibliography

[1] A. Abdulle and Y. Bai. Reduced basis finite element heterogeneous multiscale method for high-order discretizations of elliptic homogenization problems. *Journal of Computational Physics*, 231(21):7014–7036, 2012.

[2] G. Acosta and R.G. Durán. An optimal Poincaré inequality in L1 for convex domains. *Proceedings of the American Mathematical Society*, 132(1):195–202, 2004.

[3] M. Ainsworth and J.T. Oden. *A posteriori error estimation in finite element analysis*, volume 37. John Wiley & Sons, Hoboken, NJ, 2011.

[4] R.C. Almeida and J.T. Oden. Solution verification, goal-oriented adaptive methods for stochastic advection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 199(37):2472–2486, 2010.

[5] M. Arroyo, L. Heltai, D. Millán, and A. DeSimone. Reverse engineering the euglenoid movement. *Proceedings of the National Academy of Sciences*, 109(44):17874–17879, 2012.

[6] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla. An assessment of discretizations for convection-dominated convection–diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 200(47):3395–3409, 2011.

[7] I. Babuška, K.M. Liu, and R. Tempone. Solving stochastic partial differential equations based on the experimental data. *Mathematical Models and Methods in Applied Sciences*, 13(3):415–444, 2003.

[8] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45(3):1005–1034, 2007.

[9] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Review*, 52(3):317, 2010.

[10] I. Babuška, R. Tempone, and G.E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 42(2):800–825, 2005.

[11] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera. An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique, Analyse Numérique*, 339(9):667–672, 2004.

[12] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: A numerical comparison. In J.S. Hesthaven and E.M. Rønquist, editors, *Spectral and High Order Methods for Partial Differential Equations*, pages 43–62. Springer-Verlag, Berlin, 2011.

[13] J. Beck, R. Tempone, F. Nobile, and L. Tamellini. On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods. *Mathematical Models and Methods in Applied Sciences*, 22(09), 2012.

[14] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta numerica*, 10(1):1–102, 2001.

[15] R. Becker and B. Vexler. Optimal control of the convection-diffusion equation using stabilized finite element methods. *Numerische Mathematik*, 106(3):349–367, 2007.

[16] G. Berkooz, P. Holmes, and J.L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual review of fluid mechanics*, 25(1):539–575, 1993.

[17] C. Bertoglio, P. Moireau, and J-F Gerbeau. Sequential parameter estimation for fluid–structure problems: Application to hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 28(4):434–455, 2012.

[18] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox, and Y. Marzouk. *Large-scale inverse problems and quantification of uncertainty*. John Wiley & Sons, Hoboken, NJ, 2011.

[19] M. Bieri and C. Schwab. Sparse high order FEM for elliptic sPDEs. *Computer Methods in Applied Mechanics and Engineering*, 198(13):1149–1170, 2009.

[20] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM Journal of Mathematical Analysis*, 43(3):1457–1472, 2011.

[21] G. Blatman and B. Sudret. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique*, 336(6):518–523, 2008.

[22] P.B. Bochev and M.D. Gunzburger. Least-squares finite-element methods for optimization and control problems for the Stokes equations. *Computers & Mathematics with Applications*, 48(7):1035–1057, 2004.

[23] P.B. Bochev and M.D. Gunzburger. *Least-squares finite element methods*, volume 166. Springer, New York, 2009.

[24] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N.C. Nguyen, and A.T. Patera. Reduced basis techniques for stochastic problems. *Archives of Computational Methods in Engineering*, 17:435–454, 2010.

[25] S. Boyaval, C. LeBris, Y. Maday, N.C. Nguyen, and A.T. Patera. A reduced basis approach for variational problems with stochastic parameters: Application to heat conduction with variable Robin coefficient. *Computer Methods in Applied Mechanics and Engineering*, 198(41-44):3187–3206, 2009.

[26] M. Braack. Optimal control in fluid mechanics by finite elements with symmetric stabilization. *SIAM Journal on Control and Optimization*, 48:672, 2009.

[27] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis*, 8(2):129–151, 1974.

[28] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Springer, New York, 1991.

[29] C.G. Bucher and U. Bourgund. A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, 7(1):57–66, 1990.

[30] A. Buffa, Y. Maday, A. Patera, C. Prudhomme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46:595–603, 2011.

[31] T. Bui-Thanh, K. Willcox, and O. Ghattas. Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computing*, 30(6):3270–3288, 2008.

[32] T. Bui-Thanh, K. Willcox, O. Ghattas, and B. van Bloemen Waanders. Goal-oriented, model-constrained optimization for reduction of large-scale systems. *Journal of Computational Physics*, 224(2):880–896, 2007.

[33] H.J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13(1):147–269, 2004.

[34] G.T. Buzzard and D. Xiu. Variance-based global sensitivity analysis via sparse-grid interpolation and cubature. *Communications in Computational Physics*, 9:542–67, 2011.

