

A posteriori error estimation for partial differential equations with random input data

THÈSE N° 7260 (2016)

PRÉSENTÉE LE 2 NOVEMBRE 2016

À LA FACULTÉ DES SCIENCES DE BASE

CALCUL SCIENTIFIQUE ET QUANTIFICATION DE L'INCERTITUDE - CHAIRE CADMOS

CHAIRE D'ANALYSE ET DE SIMULATION NUMÉRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Diane Sylvie GUIGNARD

acceptée sur proposition du jury:

Prof. K. Hess Bellwald, présidente du jury
Prof. F. Nobile, Prof. M. Picasso, directeurs de thèse
Prof. G. Rozza, rapporteur
Prof. S. Prudhomme, rapporteur
Prof. J. Hesthaven, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

To my beloved family:
René, Christine, Simon, Aline.

Acknowledgements

A doctoral thesis is a beautiful journey, although long and tough, that is now coming to an end. This accomplishment would not have been possible without the support of many people that I would like to thank here. My apologies to those I will forget.

First of all, I am truly indebted to my two supervisors Prof. Marco Picasso and Prof. Fabio Nobile. Bringing together your respective areas of expertise, you gave me the chance to work on very interesting subjects. I am extremely grateful to both of you for your constant support, your availability and also for your confidence in me, allowing me a great deal of freedom. It has been a real pleasure to work with you during these years and everything I have learned from you is priceless.

Next, I would like to thank Prof. Gianluigi Rozza, Prof. Serge Prudhomme and Prof. Jan Hesthaven who honoured me by being part of my jury and to Prof. Kathryn Hess Bellwald for presiding it.

A nice working environment is, to my mind, a key ingredient to succeed in the completion of a Ph.D. thesis, reason why I express my gratitude to all my colleagues from EPFL. Special thanks to my office mates Gilles Steiner, Jonas Ballani, Sylvain Vallaghé and Alexandre Caboussat, as well as to the (past and present) members of CSQI and GR-PI Laurent Michel, Stéphane Flotron, Michel Flueck, Viljami Laurmaa, Jonathan Rochat, Thomas Foetisch, Samuel Dubuis, Francesca Bonizzoni, Lorenzo Tamellini, Francesco Tesei, Eleonora Musharbash, Michele Pisaroni, Sebastian Krumscheid, Matthieu Martin, ...

I would like to warmly thank Monica, Jonathan, Thierry, Younes, Dimitri, Rafael, Nicolas and Valentin who made my studies at EPFL particularly pleasant. I also thank all the students for whom I had the fortune to be the teaching assistant; I deeply enjoyed this role.

I am profoundly grateful to all my friends outside EPFL, especially those from my neighbourhood and from gymnastics which play a major role in my life. Anjuli, Evelyne, Christel, Norbert, Charlotte, Margaux, Barbara, Morena, just to name a few, you are like my second family.

Last but not least, I would like to thank from the bottom of my heart my beloved family, in particular my sister Aline, my brother Simon and my parents René and Christine to whom I dedicate this thesis.

Lausanne, September 2016

Diane Guignard

Abstract

This thesis is devoted to the derivation of error estimates for partial differential equations with random input data, with a focus on *a posteriori* error estimates which are the basis for adaptive strategies. Such procedures aim at obtaining an approximation of the solution with a given precision while *minimizing* the computational costs. If several sources of error come into play, it is then necessary to balance them to avoid unnecessary work.

We are first interested in problems that contain small uncertainties approximated by finite elements. The use of perturbation techniques is appropriate in this setting since only few terms in the power series expansion of the exact random solution with respect to a parameter characterizing the amount of randomness in the problem are required to obtain an accurate approximation. The goal is then to perform an error analysis for the finite element approximation of the expansion up to a certain order. First, an elliptic model problem with random diffusion coefficient with affine dependence on a vector of independent random variables is studied. We give both *a priori* and *a posteriori* error estimates for the first term in the expansion for various norms of the error. The results are then extended to higher order approximations and to other sources of uncertainty, such as boundary conditions or forcing term. Next, the analysis of nonlinear problems in random domains is proposed, considering the one-dimensional viscous Burgers' equation and the more involved incompressible steady-state Navier-Stokes equations. The *domain mapping method* is used to transform the equations in random domains into equations in a fixed reference domain with random coefficients. We give conditions on the mapping and the input data under which we can prove the well-posedness of the problems and give *a posteriori* error estimates for the finite element approximation of the first term in the expansion. Finally, we consider the heat equation with random Robin boundary conditions. For this parabolic problem, the time discretization brings an additional source of error that is accounted for in the error analysis.

The second part of this work consists in the analysis of a random elliptic diffusion problem that is approximated in the physical space by the finite element method and in the stochastic space by the stochastic collocation method on a sparse grid. Considering a random diffusion coefficient with affine dependence on a vector of independent random variables, we derive a residual-based *a posteriori* error estimate that controls the two sources of error. The stochastic error estimator is then used to drive an adaptive sparse grid algorithm which aims at alleviating the so-called *curse of dimensionality* inherent to tensor grids. Several numerical examples are given to illustrate the performance of the adaptive procedure.

Abstract

Key words: PDEs with random inputs, uncertainty quantification, *a priori* and *a posteriori* error analysis, finite elements, perturbation techniques, stochastic collocation, elliptic equations, steady Navier-Stokes equations, heat equation

Résumé

Cette thèse est consacrée à la dérivation d'estimations d'erreur pour des équations aux dérivées partielles contenant des données aléatoires. Un accent particulier est mis sur les estimateurs *a posteriori* qui sont à la base d'algorithmes adaptatifs. Ces derniers visent à obtenir une approximation de la solution avec une certaine précision tout en *minimisant* le coût du calcul. Lorsque plusieurs sources d'erreurs entrent en jeu, il est judicieux de les équilibrer afin d'éviter tout travail inutile.

Nous nous intéressons pour commencer à des problèmes contenant de petites incertitudes résolus par la méthode des éléments finis. Dans ce cas, l'utilisation de méthodes dites de perturbation est indiquée car une bonne approximation de la solution peut être obtenue avec peu de termes dans le développement en série de puissances de la solution exacte par rapport à un paramètre contrôlant le niveau d'incertitude du problème. Le but principal de ce travail est d'effectuer une analyse d'erreur pour l'approximation par éléments finis du développement à un certain ordre. Nous considérons pour commencer un problème modèle elliptique avec un coefficient de diffusion aléatoire qui dépend de manière affine d'un vecteur de variables aléatoires indépendantes. Des estimations d'erreur *a priori* et *a posteriori* sont données pour le premier terme dans le développement de la solution en considérant différentes normes de l'erreur. Les résultats obtenus sont alors généralisés pour des approximations d'ordres supérieurs ainsi que pour des problèmes contenant d'autres sources d'incertitudes, comme par exemple les conditions au bord ou le terme de force. L'étude se poursuit en considérant des problèmes non-linéaires définis sur des domaines aléatoires, tout d'abord l'équation de Burgers à une dimension d'espace puis les équations de Navier-Stokes stationnaires incompressibles. Les problèmes sont reformulés sur un domaine fixe de référence à l'aide d'une transformation introduisant alors des coefficients aléatoires dans les équations. Nous donnons des conditions sur la transformation et les données sous lesquelles les problèmes sont bien posés et nous donnons des estimations d'erreur pour le premier terme du développement. Finalement, nous considérons le problème de la chaleur avec des conditions au bord de type Robin qui contiennent des incertitudes. Pour ce problème parabolique, la discrétisation temporelle ajoute une source supplémentaire d'erreur qui est prise en compte dans l'analyse d'erreur.

Dans la deuxième partie de ce travail, nous analysons un problème de diffusion elliptique avec coefficient aléatoire résolu approximativement par la méthode des éléments finis en espace physique et par la méthode de collocation stochastique avec grille fine en espace stochastique. En considérant un coefficient de diffusion dépendant de manière affine d'un vecteur de

Résumé

variables aléatoires indépendantes, nous donnons un estimateur d'erreur *a posteriori* basé sur le résidu qui contrôle les deux sources d'erreur. L'estimateur contrôlant l'erreur stochastique est ensuite utilisé dans un algorithme construisant de manière adaptative une grille peu dense, permettant ainsi de palier au problème *du fléau de la dimension* dont souffrent les grilles de type tenseuriel. Plusieurs exemples numériques sont donnés pour illustrer les performances de l'algorithme adaptatif.

Mots clefs : EDP avec données aléatoires, quantification des incertitudes, analyse d'erreur *a priori* et *a posteriori*, éléments finis, technique de perturbation, collocation stochastique, équations elliptiques, Navier-Stokes stationnaire, équation de la chaleur

Contents

Acknowledgements	i
Abstract (English/Français)	iii
List of figures	ix
List of tables	xiii
Introduction	1
1 Elliptic model problems with random diffusion coefficient	7
1.1 Problem statement	9
1.2 Methodology	14
1.3 Error analysis for the first order approximation	16
1.3.1 <i>A priori</i> error analysis	17
1.3.2 <i>A posteriori</i> error analysis	22
1.4 Error analysis for higher order approximations	30
1.4.1 Second order approximation	30
1.4.2 Generalization	33
1.5 Extension to nonlinear problems	34
1.6 Computational costs	38
1.7 Numerical results	40
1.7.1 1D problems	40
1.7.2 2D problems	55
1.7.3 Comparison with the stochastic collocation method	57
1.A Derivation of problems (1.20), (1.21) and (1.22)	61
1.B Upper and lower bounds for the error $u - u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm	64
1.C Estimation of the interpolation constant	70
2 Elliptic model problems with other sources of uncertainty	73
2.1 Neumann random boundary conditions	73
2.2 Two sources of uncertainty	78
2.3 Numerical results	85

Contents

3	PDEs in random domains	91
3.1	Steady-state viscous Burgers' equation in random intervals	93
3.1.1	Deterministic case	94
3.1.2	Random case	98
3.1.3	Numerical results	104
3.2	Steady-state incompressible Navier-Stokes equations in random domains . . .	107
3.2.1	Problem statement	107
3.2.2	Formulation on a reference domain	108
3.2.3	Well-posedness of the problem	111
3.2.4	Specific form of the random mapping	116
3.2.5	<i>A posteriori</i> error analysis	119
3.2.6	Numerical results	128
3.A	Derivation of problems (3.58) and (3.59)	142
3.B	Choice of the norm	143
3.C	Proof of some properties	144
4	Time-dependent heat equation with random Robin boundary conditions	149
4.1	Problem statement	149
4.2	Numerical approximation	153
4.3	<i>A posteriori</i> error analysis	154
4.4	Numerical results	157
5	Error analysis for the stochastic collocation method	165
5.1	Problem statement	166
5.2	Stochastic collocation finite element method	167
5.3	Residual-based <i>a posteriori</i> error estimate	170
5.3.1	An abstract reformulation of the problem	174
5.4	Adaptive algorithms	176
5.5	Numerical results	180
5.A	Miscellaneous results	187
	Conclusions and perspectives	191
	Bibliography	201
	Curriculum Vitae	203

List of Figures

1.1	Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L_P^2(\Omega; H_0^1(D))$ -norm w.r.t h with $\varepsilon = 32h$	42
1.2	Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and u_h^1 in $L_P^2(\Omega; H_0^1(D))$ -norm w.r.t ε with $h = \varepsilon^2/32$	42
1.3	Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L_P^2(\Omega; L^2(D))$ -norm w.r.t h with ε fixed to $32h^2$	45
1.4	Five realizations of the random diffusion coefficient a given in (1.88) with $\varepsilon = 1$ (left) and the corresponding solution for $f = f_2$ (right).	46
1.5	Convergence rate for problem (1.11) with $f = f_1$ for $\varepsilon = 4, 1, 0.25, 0.0625$. Log log scale plot of the error in $L_P^2(\Omega; H_0^1(D))$ -norm w.r.t h	47
1.6	Repartition of the nodes for $\varepsilon = 1$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.01$ (bottom) in the case $T = 2$. Left: strategy (1.91), right: strategy (1.92).	52
1.7	Expected value (left) and standard deviation (right) of the solution with $\varepsilon = 0.5$ for the first example.	55
1.8	Plot of the functions a_j , $j = 1, \dots, 9$, constructed by tensorization of one-dimensional KL functions.	57
1.9	Expected value (left) and standard deviation (right) of the solution with $\varepsilon = 0.1$ for the second example.	58
1.10	Time to solve the linear problem (1.11) and the nonlinear problem (1.78) with accuracy of order 2 in ε using the SC-FEM and the <i>perturbation method</i>	59
1.11	Log log scale plot of the computational time w.r.t. the error in $L_P^2(\Omega; H_0^1(D))$ -norm using the SC-FEM with Smolyak and Clenshaw-Curtis abscissas and the <i>perturbation method</i>	60
1.12	Notation for an element K in \mathcal{T}_h (left) and illustration of the domains w_K (middle) and w_e (right).	65
1.13	Structured (left) and Delaunay (right) triangulations of D with $N = 16$	71
2.1	Six realizations of the random forcing term f given in (2.39) with $\delta = 0.5$ and $M = 6$ (left) and the corresponding solution (right).	86

List of Figures

2.2	Six realizations of the random forcing term f given in (2.39) with $\delta = 0.5$ and $M = 50$ (left) and the corresponding solution (right).	86
2.3	Seven realizations of the random forcing term f given in (2.40) with $\delta = 0.5$ and $M = 50$ (left) and the corresponding solution (right).	87
3.1	Illustration and notation for the domain mapping approach.	92
3.2	Two strategies s_1 and s_2 for the (strong) formulation on the reference domain.	92
3.3	Function f and corresponding solution u for various values of s	105
3.4	Function g and corresponding solution u for various values of s	106
3.5	Geometry with prescribed boundary conditions for the first example.	129
3.6	Functions $\varphi_1(\xi_1)$, $\xi_1 \in [0, 2.2]$ (left) and $\varphi_2(\xi_2)$, $\xi_2 \in [0, 0.41]$ (right) defined in (3.79).	130
3.7	Velocity magnitude, components u_1 and u_2 and pressure for the first problem in the case $\varepsilon = 0$ and $\nu = 0.001$	131
3.8	Velocity magnitude for $\nu = 0.001$ in the case $\varepsilon = 0$ (top) and $\varepsilon = 0.05$ with $Y = 1$ computed on D_ω (middle) and on D (bottom) for the first example.	132
3.9	Effectivity index with respect to the viscosity ν for the two error estimators η and $\hat{\eta}$ defined in (3.73) and (3.77) for the first example.	134
3.10	Geometry with prescribed boundary conditions for the second example.	135
3.11	Function $g = g(\xi_1, \xi_2)$ defined in (3.81).	136
3.12	From left to right: velocity magnitude, components u_1 and u_2 and pressure for the second problem in the case $\varepsilon = 0$ and $\nu = 0.05$	137
3.13	Vorticity of the velocity and pressure for $\nu = 0.05$ in the case $\varepsilon = 0$ (left) and $\varepsilon = 0.01$ with $Y = 1$ computed D_ω (middle) and on D (right) for the second example.	138
3.14	Effectivity index with respect to the viscosity ν for the two error estimators η and $\hat{\eta}$ defined in (3.73) and (3.77) for the second example.	138
3.15	Vorticity of the velocity and pressure for $\nu = 0.05$ in the case $\varepsilon = 0$ (left) and $\varepsilon = 0.01$ with $Y = 1$ computed D_ω (middle) and on D (right) for the second example with $L = 2$	140
4.1	Geometry with label for each part of the boundary.	152
5.1	Non-downward closed set (left), downward closed set (middle) and multi-index set with its margin and reduced margin (right).	170
5.2	Geometry of the problem (left), expected value (middle) and standard deviation (right) of the solution for the case $\gamma_1 = \gamma_2 = 1$	181
5.3	Evolution of I during the adaptive process for the case $\gamma_1 = \gamma_2 = 1$. From left to right and top to bottom: iterations 3,5,8,10,14 and order of selection of the multi-indices.	181
5.4	Final multi-index set I (left), final sparse grid (middle) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = \gamma_2 = 1$	182