[35] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer-Verlag, Berlin, 2006.

[36] C. Canuto, Y. Maday, and A. Quarteroni. Analysis of the combined finite element and Fourier interpolation. *Numerische Mathematik*, 39(2):205–220, 1982.

[37] Y. Cao, J. Zhu, I.M. Navon, and Z. Luo. A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, 53(10):1571–1583, 2007.

[38] O. Cappé, R. Douc, A. Guillin, J.M. Marin, and C.P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

[39] K. Carlberg and C. Farhat. A low-cost, goal-oriented compact proper orthogonal decomposition basis for model reduction of static systems. *International Journal for Numerical Methods in Engineering*, 86(3):381–402, 2011.

[40] S. Chaturantabut and D.C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, 2010.

[41] P. Chen and A. Quarteroni. Accurate and efficient evaluation of failure probability for partial differential equations with random input data. *Computer Methods in Applied Mechanics and Engineering*, 267(0):233–260, 2013.

[42] P. Chen and A. Quarteroni. A new algorithm for high-dimensional uncertainty quantification problems based on dimension-adaptive and reduced basis methods. *Submitted*, 2014.

[43] P. Chen and A. Quarteroni. Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraints. *To appear in SIAM/ASA Journal on Uncertainty Quantification,*, 2014.

[44] P. Chen, A. Quarteroni, and G. Rozza. Fast solver for optimal control problems constrained by stochastic PDE with general statistical observation data. *in preparation*, 2013.

[45] P. Chen, A. Quarteroni, and G. Rozza. Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations. *Submitted*, 2013.

[46] P. Chen, A. Quarteroni, and G. Rozza. Simulation-based uncertainty quantification of human arterial network hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 29(6):698–721, 2013.

[47] P. Chen, A. Quarteroni, and G. Rozza. Stochastic optimal Robin boundary control problems of advection-dominated elliptic equations. *SIAM Journal on Numerical Analysis*, 51(5):2700 – 2722, 2013.

[48] P. Chen, A. Quarteroni, and G. Rozza. A weighted empirical interpolation method: A priori convergence analysis and applications. *ESAIM: Mathematical Modelling and Numerical Analysis, in press, EPFL, MATHICSE Report 05*, 2013.

[49] P. Chen, A. Quarteroni, and G. Rozza. A weighted reduced basis method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 51(6):3163 – 3185, 2013.

[50] P. Chen, A. Quarteroni, and G. Rozza. Comparison between reduced basis and stochastic collocation methods for elliptic problems. *Journal of Scientific Computing*, 59:187–216, 2014.

[51] Y. Chen, J.S. Hesthaven, Y. Maday, and J. Rodríguez. Certified reduced basis methods and output bounds for the harmonic Maxwell's equations. *SIAM Journal on Scientific Computing*, 32(2):970–996, 2010.

[52] M. Cheng, T.Y. Hou, and Z. Zhang. A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations i: Derivation and algorithms. *Journal of Computational Physics*, 242:843–868, 2013.

[53] M. Cheng, T.Y. Hou, and Z. Zhang. A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations ii: Adaptivity and generalizations. *Journal of Computational Physics*, 2013.

[54] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic SPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010.

[55] A. Cohen, R. Devore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Analysis and Applications*, 9(01):11–47, 2011.

[56] S.S. Collis and M. Heinkenschloss. Analysis of the streamline upwind Petrov Galerkin method applied to the solution of optimal control problems. *CAAM TR02-01*, 2002.

[57] National Research Council. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. The National Academies Press, Washington, D.C., 2012.

[58] L. Dedè. Reduced basis method and a posteriori error estimation for parametrized linear-quadratic optimal control problems. *SIAM Journal on Scientific Computing*, 32(2):997–1019, 2010.

[59] L. Dedè. Reduced basis method and error estimation for parametrized optimal control problems with control constraints. *Journal of Scientific Computing*, 50(2):287–305, 2012.

[60] S. Deparis and G. Rozza. Reduced basis method for multi-parameter-dependent steady Navier–Stokes equations: Applications to natural convection in a cavity. *Journal of Computational Physics*, 228(12):4359–4378, 2009.

[61] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer, New York, 1993.

[62] J. Dick, F.Y. Kuo, and I.H. Sloan. High-dimensional integration–the Quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[63] M. Drohmann, B. Haasdonk, and M. Ohlberger. Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation. *SIAM Journal on Scientific Computing*, 34(2):A937–A969, 2012.

[64] R. Durrett. *Probability: theory and examples*. Cambridge University Press, Cambridge, UK, 2010.

[65] J.L. Eftang, A.T. Patera, and E.M. Rønquist. An “hp" certified reduced basis method for parametrized elliptic partial differential equations. *SIAM Journal on Scientific Computing*, 32(6):3170–3200, 2010.

[66] H. Elman and Q. Liao. Reduced basis collocation methods for partial differential equations with random coefficients. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):192–217, 2013.

[67] O.G. Ernst, C.E. Powell, D.J. Silvester, and E. Ullmann. Efficient solvers for a linear stochastic galerkin mixed formulation of diffusion problems with random data. *SIAM Journal of Scientific Computing*, 31(2):1424–1447, 2009.

[68] L.C. Evans. *Partial Differential Equations*, volume 19. Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2009.

[69] L. Faravelli. Response-surface approach for reliability analysis. *Journal of Engineering Mechanics*, 115(12):2763–2781, 1989.

[70] P. Feldmann and R.W. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 14(5):639–649, 1995.

[71] C. Feuersänger. *Sparse Grid Methods for Higher Dimensional Approximation*. Dissertation, Institut für Numerische Simulation, Universität Bonn, September 2010.

[72] G.S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York, 1996.

[73] J. Foo and G.E. Karniadakis. Multi-element probabilistic collocation method in high dimensions. *Journal of Computational Physics*, 229(5):1536–1557, 2010.

[74] J. Foo, X. Wan, and G.E. Karniadakis. The multi-element probabilistic collocation method (ME-PCM): error analysis and applications. *Journal of Computational Physics*, 227(22):9572–9595, 2008.

[75] L. Formaggia, A. Quarteroni, and A. Veneziani, editors. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system*, volume 1. MS&A, Springer, Milano, 2009.

[76] P. Frauenfelder, C. Schwab, and R.A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Computer Methods in Applied Mechanics and Engineering*, 194(2-5):205–228, 2005.

[77] A.C. Galeao, R.C. Almeida, S.M.C. Malta, and A.F.D. Loula. Finite element analysis of convection dominated reaction–diffusion problems. *Applied Numerical Mathematics*, 48(2):205–222, 2004.

[78] Z. Gao and J.S. Hesthaven. On ANOVA expansions and strategies for choosing the anchor point. *Applied Mathematics and Computation*, 217(7):3274–3285, 2010.

[79] T. Gerstner and M. Griebel. Dimension–adaptive tensor–product quadrature. *Computing*, 71(1):65–87, 2003.

[80] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: a Spectral Approach*. Dover Civil and Mechanical Engineering, Courier Dover Publications, Springer-Verlag, New York, 1991.

[81] M.B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[82] M.B. Giles and E. Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11(1):145–236, 2002.

[83] C. Gittelson and C. Schwab. Sparse tensor discretizations of high-dimensional pdes. *Acta Numerica*, 2011.

[84] R. Glowinski and J.L. Lions. *Exact and approximate controllability for distributed parameter systems*. Cambridge University Press, Cambridge, UK, 1996.

[85] M.A. Grepl and M. Kärcher. Reduced basis a posteriori error bounds for parametrized linear-quadratic elliptic optimal control problems. *Comptes Rendus Mathematique*, 349(15):873–877, 2011.

[86] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Patera. Efficient reduced-basis treatment of non-affine and nonlinear partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(03):575–605, 2007.

[87] M.A. Grepl and A.T. Patera. A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 39(01):157–181, 2005.

[88] M. Griebel. Sparse grids and related approximation schemes for higher dimensional problems. In *Proceedings of the conference on Foundations of Computational Mathematics*, Santander, Spain, 2005.

[89] S. Gugercin and A.C. Antoulas. A survey of model reduction by balanced truncation and some new results. *International Journal of Control*, 77(8):748–766, 2004.

[90] M. Gunzburger and A. Labosvky. An efficient and accurate numerical method for high-dimensional stochastic partial differential equations. *submitted*, 2012.

[91] M.D. Gunzburger, H.C. Lee, and J. Lee. Error estimates of stochastic optimal neumann boundary control problems. *SIAM Journal on Numerical Analysis*, 49:1532–1552, 2011.

[92] M.D. Gunzburger and S. Manservisi. Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control. *SIAM Journal on Numerical Analysis*, 37(5):1481–1512, 2000.

[93] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(02):277–302, 2008.

[94] M. Heinkenschloss and D. Leykekhman. Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. *SIAM Journal on Numerical Analysis*, 47(6):4607–4638, 2010.

[95] S. Heinrich. Multilevel Monte Carlo methods. *Large-Scale Scientific Computing*, 2179:58–67, 2001.

[96] J.S. Hesthaven and S. Zhang. On the use of ANOVA expansions in reduced basis methods for high-dimensional parametric partial differential equations. *Brown Division of Applied Math Scientific Computing Tech Report*, 2011.

[97] M. Hinze, N. Yan, and Z. Zhou. Variational discretization for optimal control governed by convection dominated diffusion equations. *Journal of Computational Mathematics*, 27(2-3):237–253, 2009.

[98] M. Holtz. *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*. Dissertation, Institut für Numerische Simulation, Universität Bonn, 2008.