5.5	Final multi-index set I (left), final sparse grid (middle) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = 1$ and $\gamma_2 = 0$	183
5.6	Evolution of the multi-index set I during the adaptive process for the case $\gamma_1 = 1$ and $\gamma_2 = 0.1$. From left to right and top to bottom: iterations 4,6,8 and order of selection of the multi-indices.	183
5.7	Final multi-index set I (left) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = 1$ and $\gamma_2 = 0.1$	184
5.8	Geometry of the problem for $N = 8$ with indication of the coefficients γ_n , $n = 1, \dots, 8$ (left) and error and estimator with respect to the number of points in logarithmic scale for the 55 first iterations (right).	184
5.9	Projection of the multi-index set I obtained after 55 iterations on (y_1, y_4) (left), (y_1, y_5) (middle) and (y_1, y_7) (right).	185
5.10	Error and estimator with respect to the number of points in logarithmic scale (left) and projection of the final multi-index set on (y_1, y_3) (right) for the case $N = 3$	186
5.11	Error and estimator with respect to the number of points in logarithmic scale (left) and projection of the final multi-index set on (y_1, y_5) (right) for the case $N = 5$	186

List of Tables

1.1	Computational costs for the SC-FEM and the <i>perturbation method</i>	40
1.2	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ for $h = 2^{-7}$ and various ε for both cases f_1 and f_2	43
1.3	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ for $\varepsilon = 0.25$ and various $h = 1/N$ for both cases f_1 and f_2	44
1.4	Error and estimators for the approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$ with h fixed (top) and ε fixed (bottom) for the case $f = f_2$	44
1.5	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ for $h = 2^{-8}$ (top) and $\varepsilon = 0.5$ (bottom).	46
1.6	Value of $h = N_h^{-1}$ with respect to ε such that (1.89) holds with $T = 2$	48
1.7	Adaptive partition of D such that (1.89) holds with $T = 2$ when criterion (1.91) is used.	51
1.8	Adaptive partition of D such that (1.89) holds with $T = 2$ when criterion (1.92) is used.	51
1.9	Dörfler strategy such that $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ holds with $T = 2$ and $\theta = 0.5$	53
1.10	Dörfler strategy such that $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ holds with $T = 2$ in the case $\varepsilon = 0.01$	54
1.11	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ with $n = 64$ for the first example.	56
1.12	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ with $\varepsilon = 0.5$ for the first example.	56
1.13	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ with $n = 128$ for the second example.	58
1.14	Error $\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$, estimators η_1, η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\ u - u_{0,h}\ _{L_p^2(\Omega; H_0^1(D))}$ with $\varepsilon = 0.05$ for the second example.	58
1.15	Error, estimator and effectivity index for the Poisson problem.	71
2.1	Efficiency of the two error estimator η and $\hat{\eta}$ for the case (2.39) with $h = 2^{-7}$ ($\eta_h = 7.8125e-3$).	87
2.2	Efficiency of the two error estimator η and $\hat{\eta}$ for the case (2.40) with $h = 2^{-7}$ ($\eta_h = 7.8125e-3$).	88
2.3	for $h = 2^{-5}$ ($\eta_h = 3.125e-2$)	89
2.4	for $h = 2^{-7}$ ($\eta_h = 7.8125e-3$)	89

List of Tables

3.1	Error, estimator and effectivity index for the deterministic Burgers' equation with mesh size $2^{-2} \leq h \leq 2^{-10}$	105
3.2	Error, estimators and effectivity index for the Burgers' equation in random intervals for the first case f with $\varepsilon = 0.005$ and 0.00125	106
3.3	Error, estimators and effectivity index for the Burgers' equation in random intervals for the second case g with $\varepsilon = 0.01$ and 0.0025	106
3.4	Error, error estimator and effectivity index for the deterministic case ($\varepsilon = 0$) and various viscosities for the first example.	132
3.5	The error, the two contributions η_h and η_ε of the error estimator in (3.73) and the effectivity index for $\nu = 0.001$ and $\nu = 1$ for the first example.	133
3.6	Error, error estimator and effectivity index for the deterministic case ($\varepsilon = 0$) and various viscosities for the second example.	137
3.7	The error, the two contributions η_h and η_ε of the estimator in (3.73) and the effectivity index for $\nu = 0.05$ and $\nu = 1$	139
3.8	Effectivity index of the two error estimators in the case $\nu = 0.05$ for the second example with $L = 2$	141
4.1	Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0$	159
4.2	Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0$	159
4.3	Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.4$	160
4.4	Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.2$	161
4.5	Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.1$	161
4.6	Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0.5$	162
4.7	Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0.25$	162

Introduction

Partial differential equations (PDEs) are widely used for modelling problems in many fields such as physics, biology or engineering. Nowadays, uncertainty is often included in mathematical models arising from the simulation of complex systems. The uncertainty can reflect an intrinsic variability of the system (aleatory uncertainty) or our inability to adequately characterize all the inputs (epistemic uncertainty), due for instance to experimental measurements. It can occur in the coefficients, the forcing term, the geometry, the boundary conditions, the initial condition or combinations of them. A possible way to describe the uncertainties present in the model is to use a probability framework. In such a setting, the uncertain input data are characterized with random variables, or more generally random fields, yielding PDEs with random input. In a forward uncertainty quantification (UQ) problem, the goal is then to determine the effect of the uncertainty on the solution or a specific quantity of interest.

Several methods have been developed to tackle the numerical approximation of such problems in both the deterministic and, more recently, the stochastic variables. We give a short overview of the available methods, pointing to some references for an in-depth description, but we have no pretension to be exhaustive.

The best known and most commonly used methods for solving deterministic problems numerically are the finite difference [112, 117], the finite element [31, 49, 61] and the finite volume [85] methods, for which the theory is at a mature stage. Many other methods have been developed, either new methods or extension of the ones mentioned above, such as discontinuous Galerkin [105], boundary element [30], meshless [86] or extended finite element methods [81]. The selection of the method depends upon the type of problems to solve: elliptic, parabolic or hyperbolic.

For the approximation of random PDEs, the most popular method is certainly the Monte-Carlo method (see [63] for instance) which consists in solving the equations for i.i.d. realizations of the random input. The main drawback of this method is its well-known slow convergence rate with respect to the sample size K , namely of $\mathcal{O}(1/\sqrt{K})$. However, the convergence is independent of the dimension of the random space and this method is very easy to use in practice. To improve the convergence rate of the method, some extensions have been introduced such as the quasi-Monte Carlo [54, 55] and the multi-level Monte-Carlo [68] methods. Other than MC type methods, we mention the stochastic spectral methods comprising the Stochastic

Galerkin (SG) [67, 90] and the Stochastic Collocation (SC) [7, 97, 124] methods. These methods exploit the possible regularity of the solution with respect to the random input combining the generalized Polynomial Chaos (gPC) expansion of the solution with a Galerkin projection or an interpolation procedure. Finally, in the framework of PDEs with small uncertainties, the perturbation or Neumann series expansion methods [6, 37, 82, 127] appear to be an appropriate choice. For all these methods, an approximation in the physical space can be obtained using any deterministic method mentioned above. In particular, in this thesis we focus on the finite element method.

When a numerical method is used to solve a problem for which the exact solution is not at hand, approximation errors are introduced. An error analysis should then be done to appropriately estimate the various sources of error. In an *a priori* error analysis, the convergence of the method is assessed under suitable regularity assumptions on the exact solution. The *a priori* error estimate gives useful information about the asymptotic behaviour of the numerical approximation when the various discretization parameters vary. However, this theoretical bound usually depend on the unknown solution and is thus not a computable quantity. In *a posteriori* error analysis, the goal is to provide computable error estimators that depend only on the numerical approximation and the input data and that are localized in space. Having such error estimators available can be necessary in many situations. Indeed, if the solution presents local features evolving at fine scale, such as shocks, boundary layers or singularities due to re-entrant corners in physical space, very fine approximation spaces are required to capture them. However, this becomes quickly numerically unaffordable due to the limitations in computer power and memory. A remedy is then to use adaptive strategies based on a (reliable and efficient) *a posteriori* error estimator, refining only where needed, to get satisfactory accuracy in the approximation while limiting the computational effort. When several sources of error are affecting the numerical solution, the estimator should also furnish an estimation of the contribution of each error component to the total error, so that it can be used for balancing the errors.