[99] L.S. Hou, J. Lee, and H. Manouzi. Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs. *Journal of Mathematical Analysis and Applications*, 384(1):87–103, 2011.

[100] X. Hu, G. Lin, T.Y. Hou, and P. Yan. An adaptive ANOVA-based data-driven stochastic method for elliptic PDE with random coefficients. *Technical Report 30, Applied and Computational Mathematics, California Institute of Technology*, 2012.

[101] D.B.P. Huynh, D.J. Knezevic, Y. Chen, J.S. Hesthaven, and A.T. Patera. A natural-norm successive constraint method for inf-sup lower bounds. *Computer Methods in Applied Mechanics and Engineering*, 199(29):1963–1975, 2010.

[102] D.B.P Huynh, G. Rozza, S. Sen, and A.T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *Comptes Rendus Mathematique, Analyse Numérique*, 345(8):473–478, 2007.

[103] R.L. Iman. *Latin hypercube sampling*. John Wiley & Sons, Hoboken, NJ, 2008.

[104] K. Ito and S.S. Ravindran. A reduced basis method for control problems governed by PDEs. In *Control and estimation of distributed parameter systems*, pages 153–168. Springer, 1998.

[105] J.D. Jakeman, R. Archibald, and D. Xiu. Characterization of discontinuities in high-dimensional stochastic problems on adaptive sparse grids. *Journal of Computational Physics*, 230(10):3977–3997, 2011.

[106] Junseok K. Phase-field models for multi-component fluid flows. *Communications in Computational Physics*, 12(3):613–661, 2012.

[107] M. Kärcher and M. Grepl. A certified reduced basis method for parametrized elliptic optimal control problems. *Accepted in ESAIM: Control, Optimisation and Calculus of Variations*, 2013.

[108] K. Karhunen. Über lineare methoden in der wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae, Series A.I. Mathematica-Phys.*, 37:1–79, 1947.

[109] M. Kleiber and T.D. Hien. *The stochastic finite element method*. John Wiley & Sons, Hoboken, NJ, 1992.

[110] A. Klimke. *Uncertainty modeling using fuzzy arithmetic and sparse grids. Universität Stuttgart*. PhD thesis, Universität Stuttgart, Germany, 2006.

[111] D.J. Knezevic and J.W. Peterson. A high-performance parallel implementation of the certified reduced basis method. *Computer Methods in Applied Mechanics and Engineering*, 200(13):1455–1466, 2011.

[112] D.P. Kouri, D. Heinkenschloos, M. Ridzal, and B.G. Van Bloemen Waanders. A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty. *SIAM Journal on Scientific Computing*, 35(4):1847–1879, 2012.

[113] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(1):1–23, 2008.

[114] A. Kunoth and C. Schwab. Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs. *ETH SAM Report*, 2011.

[115] T. Lassila, A. Manzoni, A. Quarteroni, and G. Rozza. Generalized reduced basis methods and n-width estimates for the approximation of the solution manifold of parametric PDEs. *Analysis and Numerics of Partial Differential Equations Series: Springer INdAM Series, Vol. 4, Brezzi, F.; Colli Franzone, P.; Gianazza, U.; Gilardi, G. (Eds.)*, 2013.

[116] T. Lassila, A. Manzoni, and G. Rozza. On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46:1555–1576, 2012.

[117] T. Lassila, A. Quarteroni, and G. Rozza. A reduced basis model with parametric coupling for fluid-structure interaction problems. *SIAM Journal on Scientific Computing*, 34(2):1187–1213, 2012.

[118] T. Lassila and G. Rozza. Parametric free-form shape design with PDE models and reduced basis method. *Computer Methods in Applied Mechanics and Engineering*, 199(23):1583–1592, 2010.

[119] OP Le Maître, OM Knio, HN Najm, and RG Ghanem. Uncertainty propagation using Wiener-Haar expansions. *Journal of Computational Physics*, 197(1):28–57, 2004.

[120] J. Li, J. Li, and D. Xiu. An efficient surrogate-based method for computing rare failure probability. *Journal of Computational Physics*, 230(24):8683–8697, 2011.

[121] J. Li and D. Xiu. Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, 229(23):8966–8980, 2010.

[122] C. Lieberman, K. Willcox, and O. Ghattas. Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing*, 32(5):2523–2542, 2010.

[123] J.L. Lions. *Optimal control of systems governed by partial differential equations*. Springer-Verlag, Berlin, 1971.

[124] M. Loève. *Probability Theory*, volume II. 4th ed. Graduate Texts in Mathematics 46, Springer-Verlag, New York, 1978.

[125] W.L. Loh. On Latin hypercube sampling. *The Annals of Statistics*, 24(5):2058–2080, 1996.

[126] G. Lube and G. Rapin. Residual-based stabilized higher-order fem for advection-dominated problems. *Computer methods in applied mechanics and engineering*, 195(33):4124–4138, 2006.