The derivation of *a posteriori* error estimate controlling the finite element error started in the late seventies with the work by Babuška and Rheinboldt [8], where a residual-based error estimate is derived. Since then, many different types of *a posteriori* error estimates for the FEM have been introduced, such as error estimators obtained by solving local problems [1, 41, 83] or hierarchical [14], post-processed [128] and goal-oriented [13, 22, 100] error estimators, just to mention a few. We refer to Verfürth [118], Ainsworth and Oden [3] or Grätsch and Bathe [73] for a review of these different *a posteriori* error estimation techniques. Concerning the error estimation of methods for solving random PDEs, *a posteriori* error estimators in the *energy* norm for the stochastic Galerkin finite element method (SG-FEM) are derived in [24, 58, 59], where adaptive refinement algorithms are proposed for both stochastic and physical spaces. In the algorithm proposed in [59], the refined mesh is the same for all generalized polynomial chaos (gPC) modes, contrary to the one in [58] where the refinement procedure is applied independently for each mode. In [24], the adaptive procedure is driven by the two-sided estimates the authors obtained for the error reduction when the finite element subspace,

respectively the stochastic approximation space, is enriched. Concerning the stochastic collocation finite element method (SC-FEM), *a priori* error estimates are given in [7, 20] but, to our knowledge, no *a posteriori* error estimate for the whole solution in suitable norms has been derived yet. Recently, *a posteriori* error estimates for a specific quantity of interest have been developed. Goal-oriented error estimates can be found in [33, 35, 92] for the SG method and in [4] for the SC method.

We can distinguish two parts in this thesis. In the first part, which encompasses Chapters 1, 2, 3 and 4, we consider PDEs with small uncertainties affecting the coefficients, the forcing term, the physical domain, the boundary conditions or combinations of them. The assumption of small uncertainties naturally leads to the choice of perturbation techniques for the approximation of the stochastic space. Indeed, if the level of uncertainty is small, then only few terms in the power series expansion of the solution with respect to a parameter ε characterizing the amount of randomness of the problem will be needed to obtain an accurate approximation. With this technique, we are reduced to solve only deterministic problems whose solutions can be computed approximately with for instance the finite element method. The main goal of this thesis is then to derive error estimates that control the two sources of error: the stochastic error due to the truncation in the expansion of the solution and the spatial error coming from the finite element approximation of the continuous deterministic problems.

To have a general idea of the methodology, let us consider an abstract problem of the form: find $u(\cdot, \mathbf{Y}(\omega)) \in V$ such that almost surely

$$\mathcal{A}(u, v; \mathbf{Y}(\omega)) = F(v; \mathbf{Y}(\omega)) \quad \forall v \in V$$

where \mathbf{Y} is a random vector used to characterize the randomness in the input data, whose variability is controlled by a (small) parameter ε . Here, V is a given Hilbert space, \mathcal{A} is a bilinear form on $V \times V$ and F is a linear functional on V , the latter two being parametrized by the random vector \mathbf{Y} . The solution u of this problem also depends on \mathbf{Y} and, adopting a perturbation approach, it is then expanded as

$$u(\mathbf{x}, \mathbf{Y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{Y}(\omega)) + \varepsilon^2 u_2(\mathbf{x}, \mathbf{Y}(\omega)) + \dots$$

Considering a finite element space $V_h \subset V$, the first term in the expansion is approximated by $u_{0,h} \in V_h$, the solution of

$$\mathcal{A}(u_{0,h}, v_h; \mathbf{y}_0) = F(v_h; \mathbf{y}_0) \quad \forall v_h \in V_h$$

with $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}]$. Defining the residual for $u_{0,h}$ by

$$R(v; \mathbf{Y}(\omega)) := F(v; \mathbf{Y}(\omega)) - \mathcal{A}(u_{0,h}, v; \mathbf{Y}(\omega)),$$

the first step in the residual-based error estimation, that separates the two sources of error, is

Introduction

then

$$\mathcal{A}(u - u_{0,h}, v; \mathbf{Y}(\omega)) = F(v; \mathbf{Y}(\omega)) - \mathcal{A}(u_{0,h}, v; \mathbf{Y}(\omega)) = \text{I} + \text{II}$$

with

$$\text{I} := F(v; \mathbf{y}_0) - \mathcal{A}(u_{0,h}, v; \mathbf{y}_0) = R(v; \mathbf{y}_0)$$

$$\text{II} := F(v; \mathbf{Y}(\omega)) - F(v; \mathbf{y}_0) - \mathcal{A}(u_{0,h}, v; \mathbf{Y}(\omega)) + \mathcal{A}(u_{0,h}, v; \mathbf{y}_0) = R(v; \mathbf{Y}(\omega)) - R(v; \mathbf{y}_0).$$

The two terms can then be bounded separately. The first term I is nothing else than the residual for $u_{0,h}$ that can be bounded using a standard procedure as described by Verfürth in [118]. It yields an *a posteriori* error estimator that is localized on each element of the spatial mesh which can be used for mesh refinement. The second term is the one controlling the randomness. In this work, we will apply this methodology to a wide range of problems, as detailed in the thesis outline given below.

A different perspective is considered in the second part of this thesis, constituted of Chapter 5. Dropping the assumption of small uncertainty, and thus making perturbation techniques unsuitable, we use the stochastic collocation method to solve the random PDE. For the abstract problem considered above, this method, combined with the finite element method for the physical space discretization, consists in solving

$$\mathcal{A}(u_h(\cdot, \mathbf{y}_k), v_h; \mathbf{y}_k) = F(v_h; \mathbf{y}_k) \quad \forall v_h \in V_h$$

for a given set of collocation points \mathbf{y}_k , $k = 1, \dots, N_c$, in the stochastic space and building a global polynomial approximation

$$u_{h,N_c}(\mathbf{x}, \mathbf{Y}(\omega)) = \sum_{k=1}^{N_c} u_h(\mathbf{x}, \mathbf{y}_k) L_k(\mathbf{Y}(\omega))$$

for suitable multivariate polynomials L_k . The goal is then to estimate the error due to this method when combined with the finite element method for the spatial discretization. We propose a residual-based *a posteriori* error estimate for an elliptic diffusion problem. It consists of two terms controlling each source of error, the SC and the FE error. The stochastic estimator is then used to drive an adaptive sparse grid algorithm.

The precise outline of this thesis is as follows.

Thesis outline

We start in **Chapter 1** with an in-depth analysis of a second order elliptic differential equation with random diffusion coefficient. We present the methodology we are using, namely a perturbation technique for the stochastic space approximation and the finite element method for the physical space discretization. We provide then *a priori* and *a posteriori* error analysis in various norms and for several approximations. Extension to some class of nonlinear problems

and a comparison in terms of computational costs with the stochastic collocation method are also provided. Many numerical experiments are presented to illustrate the theoretical findings.

The results are then extended in **Chapter 2** where other sources of uncertainty are considered. More precisely, we consider first the case of random Neumann boundary conditions and then the combination of two uncertain inputs, the diffusion coefficient and the forcing term, described by two independent sets of random variables.

In **Chapter 3**, we consider nonlinear partial differential equations defined in random domains. Using the so-called *domain mapping method*, we use a random mapping to transform these equations into PDEs on a fixed reference domain with random coefficients. We start with the analysis of the one-dimensional steady-state viscous Burgers' equation in random intervals and consider then the more involved steady-state incompressible Navier-Stokes equations in random domains. We show the well-posedness of these problems, under suitable conditions on the mapping and the input data, and perform *a posteriori* error estimation for the finite element approximation of the first term in the expansion.

A time dependent parabolic problem is analysed in **Chapter 4**, considering the heat equation with random Robin boundary conditions. For the stochastic space, physical space and time discretizations, we use a perturbation technique, the finite element method and the (implicit) backward Euler scheme, respectively. We give an *a posteriori* error estimate for the first order approximation, which is here constituted of three parts controlling each source of error.

We conclude this thesis with an adaptive sparse grid algorithm for the stochastic collocation finite element method in **Chapter 5**. Considering again the diffusion model problem with random diffusion coefficient, that is assumed to depend affinely on a finite number of random variables, we derive an *a posteriori* error estimate for the total error that provides a guaranteed upper bound for the error. We propose then an algorithm that adaptively construct the multi-index set underlying the sparse grid and give numerical results to illustrate its performances.

Note: all the one-dimensional numerical experiments have been carried out using MATLAB Released R2012a, while the 2D numerical results have been obtained using either FreeFem++ 3.21 [78] or MATLAB.

1 Elliptic model problems with random diffusion coefficient

This chapter is mainly based on the paper [74] with respect to which we have done minor changes in the notation, essentially the distinction between a random vector $\mathbf{Y}: \Omega \rightarrow \Gamma \subset \mathbb{R}^L$ and a realization $\mathbf{y} \in \Gamma$. Moreover, we have added the following complements. First, a general statement of the model problem under consideration in Section 1.1. Additional numerical results are provided in Section 1.7. In particular, we present adaptive algorithms with non-uniform refinement which balances the two sources of error, namely the physical space discretization and the uncertainty. We give in Appendix some details about the derivation of the various deterministic problems for the first three terms in the expansion of the random solution, and state a precise link between each component of such terms and the derivatives of u with respect to the stochastic space variable. Finally, a detailed proof of the upper and lower bounds of a certain error estimator and estimates of the interpolation constant closes this chapter.

Introduction

In this chapter, we are focusing on PDEs with small uncertainties (for instance the linear model problem $-\operatorname{div}(a\nabla u) = f$ with $a = a_0 + \varepsilon(a_1 Y_1 + \dots + a_L Y_L)$ where ε is small and Y_1, \dots, Y_L are random variables). Following a different path than Monte-Carlo type, stochastic Galerkin or stochastic collocation methods, we adopt a perturbation approach, see e.g. [37, 82], which is appropriate for problems with small variability. We thus expand the stochastic solution u as

$$u(\mathbf{x}, \omega) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \omega) + \mathcal{O}(\varepsilon^2) \quad (1.1)$$

where ε is a parameter controlling the magnitude of uncertainty in the input which is assumed to be small. Uncoupled problems can be derived to find the deterministic part u_0 and the stochastic one u_1 (and higher order terms), the error analysis being performed in various norms. The main goal is then to derive *a posteriori* estimates for the error between the exact (random) solution u and certain approximations to be defined. For instance, if we write $u_{0,h}$ for the FE approximation of u_0 , then we will show that the error $u - u_{0,h}$ splits into two parts.

Chapter 1. Elliptic model problems with random diffusion coefficient

More precisely, we will derive an *a posteriori* error estimator η composed of two deterministic computable quantities η_1 and η_2 such that the following upper bound for the error holds

$$\|u - u_{0,h}\| \leq C\eta, \quad \eta = (\eta_1^2 + \eta_2^2)^{\frac{1}{2}},$$

with the norm $\|\cdot\|$ to be defined and where C is a constant depending only on the domain D , the mesh and a (deterministic) ellipticity constant. Therefore, by solving only one deterministic problem we can obtain an upper bound of the error due to space discretization (η_1) and the error due to uncertainty (η_2). This estimator can then be used to determine a mesh size yielding comparable accuracy in h and ε . The same kind of results can be obtained for $\|u - (u_{0,h} + \varepsilon u_{1,h})\|$, yielding a better accuracy in ε , and then for higher order terms.

We mention that the *a posteriori* error estimator that we obtain for $u - u_{0,h}$ for the elliptic model problem (1.2) has similarities with the one derived in [26], although the context of this paper is quite different from the one considered here. In [26] the authors derive an adaptive finite element method (AFEM) for elliptic PDEs with discontinuous coefficients. The proposed algorithm takes into account the error due to FE approximation but also the effect of replacing the discontinuous input data by some piecewise polynomial approximation, which plays the same role as a_0 in our setting. More precisely, before applying a standard AFEM to the problem, the mesh is first refined so that the discontinuous input are approximated by piecewise polynomials with a prescribed accuracy. The specific form of the uncertain input we consider here, see (1.12), allows us to increase the accuracy in ε by adding terms in the expansion (1.1) of u .

This chapter is organized as follows. The model problem, a second-order elliptic diffusion problem with homogeneous Dirichlet boundary conditions and random diffusion coefficient, is stated in Section 1.1. The diffusion coefficient is assumed, among others, to be expanded as a finite sum which depends on independent random variables with zero mean and unit variance. The methodology we are using to approximate the solution is given in Section 1.2. Error analysis in the H_0^1 and L^2 norms in the physical space, as well as goal-oriented error estimation, is performed in Section 1.3 where the exact (random) solution u is approximated by the (deterministic) FE approximation of u_0 . In Section 1.4, we consider the error between u and the FE approximation of $u_0 + \varepsilon u_1$, before giving a generalization for an approximation of arbitrary order in ε . The theory is then extended to nonlinear problems in Section 1.5. In Section 1.6, a comparison of the computational costs for the stochastic collocation method and the one presented here is performed. Section 1.7 is devoted to numerical examples used to illustrate and validate the theoretical results. Finally, a few complements are given in Appendix.

1.1 Problem statement

We start with a general and detailed description of the problem under consideration in this chapter, namely an elliptic diffusion PDE with random diffusion coefficient. In this description, we will make some distinctions in notation that will no longer be used in the next sections for ease of presentation.

General problem statement

Let D be a bounded polyhedral domain in \mathbb{R}^d , $d = 1, 2, 3$, and (Ω, \mathcal{F}, P) a complete probability space, where Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ -algebra of events and $P : \mathcal{F} \rightarrow [0, 1]$ is a probability measure. For any $p \in [1, \infty)$, let $L_p^p(\Omega)$ be the space of real-valued random variables Y on (Ω, \mathcal{F}, P) that are p -integrable with respect to P , i.e. such that $\int_\Omega |Y(\omega)|^p dP(\omega) < \infty$. Moreover, if $Y \in L_1^1(\Omega)$, we denote its expected value (or mean) by $\mathbb{E}[Y] = \int_\Omega Y(\omega) dP(\omega)$. The following problem is considered.

Find $u : D \times \Omega \rightarrow \mathbb{R}$ such that P -almost everywhere in Ω (in other words almost surely in Ω):

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad (1.2)$$

where a is a random field on (Ω, \mathcal{F}, P) over $L^\infty(D)$. For simplicity, the right-hand side f is assumed to be deterministic, $f \in L^2(D)$, but the case of stochastic forcing term could be considered as well adding no real difficulty, see Chapter 2. Note that the divergence and gradient operators apply only on \mathbf{x} , the physical space variable. Let $H_0^1(D)$ be endowed with the following norm

$$\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)} = \left(\int_D |\nabla v|^2 \right)^{\frac{1}{2}}.$$

The problem (1.2) can be written in weak form as:

find $u \in L_P^2(\Omega) \otimes H_0^1(D)$ such that

$$\mathbb{E} \left[\int_D a \nabla u \cdot \nabla v d\mathbf{x} \right] = \mathbb{E} \left[\int_D f v d\mathbf{x} \right] \quad \forall v \in L_P^2(\Omega) \otimes H_0^1(D). \quad (1.3)$$

Since the tensor product space $L_P^2(\Omega) \otimes H_0^1(D)$ is isomorphic (see for instance [10]) to the Bochner space

$$L_P^2(\Omega; H_0^1(D)) := \left\{ v : \Omega \rightarrow H_0^1(D) \mid v \text{ is strongly measurable and } \|v\|_{L_P^2(\Omega; H_0^1(D))} < \infty \right\} \quad (1.4)$$

where

$$\|v\|_{L_P^2(\Omega; H_0^1(D))}^2 := \int_\Omega \|\nabla v(\cdot, \omega)\|_{L^2(D)}^2 dP(\omega) = \mathbb{E} \left[\|\nabla v\|_{L^2(D)}^2 \right],$$

we can see the weak solution u of problem (1.2) as a function $u : \Omega \rightarrow H_0^1(D)$. The correspond-

Chapter 1. Elliptic model problems with random diffusion coefficient

ing pointwise weak formulation, equivalent to (1.3), is then given by:

find $u(\cdot, \omega) \in H_0^1(D)$ such that

$$\int_D a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in H_0^1(D), P\text{-a.e. in } \Omega. \quad (1.5)$$

If the diffusion coefficient a is (uniformly) bounded from below and from above, namely

$$\exists 0 < a_{\min} \leq a_{\max} < \infty : \quad P(\omega \in \Omega : a_{\min} \leq a(\mathbf{x}, \omega) \leq a_{\max} \quad \forall \mathbf{x} \in \bar{D}) = 1, \quad (1.6)$$

then we can show, by a straightforward application of Lax-Milgram's Lemma, that problem (1.5) is well-posed. More precisely, there exists a unique solution $u \in L_P^2(\Omega; H_0^1(D))$ which satisfies the *a priori* estimate

$$\|u\|_{L_P^2(\Omega; H_0^1(D))} \leq \frac{C_P}{a_{\min}} \|f\|_{L^2(D)}$$

with $C_P = C_P(D)$ the Poincaré constant.