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

[128] X. Ma and N. Zabaras. An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations. *Journal of Computational Physics*, 229(10):3884–3915, 2010.

[129] Y. Maday, N.C. Nguyen, A.T. Patera, and G.S.H. Pau. A general, multipurpose interpolation procedure: the magic points. *Communications on Pure and Applied Analysis*, 8(1):383–404, 2009.

[130] Y. Maday, A.T. Patera, and G. Turinici. Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. *Comptes Rendus Mathematique*, 335(3):289–294, 2002.

[131] Y. Maday, A.T. Patera, and G. Turinici. A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. *Journal of Scientific Computing*, 17(1):437–446, 2002.

[132] A. Manzoni. *Reduced models for optimal control, shape optimization and inverse problems in haemodynamics*. PhD thesis, EPFL, Lausanne, 2012.

[133] A. Manzoni, T. Lassila, A. Quarteroni, and G. Rozza. A reduced-order strategy for solving inverse Bayesian shape identification problems in physiological flows. In *5th International Conference on High Performance Scientific Computing - Modeling, Simulation and Optimization of Complex Processes*, Hanoi, Vietnam, in press, 2014. Springer Heildeberg.

[134] A. Manzoni, A. Quarteroni, and G. Rozza. Model reduction techniques for fast blood flow simulation in parametrized geometries. *International Journal for Numerical Methods in Biomedical Engineering*, 28(6-7):604–625, 2012.

[135] A.L. Marsden. Optimization in cardiovascular modeling. *Annual Review of Fluid Mechanics*, 46:519–546, 2014.

[136] Y.M. Marzouk and H.N. Najm. Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[137] Y.M. Marzouk, H.N. Najm, and L.A. Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, 2007.

[138] H.G. Matthies and A. Keese. Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 194(12-16):1295–1331, 2005.

[139] G.J. McRae, M.A. Tatang, et al. *Direct incorporation of uncertainty in chemical and environmental engineering systems*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, US, 1995.

[140] G. Migliorati, F. Nobile, E. Von Schwerin, and R. Tempone. Approximation of quantities of interest in stochastic PDEs by the random discrete L^2 projection on polynomial spaces. *SIAM Journal on Scientific Computing*, 35(3):A1440–A1460, 2013.

[141] B. Moore. Principal component analysis in linear systems: controllability, observability, and model reduction. *Automatic Control, IEEE Transactions on*, 26(1):17–32, 1981.

[142] F. Negri. Reduced basis method for parametrized optimal control problems governed by PDEs, Master Thesis, EPFL. 2011.

[143] F. Negri, G. Rozza, and A. Manzoni. Certified reduced basis method for parametrized optimal control problems governed by Stokes equations. *submitted*, 2013.

[144] F. Negri, G. Rozza, A. Manzoni, and A. Quarteroni. Reduced basis method for parametrized elliptic optimal control problems. *SIAM Journal on Scientific Computing*, 35(5):A2316–A2340, 2013.

[145] N.C. Nguyen, G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for parametrized parabolic PDEs; application to real-time Bayesian parameter estimation. *Biegler, Biros, Ghattas, Heinkenschloss, Keyes, Mallick, Tenorio, van Bloemen Waanders, and Willcox, editors, Computational Methods for Large Scale Inverse Problems and Uncertainty Quantification, John Wiley & Sons, UK*, 2009.

[146] N.C. Nguyen, G. Rozza, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for the time-dependent viscous burgers' equation. *Calcolo*, 46(3):157–185, 2009.

[147] H. Niederreiter. *Quasi-Monte Carlo Methods*. John Wiley & Sons, Hoboken, NJ, 1992.

[148] F. Nobile, R. Tempone, and C.G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2411–2442, 2008.

[149] F. Nobile, R. Tempone, and C.G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

[150] A.K. Noor and J.M. Peters. Reduced basis technique for nonlinear analysis of structures. *AIAA Journal*, 18(4):455–462, 1980.

[151] A. Nouy. A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 196(45-48):4521–4537, 2007.

[152] A. Nouy. Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. *Archives of Computational Methods in Engineering*, 16(3):251–285, 2009.

[153] A. Nouy. Proper generalized decompositions and separated representations for the numerical solution of high dimensional stochastic problems. *Archives of Computational Methods in Engineering*, 17(4):403–434, 2010.

[154] A. Nouy, A. Clement, F. Schoefs, and N. Moës. An extended stochastic finite element method for solving stochastic partial differential equations on random domains. *Computer Methods in Applied Mechanics and Engineering*, 197(51-52):4663–4682, 2008.

[155] J.T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Computers & Mathematics with Applications*, 41(5):735–756, 2001.

[156] J.T. Oden and K.S. Vemaganti. Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials: I. error estimates and adaptive algorithms. *Journal of Computational Physics*, 164(1):22–47, 2000.

[157] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, Heidelberg, New York, 2010.

[158] A.T. Patera and G. Rozza. Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations. *Copyright MIT, http://augustine.mit.edu*, 2007.

[159] T.N.L. Patterson. The optimum addition of points to quadrature formulae. *Mathematics of Computation*, 22(104):847–856, 1968.

[160] A. Pinkus. *N-widths in Approximation Theory*. Springer, Berlin, 1985.

[161] A. Quarteroni. *Numerical Models for Differential Problems*. Springer, Milano, MS&A, vol. 8, 2013.

[162] A. Quarteroni and G. Rozza. Numerical solution of parametrized Navier–Stokes equations by reduced basis methods. *Numerical Methods for Partial Differential Equations*, 23(4):923–948, 2007.

[163] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1(1):1–49, 2011.

[164] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Springer, New York, 2007.

[165] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer, Berlin and New York, 1994.

[166] R. Rackwitz. Reliability analysis – a review and some perspectives. *Structural Safety*, 23(4):365–395, 2001.

[167] S.S. Ravindran. A reduced-order approach for optimal control of fluids using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, 34(5):425–448, 2000.

[168] T. Rees, H.S. Dollar, and A.J. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010.

[169] T. Rees and A.J. Wathen. Preconditioning iterative methods for the optimal control of the Stokes equations. *SIAM Journal on Scientific Computing*, 33(5):2903–2926, 2011.

[170] F. Riesz and B. Sz.-Nagy. *Functional Analysis*. Dover Publications, New York, 1990.

[171] C.P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 2. Springer, New York, 1999.

[172] E. Rosseel and G.N. Wells. Optimal control with stochastic PDE constraints and uncertain controls. *Computer Methods in Applied Mechanics and Engineering*, 213 - 216(0):152 – 167, 2012.

[173] D.V. Rovas, L. Machiels, and Y. Maday. Reduced-basis output bound methods for parabolic problems. *IMA journal of numerical analysis*, 26(3):423–445, 2006.

[174] G. Rozza. Reduced-basis methods for elliptic equations in sub-domains with a posteriori error bounds and adaptivity. *Applied Numerical Mathematics*, 55(4):403–424, 2005.

[175] G. Rozza. *Shape design by optimal flow control and reduced basis techniques: Applications to bypass configurations in haemodynamics*. PhD thesis, EPFL, Lausanne, 2005.

[176] G. Rozza. Reduced basis methods for Stokes equations in domains with non-affine parameter dependence. *Computing and Visualization in Science*, 12(1):23–35, 2009.

[177] G. Rozza, D.B.P. Huynh, and A. Manzoni. Reduced basis approximation and a posteriori error estimation for Stokes flows in parametrized geometries: roles of the inf-sup stability constants. *Numerische Mathematik*, 125(1):1–38, 2013.

[178] G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):229–275, 2008.

[179] G. Rozza, A. Manzoni, and F. Negri. Reduced strategies for pde-constrained optimization problems in haemodynamics. In *Proceedings of the 6th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), Vienna, Austria*, 2012.

[180] G. Rozza and K. Veroy. On the stability of the reduced basis method for stokes equations in parametrized domains. *Computer methods in applied mechanics and engineering*, 196(7):1244–1260, 2007.

[181] Y. Saad. *Iterative methods for sparse linear systems*, volume 620. PWS publishing company Boston, 1996.

[182] A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity Analysis*, volume 134. John Wiley & Sons, Hoboken, NJ, 2000.

[183] T.P. Sapsis and P.F.J. Lermusiaux. Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D: Nonlinear Phenomena*, 238(23):2347–2360, 2009.

[184] T.P. Sapsis and P.F.J. Lermusiaux. Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D: Nonlinear Phenomena*, 241(1):60–76, 2012.

[185] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3):752–773, 2007.

[186] W. Schoutens. *Stochastic processes and orthogonal polynomials*. Springer-Verlag. New York, 2000.

[187] G.I. Schuëller, H.J. Pradlwarter, and P.S. Koutsourelakis. A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic Engineering Mechanics*, 19(4):463–474, 2004.

[188] C. Schwab and A.M. Stuart. Sparse deterministic approximation of bayesian inverse problems. *Inverse Problems*, 28(4):045003, 2012.

[189] C. Schwab and R. A. Todor. Karhunen–Loève approximation of random fields by generalized fast multipole methods. *Journal of Computational Physics*, 217(1):100–122, 2006.

[190] C. Schwab and R.A. Todor. Sparse finite elements for elliptic problems with stochastic loading. *Numerische Mathematik*, 95(4):707–734, 2003.

[191] S.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Doklady Akademii Nauk SSSR*, volume 4, pages 240–243, 1963.

[192] I.M. Sobol. Theorems and examples on high dimensional model representation. *Reliability Engineering & System Safety*, 79(2):187–193, 2003.

[193] G. Stefanou. The stochastic finite element method: past, present and future. *Computer Methods in Applied Mechanics and Engineering*, 198(9-12):1031–1051, 2009.