Remark 1.1.1. *With the above assumptions, the solution belongs to $L_P^k(\Omega; H_0^1(D))$ for any $k \in [1, \infty]$. This is also true in the more general case $f \in L_P^{kp}(\Omega; H_0^1(D))$ and $a(x, \omega) \geq a_{\min}(\omega) > 0$ a.e. in D and a.s. in Ω with $\frac{1}{a_{\min}} \in L_P^{kq}(\Omega)$, where $\frac{1}{p} + \frac{1}{q} = 1$ (see [7]).*

We further assume that the random coefficient is well approximated by the finite expansion

$$a(\mathbf{x}, \omega) \approx a_L(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L a_j(\mathbf{x}) Y_j(\omega) \quad \text{with} \quad a_0(\mathbf{x}) = \mathbb{E}[a(\mathbf{x}, \cdot)], \quad (1.7)$$

where $\{Y_j\}_{j=1}^L$ are independent random variables with zero mean and unit variance.

Remark 1.1.2. *The characterization (1.7) of the random input can be achieved using for instance a truncated Karhunen-Loève type expansion (see [87, 88]) if the mean and the two-point correlation (or equivalently the covariance) of a is known. In this case, the functions a_j , $j = 1, \dots, L$, in (1.7) write $a_j(\mathbf{x}) = \sqrt{\lambda_j} \varphi_j(\mathbf{x})$ with $\{\lambda_j, \varphi_j\}$ the eigenpairs of the (compact and self-adjoint) integral operator associated with the covariance kernel $V : D \times D \rightarrow \mathbb{R}$ given by*

$$V(\mathbf{x}, \mathbf{x}') := \frac{1}{\varepsilon^2} \mathbb{E}[(a(\mathbf{x}, \omega) - a_0(\mathbf{x}))(a(\mathbf{x}', \omega) - a_0(\mathbf{x}'))].$$

Notice that, in general, the family of random variables appearing in the KL expansion of an arbitrary random field a are only uncorrelated (see [111]), but not necessarily independent.

The problem (1.2) is then approximated by:

find $u_L : D \times \Omega \rightarrow \mathbb{R}$ such that P -a.e. in Ω the following equation holds

$$\begin{cases} -\operatorname{div}(a_L(\mathbf{x}, \omega) \nabla u_L(\mathbf{x}, \omega)) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u_L(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad (1.8)$$

which admits a unique weak solution $u_L \in L_P^2(\Omega; H_0^1(D))$ under the assumption

$$\exists 0 < a_{L,\min} \leq a_{L,\max} < \infty : P(\omega \in \Omega : a_{L,\min} \leq a_L(\mathbf{x}, \omega) \leq a_{L,\max}, \forall \mathbf{x} \in \tilde{D}) = 1.$$

The stochasticity of the problem (1.8) for u_L can therefore be parametrized by the random vector $\mathbf{Y} = (Y_1, \dots, Y_L)$. Indeed, with the definition of a_L given in (1.7) we have $a_L(\mathbf{x}, \omega) = \tilde{a}_L(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega))$ and thus $u_L(\mathbf{x}, \omega) = \tilde{u}_L(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega))$ thanks to the Doob-Dynkin Lemma (see [6, p.6] for instance). We can therefore derive a parametric *deterministic* weak formulation of (1.8). Let $\Gamma = \Gamma_1 \times \Gamma_2 \times \dots \times \Gamma_L$ where Γ_j denotes the bounded image in \mathbb{R} of the random variable Y_j , i.e. $\Gamma_j := Y_j(\Omega)$, for $j = 1, \dots, L$. Moreover, let $\rho_j : \Gamma_j \rightarrow \mathbb{R}^+$ denote the probability density function of Y_j , $j = 1, \dots, L$. Thanks to the independence of the random variables, the joint density function $\rho : \Gamma \rightarrow \mathbb{R}^+$ of the random vector \mathbf{Y} factorizes as $\rho(\mathbf{y}) = \prod_{j=1}^L \rho_j(y_j)$ for all $\mathbf{y} = (y_1, \dots, y_L) \in \Gamma$. We can thus replace the probability space (Ω, \mathcal{F}, P) by its image $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $B(\Gamma)$ denotes the Borel σ -algebra defined on Γ and $\rho(\mathbf{y})d\mathbf{y}$ the probability measure of \mathbf{Y} . For any measurable function $\tilde{g}_L : \Gamma \rightarrow \mathbb{R}$ defined on $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, the expectation of the random variable $g_L = \tilde{g}_L \circ \mathbf{Y} : \Omega \rightarrow \mathbb{R}$ is then given by

$$\mathbb{E}[g_L] = \int_{\Omega} g_L(\omega) dP(\omega) = \int_{\Omega} \tilde{g}_L(\mathbf{Y}(\omega)) dP(\omega) = \int_{\Gamma} \tilde{g}_L(\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y}.$$

Remark 1.1.3. *The error analysis for $u - u_0$ with u_0 the first term in the expansion, see (1.1), is exactly the same as the one performed in Section 1.3 if the random variables are assumed uncorrelated instead of independent, i.e. such that $\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i] \mathbb{E}[Y_j]$ for any $i, j = 1, \dots, L$ with $i \neq j$. For the higher order approximations, however, few changes have to be made to the analysis given in Section 1.4. Moreover, the definitions given above are not restricted to continuous random variables but also hold for discrete random variables. In such a case, we consider a generalized probability density function defined via Dirac delta functions. For instance, the density function of a random variable Y_j taking value ± 1 with probability $\frac{1}{2}$ would be*

$$\rho_j(y_j) = \frac{1}{2}(\delta(y_j + 1) + \delta(y_j - 1)).$$

Such random variables will be considered in the numerical results of Section 1.7.

The (parametric, pointwise) weak formulation of problem (1.8) reads:

find $\tilde{u}_L : \Gamma \rightarrow H_0^1(D)$ such that

$$\int_D \tilde{a}_L(\mathbf{x}, \mathbf{y}) \nabla \tilde{u}_L(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma, \quad (1.9)$$

Chapter 1. Elliptic model problems with random diffusion coefficient

where $\tilde{a}_L(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L a_j(\mathbf{x}) y_j$. Thanks again to Lax-Milgram's lemma, we know that there exists a unique solution $\tilde{u}_L \in L^2_\rho(\Gamma; H^1_0(D))$ of problem (1.9) which satisfies

$$\|\tilde{u}_L\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq \frac{C_P}{a_{\min}} \|f\|_{L^2(D)},$$

where similarly to (1.4) we define

$$L^2_\rho(\Gamma; H^1_0(D)) := \left\{ v : \Gamma \rightarrow H^1_0(D) \mid v \text{ is strongly measurable and } \|v\|_{L^2_\rho(\Gamma; H^1_0(D))} < \infty \right\} \quad (1.10)$$

with

$$\|v\|_{L^2_\rho(\Gamma; H^1_0(D))}^2 := \int_\Gamma \|\nabla v(\cdot, \mathbf{y})\|_{L^2(D)}^2 \rho(\mathbf{y}) d\mathbf{y}.$$

Notice that the weak solution u_L of problem (1.8) and the solution \tilde{u}_L of problem (1.9) are related by

$$u_L(x, \omega) = \tilde{u}_L(x, Y_1(\omega), \dots, Y_L(\omega)) \quad \text{a.s. in } \Omega$$

and we have

$$\|u_L\|_{L^2_\rho(\Omega; H^1_0(D))} = \|\tilde{u}_L\|_{L^2_\rho(\Gamma; H^1_0(D))}.$$

For the sake of presentation \tilde{a}_L and \tilde{u}_L will be denoted again a_L and u_L , respectively, i.e. we write $a_L(\mathbf{x}, \omega) = a_L(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega))$ and $u_L(\mathbf{x}, \omega) = u_L(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega))$, when no ambiguity arises. Moreover, the goal here is not to analyse the error committed when replacing a by a_L , i.e. when the random input is approximated via L random variables. Therefore, we assume from now on that $a = a_L$, i.e. $u = u_L$. We mention that a complete analysis of the (*strong, weak*) error $u - u_L$ can be found in [44].