[194] A.M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19(1):451–559, 2010.

[195] T. Taddei, S. Perotto, and A. Quarteroni. Reduced basis techniques for nonlinear conservation laws. *MOX–Report No. 32, Politecnico di Milano*, 2013.

[196] Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127–204, 2011.

[197] H. Tiesler, R.M. Kirby, D. Xiu, and T. Preusser. Stochastic collocation for optimal control problems with stochastic PDE constraints. *SIAM Journal on Control and Optimization*, 50(5):2659–2682, 2012.

[198] R.A. Todor and C. Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA Journal of Numerical Analysis*, 27(2):232, 2007.

[199] L.N. Trefethen. Is Gauss quadrature better than Clenshaw-Curtis? *SIAM Review*, 50(1):67–87, 2008.

[200] F. Tröltzsch. *Optimal control of partial differential equations: theory, methods, and applications*, volume 112. American Mathematical Society, Providence, RI, 2010.

[201] K. Urban and B. Wieland. Affine decompositions of parametric stochastic processes for application within reduced basis methods. In *Proceedings MATHMOD, 7th Vienna International Conference on Mathematical Modelling (accepted)*, 2012.

[202] J. Walsh. An introduction to stochastic partial differential equations. *École d'Été de Probabilités de Saint Flour XIV-1984*, pages 265–439, 1986.

[203] X. Wan and G.E. Karniadakis. Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM Journal on Scientific Computing*, 28(3):901–928, 2007.

[204] K. Willcox and J. Peraire. Balanced model reduction via the proper orthogonal decomposition. *AIAA journal*, 40(11):2323–2330, 2002.

[205] D. Xiu. Fast numerical methods for stochastic computations: a review. *Communications in Computational Physics*, 5(2-4):242–272, 2009.

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

[207] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3):1118–1139, 2005.

[208] D. Xiu and G.E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2003.

[209] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numerische Mathematik*, 94(1):195–202, 2003.

[210] F. Yamazaki. Neumann expansion for stochastic finite element analysis. *Journal of Engineering Mechanics*, 114(8):1335–1354, 1988.

[211] N. Zabaras and B. Ganapathysubramanian. A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach. *Journal of Computational Physics*, 227(9):4697–4735, 2008.

[212] S. Zhang. Efficient greedy algorithms for successive constraints methods with high-dimensional parameters. *Brown Division of Applied Math Scientific Computing Tech Report*, 23, 2011.

Peng Chen

Curriculum Vitae

Personal Details

Date of Birth	November 12th, 1986
Nationality	Chinese, born in Peixian, Jiangsu
Address	EPFL-SB-MATHICSE-CMCS, Av. Piccard, Station 8, CH-1015 Lausanne, Switzerland
Email & Phone	peng.chen@epfl.ch, +41 (0)78 9241158 (mobile), +41 (0)21 69 32733 (office)
Homepage	`http://cmcs.epfl.ch/people/chen`

Academic Education

Mar.11- Feb.14	**Ph.D. in Applied Mathematics, EPFL, Switzerland**.
Thesis:	*Model Order Reduction Techniques for Uncertainty Quantification Problems*.
Advisor:	*Prof. Alfio Quarteroni. Co-advisor: Dr. Gianluigi Rozza*.
Sep.09 - Feb.11	**Master in Mathematical Sciences, EPFL, Switzerland**.
Thesis:	*The Lattice Boltzmann Method for Fluid Dynamics: Theory and Applications*.
Advisor:	*Prof. Alfio Quarteroni. Co-advisor: Dr. Matteo Astorino*.
Sep.05 - Jul.09	**Bachelor in Mathematics and Applied Mathematics, XJTU, China**.
Thesis:	*Stochastic Differential Equations and Their Applications in Finance*.
Advisor:	*Prof. Jiuquan Ren*.

Publications

preprint

[J9] P. Chen, A. Quarteroni. *A new algorithm for high-dimensional uncertainty quantification problems based on dimension-adaptive and reduced basis methods.* Submitted, 2014

[J8] P. Chen, A. Quarteroni, G. Rozza. *Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations*. Submitted, 2013.

[J7] P. Chen, A. Quarteroni. *Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraints*. To appear in SIAM/ASA Journal on Uncertainty Quantification, 2014.

published

[J6] P. Chen, A. Quarteroni, G. Rozza. *A weighted empirical interpolation method: A priori convergence analysis and applications*. ESAIM: Mathematical Modelling and Numerical Analysis, in press, online doi: 10.1051/m2an/2013128, 2013.

[J5] P. Chen, A. Quarteroni, G. Rozza. *Comparison of reduced basis and stochastic collocation methods for elliptic problems*. Journal of Scientific Computing, 59:187-216, 2014.

[J4] P. Chen, A. Quarteroni, G. Rozza. *Stochastic optimal robin boundary control problems of advection-dominated elliptic equations*. SIAM Journal on Numerical Analysis, 51(5):2700-2722, 2013.