Specific problem statement

We give now a short statement of the problem that will be analysed in the subsequent sections, indicating only the necessary assumptions and using the shorthand notation described above. We consider the following problem.

Find $u : D \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω :

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D, \end{cases} \quad (1.11)$$

where $f \in L^2(D)$ is deterministic and a is a random field on (Ω, \mathcal{F}, P) over $L^\infty(D)$ which satisfies the following assumptions (see [6, 7, 10] for instance) that ensure, among others, the well-posedness of the problem:

(A1) coercivity and continuity: a is bounded and uniformly coercive, i.e. there exist two real

constants $0 < a_{\min} \leq a_{\max} < \infty$ such that

$$P(\omega \in \Omega : a_{\min} \leq a(\mathbf{x}, \omega) \leq a_{\max}, \forall \mathbf{x} \in \overline{D}) = 1.$$

(A2) finite dimensional noise: a is parametrized by L mutually independent random variables $a(\mathbf{x}, \omega) = a(\mathbf{x}, Y_1(\omega), Y_2(\omega), \dots, Y_L(\omega))$. More precisely, we assume that a can be expanded as

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L a_j(\mathbf{x}) Y_j(\omega), \quad (1.12)$$

where the $\{Y_j\}_{j=1}^L$ are independent random variables with zero mean and unit variance, $a_j \in W^{1,\infty}(D)$ for $j = 0, \dots, L$ and $\varepsilon \in [0, \varepsilon_{\max}]$ with ε_{\max} the maximum value such that property (A1) is satisfied. The functions a_j , $j = 0, 1, \dots, L$, and the random variables Y_j , $j = 1, \dots, L$, are assumed to be independent of ε .

Notice that assuming $a_j \in L^\infty(D)$ for $j = 0, 1, \dots, L$ is enough to ensure the well-posedness of the problem. We impose here more regularity in order to avoid difficulties that are beyond the scope of this work. We refer to [23] for a derivation of *a posteriori* error estimation in the case of discontinuous coefficients. Moreover, as a consequence of assumption (A1), the random variables Y_j , $j = 1, \dots, L$, have to be bounded almost surely. In particular, they have finite moment of any order. Finally, from assumption (A2) it follows that the mean and variance of a are given by $\mathbb{E}[a](\mathbf{x}) = a_0(\mathbf{x})$ and $\text{Var}[a](\mathbf{x}) = \varepsilon^2 \sum_{j=1}^L a_j^2(\mathbf{x})$, respectively. Therefore, for fixed functions a_j , we can modify the variance of a by changing the value of ε . From assumption (A2), the solution u is a function of the random variables Y_j , i.e. $u(\mathbf{x}, \omega) = u(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega))$. Replacing (Ω, \mathcal{F}, P) by $(\Gamma, B(\Gamma), \rho(\mathbf{y}) d\mathbf{y})$, the stochastic elliptic boundary value problem (1.11) can equivalently be written in the following *deterministic* parametric form:

find $u : D \times \Gamma \rightarrow \mathbb{R}$ such that ρ -a.e. in Γ we have

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x}, \mathbf{y}) &= 0 & \mathbf{x} \in \partial D. \end{cases} \quad (1.13)$$

The (parametric, pointwise) weak form of problem (1.13) then reads:

find $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

$$\mathcal{A}(u(\cdot, \mathbf{y}), v; \mathbf{y}) = \mathcal{F}(v) \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (1.14)$$

where

$$\mathcal{A}(u(\cdot, \mathbf{y}), v; \mathbf{y}) = \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}, \quad (1.15)$$

$$\mathcal{F}(v) = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}. \quad (1.16)$$

Again, thanks to Lax-Milgram's lemma the coercivity and continuity assumptions on a ensure the well-posedness of problem (1.14), namely there exists a unique solution $u \in L^2_\rho(\Gamma; H^1_0(D))$. Indeed, since a is bounded from below and above almost surely, the bilinear form \mathcal{A} is continuous and coercive with constant of continuity and coercivity given respectively by a_{max} and a_{min} . Furthermore, the linear (deterministic) functional \mathcal{F} is continuous, with constant of continuity equal to $C_P \|f\|_{L^2(D)}$, where C_P denotes the constant in the Poincaré inequality. Therefore, the solution u of problem (1.14) satisfies

$$\|\nabla u(\cdot, \mathbf{y})\|_{L^2(D)} \leq \frac{C_P}{a_{min}} \|f\|_{L^2(D)} \quad \rho\text{-a.e. in } \Gamma. \quad (1.17)$$

Notice that the weak solution of problem (1.11) is then given by $u(\cdot, \mathbf{Y}(\omega))$ with u the parametric solution of problem (1.14) and it satisfies

$$\|\nabla u(\cdot, \mathbf{Y}(\omega))\|_{L^2(D)} \leq \frac{C_P}{a_{min}} \|f\|_{L^2(D)} \quad \text{a.s. in } \Omega. \quad (1.18)$$

Moreover, it has been proved (see for instance [7]) that solution $u = u(\mathbf{x}, \mathbf{y})$ of (1.14) is analytic with respect to each variable y_j , $j = 1, \dots, L$.

For ease of presentation, the dependence of the random variables Y_j with respect to $\omega \in \Omega$ will not necessarily be indicated in the subsequent analysis.

1.2 Methodology

In this section, we present the method we use to approximate the random (weak) solution u of problem (1.11). We use first a perturbation technique for the stochastic space approximation, yielding a collection of deterministic problems. The physical space approximation of each problem is then performed using the finite element method. More precisely, we assume from now on that ε in (1.12) is small enough that (A2) holds and expand the solution $u = u(\mathbf{x}, \mathbf{Y}(\omega))$ with respect to ε up to a certain order $N \in \mathbb{N}$

$$u(\mathbf{x}, \mathbf{Y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{Y}(\omega)) + \dots + \varepsilon^N u_N(\mathbf{x}, \mathbf{Y}(\omega)) + \mathcal{O}(\varepsilon^{N+1}). \quad (1.19)$$

Inserting the latter expansion into (1.11) with a defined in (1.12) and keeping the $\mathcal{O}(1)$ term with respect to ε yields the problem:

find $u_0 : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x}) \nabla u_0(\mathbf{x})) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u_0(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D. \end{cases} \quad (1.20)$$

Then, writing $u_1(\mathbf{x}, \mathbf{Y}(\omega)) = \sum_{j=1}^L U_j(\mathbf{x}) Y_j(\omega)$ and keeping the $\mathcal{O}(\varepsilon)$ terms in (1.11) yields the L problems:

find $U_j : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_j(\mathbf{x})\nabla u_0(\mathbf{x}) + a_0(\mathbf{x})\nabla U_j(\mathbf{x})) &= 0 & \mathbf{x} \in D \\ U_j(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad j = 1, \dots, L, \quad (1.21)$$

in which the solution u_0 of problem (1.20) is needed. Notice that for $j = 1, \dots, L$, the function U_j is related to $\frac{\partial u(\mathbf{x}, \mathbf{y}_0)}{\partial y_j}$ with $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$. Similarly, we can use the solutions U_j , $j = 1, \dots, L$, of problem (1.21) to compute the deterministic part of the next term in the expansion (1.19), which in turn is related to the second derivatives $\frac{\partial^2 u(\mathbf{x}, \mathbf{y}_0)}{\partial y_k \partial y_j}$, $j, k = 1, \dots, L$. Indeed, if we write $u_2(\mathbf{x}, \mathbf{Y}(\omega)) = \sum_{j,k=1}^L U_{jk}(\mathbf{x}) Y_j(\omega) Y_k(\omega)$, keeping the $\mathcal{O}(\varepsilon^2)$ terms in (1.11), we get the L^2 problems:

find $U_{jk} : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_j(\mathbf{x})\nabla U_k(\mathbf{x}) + a_0(\mathbf{x})\nabla U_{jk}(\mathbf{x})) &= 0 & \mathbf{x} \in D \\ U_{jk}(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad j, k = 1, \dots, L. \quad (1.22)$$

More details about the derivation of problems (1.20), (1.21) and (1.22) are given in Appendix 1.A.