[J3] P. Chen, A. Quarteroni. *Accurate and efficient evaluation of failure probability for partial differential equations with random input data*. Computer Methods in Applied Mechanics and Engineering, 267(0):233-260, 2013

[J2] P. Chen, A. Quarteroni, G. Rozza. *Simulation-based uncertainty quantification of human arterial network hemodynamics*. International Journal for Numerical Methods in Biomedical Engineering, 29(6):698-721, 2013.

[J1] P. Chen, A. Quarteroni, G. Rozza. *A weighted reduced basis method for elliptic partial differential equation with random input data*. SIAM Journal on Numerical Analysis, 51(6):3163-3185, 2013.

Conferences/Workshops/Seminars

Invited and contributed presentations in conferences and symposiums

[C5] **SIAM UQ14**, SIAM Conference on Uncertainty Quantification, Savannah, USA, March 31 - April 03, 2014.

Invited minisymposium talk: Weighted reduced basis method for optimal control problems constrained by stochastic PDEs.

[C4] **SIAM UQ14**, SIAM Conference on Uncertainty Quantification, Savannah, USA, March 31 - April 03, 2014.

Invited minisymposium talk: Reduced basis method and several extensions for uncertainty quantification problems.

[C3] **DDMOPDEC 2013**, Domain Decomposition Methods for Optimization with PDE Constraints, Monte Verita, Ascona, Switzerland, 01-06, September, 2013.

Invited minisymposium talk: Weighted reduced basis method for stochastic optimal control problems with PDE constraints.

[C2] **ENUMATH 2013**, European Conference on Numerical Mathematics and Advanced Applications, EPFL, Lausanne, Switzerland, 26-30, August, 2013.

Contributed talk: A weighted reduced basis method for elliptic partial differential equations with random input data.

[C1] **MPF 2013**, International Symposium on Modelling of Physiological Flows, Chia Laguna, Italy, 11-14 June, 2013.

Contributed talk: Uncertainty quantification of human arterial system.

Invited and contributed presentations in workshops and colloquiums

[W5] **NMUQ 2013**, Workshop: Numerical Methods for Uncertainty Quantification, Bonn University, Bonn, Germany, 13-17 May, 2013.

Poster presentation: Reduced basis methods for reliability analysis.

[W4] **SNC 2013**, Swiss Numerics Colloquium, EPFL, Lausanne, Switzerland, 05 April, 2013.

Contributed talk: Accurate and efficient evaluation of failure probability for partial differential equations with random inputs.

[W3] **WUQ 2012**, Workshop on Uncertainty Quantification, ICERM, Brown University, Providence, USA, 09-13 October, 2012.

Invited poster presentation: Uncertainty quantification of human arterial network.

[W2] **CECAM 2012**, Workshop on Reduced Basis, POD and Reduced Order Methods for model and computational reduction: towards real-time computing and visualization? EPFL, Lausanne, Switzerland, 14-16 May 2012.

Invited poster presentation: Comparison of reduced basis method and stochastic collocation method for stochastic elliptic problems.

[W1] **SNC 2012**, Swiss Numerics Colloquium, University of Bern, Bern, Switzerland, 13 April, 2012.

Poster presentation: Stochastic optimal Robin boundary control problems constrained by an advection-dominated elliptic equation.

Invited seminars

[S2] **SAM-ETHZ 2014**, Seminar for Applied Mathematics, Swiss Federal Institute of Technology in Zurich, Zurich, 10 March, 2014.

Invited seminar: Reduced basis methods for uncertainty quantification problems.

[S1] **ICMSEC 2013**, The Institute of Computational Mathematics and Scientific/Engineering Computing of Chinese Academy of Science, Beijing, 06 May, 2013.

Invited seminar: Reduced basis methods and several extensions for uncertainty quantification problems.

Other attended conferences/workshops/winter and summer schools

[C] **ICCS 2012**, International conference on computational sciences, on the Occasion of Professor Benyu Guo's 70th Birthday, Shanghai Normal University, Shanghai, China, 16-20 July, 2012.

[School] **MCSE 2012**, KAUST-CIMPA School in Applied Mathematics on Uncertainty Quantification, KAUST, Saudi-Arabia, 04-12 January, 2012.

[School] **RISM 2011**, Riemann International School of Mathematics, Multiphase and Multiphysics Problems, Verbania, Italy, 25-30 September, 2011.

Honours and Awards

2013 SIAM Student Travel Award

2009 – 2011 EPFL Excellence Fellowships.

2009 Excellent Student of Xi'an Jiaotong University.

2008 Meritorious Winner of Mathematical Contest in Modeling, SIAM.

2007 First Prize of China Undergraduate Mathematical Contest in Modeling, CSIAM.

2006, 2008 First Class National Scholarship, Ministry of Education, China.

2007 Kang PENG (President) Scholarship, Xi'an Jiaotong University, China.