Remark 1.2.1. *We will prove in the sections 1.3, 1.4.1 and 1.4.2 that*

$$u - u_0 = \mathcal{O}(\varepsilon), \quad u - (u_0 + \varepsilon u_1) = \mathcal{O}(\varepsilon^2) \quad \text{and} \quad u - (u_0 + \varepsilon u_1 + \varepsilon^2 u_2) = \mathcal{O}(\varepsilon^3).$$

The solution to the deterministic problems (1.20), (1.21) and (1.22) can be approximated using for instance the finite element method. For any $h > 0$, let \mathcal{T}_h be a family of partitions of D into d -simplices (intervals, triangles, tetrahedra) K of diameter $h_K \leq h$. Unless otherwise stated, we will always consider shape regular (see [49]) meshes of D , i.e. decompositions such that there exists a constant $c > 0$ satisfying

$$\frac{h_K}{\rho_K} \leq c \quad \forall K \in \mathcal{T}_h, \forall h > 0 \quad (1.23)$$

where $\rho_K = \sup\{\operatorname{diam}(B) : B \text{ is a ball contained in } K\}$. The condition (1.23) is equivalent to a *minimal angle* condition, namely that there exists a constant α_0 such that $\alpha_K \geq \alpha_0 > 0$ for all $K \in \mathcal{T}_h$ with α_K the smallest angle of K . Let $V_h \subset H_0^1(D)$ be the space of continuous, piecewise linear finite element functions associated to \mathcal{T}_h that vanish on ∂D , that is

$$V_h := \{v_h \in C^0(\bar{D}) : v_h|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h\} \cap H_0^1(D),$$

where \mathbb{P}_1 is the set of polynomials of degree less than or equal to 1.

In the derivation of *a priori* and *a posteriori* error estimates, we will need an interpolation operator which maps $H_0^1(D)$ to V_h , along with interpolation error bounds. We distinguish the

cases $d = 1$ and $d = 2, 3$. For the one-dimensional case, any function of $H_0^1(D)$ is continuous thanks to the Sobolev embedding theorem. Therefore, the Lagrange interpolant operator $r_h : C^0(\bar{D}) \rightarrow V_h$, which requires point evaluations, is well-defined and satisfies the following error bounds: there exists a constant $C > 0$ such that $\forall h > 0$, $\forall K \in \mathcal{T}_h$ and all $v \in H_0^1(D)$ we have

$$\|v - r_h v\|_{L^2(K)} \leq Ch_K \|v'\|_{L^2(K)} \quad (1.24)$$

and for all $v \in H^2(D)$

$$\|v - r_h v\|_{L^2(K)} + h_K \|v' - (r_h v)'\|_{L^2(K)} \leq h_K^2 \|v''\|_{L^2(K)}.$$

For the case $d = 2, 3$, the functions of $H^2(D)$ are continuous and we have the following error bound (see [31, 49] for instance) based on the Bramble-Hilbert lemma: there exists a constant $C > 0$ such that $\forall h > 0$, $\forall K \in \mathcal{T}_h$ and all $v \in H^2(K)$ we have

$$\|v - r_h v\|_{L^2(K)} + h_K \|\nabla(v - r_h v)\|_{L^2(K)} \leq Ch_K^2 |v|_{H^2(K)}. \quad (1.25)$$

In general however, such regularity might not be reached by the solution of problem (1.14), since we are seeking a solution in $H_0^1(D)$ in the physical space. In that case, we will use the Clément interpolant [50] operator $\mathcal{J}_h : H^1(D) \rightarrow V_h$ which satisfies the following interpolation results: there exists a constant $C > 0$ such that $\forall h > 0$, $\forall K, e \in \mathcal{T}_h$ and all $v \in H^1(D)$ we have

$$\|v - \mathcal{J}_h v\|_{L^2(K)} \leq Ch_K |v|_{H^1(N(K))}, \quad (1.26)$$

$$\|\nabla(v - \mathcal{J}_h v)\|_{L^2(K)} \leq C |v|_{H^1(N(K))} \quad (1.27)$$

and

$$\|v - \mathcal{J}_h v\|_{L^2(e)} \leq Ch_e^{\frac{1}{2}} |v|_{H^1(N(K_e))}, \quad (1.28)$$

where, for an internal edge e , K_e is the union of the two elements touching e and $N(K)$ (respectively $N(K_e)$) denotes the patch of elements associated to K (respectively K_e). Notice that the constant C in (1.26), (1.27) and (1.28) depends on the constant in (1.23) characterizing the mesh aspect ratio.

We will now derive *a priori* and *a posteriori* error estimates in various norms, the error being the difference between the exact solution and a certain approximate solution to be defined. We first start by giving error estimates between the exact solution u and $u_{0,h}$, the FE approximation of u_0 . Our goal is to decompose the error into two parts, the error due to the finite element approximation (h) and the error due to the uncertainty (ε).

1.3 Error analysis for the first order approximation

We consider u the (weak) solution of (1.11) and u_0 that of (1.20), i.e. the case $N = 0$ in the expansion (1.19). The error due to the stochastic truncation is of order ε . Indeed, for any

$v \in H_0^1(D)$ and a.s. in Ω we have

$$\int_D a \nabla(u(\cdot, \mathbf{Y}(\omega)) - u_0) \cdot \nabla v = \int_D f v - \int_D a(\cdot, \mathbf{Y}(\omega)) \nabla u_0 \cdot \nabla v = -\varepsilon \sum_{j=1}^L Y_j(\omega) \int_D a_j \nabla u_0 \cdot \nabla v. \quad (1.29)$$

Using the FEM, the unknown solution u_0 of problem (1.20) is approximated by $u_{0,h}$, the solution of:

$$\text{find } u_{0,h} \in V_h \text{ such that } \int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D f v_h \quad \forall v_h \in V_h. \quad (1.30)$$

In what follows, we will derive *a priori* and *a posteriori* error estimates for $u - u_{0,h}$ in various norms. In particular, the *a posteriori* error estimators, which are computable quantities, yield useful information about the two sources of error by computing only one deterministic problem.

1.3.1 *A priori* error analysis

This section is devoted to *a priori* error estimation for the strong and *weak* errors, which gives information on the asymptotic behaviour of the error. In particular, we will show that the order of the error of the mean in ε is twice the order of the strong error, while the order of the error in h is the same for both. Sections 1.3.2, 1.3.2 and 1.3.2 are instead devoted to *a posteriori* error estimates in different norms.

Strong error estimate

Let us first give error estimates on the strong error, i.e. on the error between u and $u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm. Our goal is to prove that there exists a constant $C > 0$ independent of h and ε such that

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq C(h + \varepsilon).$$

Proposition 1.3.1. *Let u and u_0 be the (weak) solutions of problems (1.11) and (1.20), respectively, and let $u_{0,h}$ be the solution of problem (1.30). If $u_0 \in H^2(D)$, then we have the *a priori* error estimate*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \sqrt{2} \left[\frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|_{H^2(D)}^2 + L \frac{\varepsilon^2 C_P^2}{a_{0,min}^2 a_{min}^2} \|f\|_{L^2(D)}^2 \sum_{j=1}^L \|a_j^2\|_{L^\infty(D)} \right]^{\frac{1}{2}} \quad (1.31)$$

where $C > 0$ is the constant that appears in (1.25). Moreover, if we assume that for a fixed value $\alpha > \frac{1}{2}$, there exists a constant M_α such that for any L we have $\sum_{j=1}^L \|a_j^2\|_{L^\infty(D)} j^{2\alpha} \leq M_\alpha$, then we

also have

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \sqrt{2} \left[\frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|_{H^2(D)}^2 + M_\alpha \frac{\varepsilon^2 C_P^2}{a_{0,min}^2 a_{min}^2} \|f\|_{L^2(D)}^2 \sum_{j=1}^{\infty} j^{-2\alpha} \right]^{\frac{1}{2}}. \quad (1.32)$$

Remark 1.3.2. The *a priori* error estimate (1.31) blows up when L tends to infinity since the second part of the estimate depends linearly on L . If we add a constraint on the functions a_j , $j = 1, \dots, L$, for instance that a_j decays as $j^{-\beta}$ with $\beta > \alpha + \frac{1}{2}$, then (1.32) holds with M_α independent of L .

Proof. Using the fact that almost surely it holds

$$\int_D a_0 \nabla u_0 \cdot \nabla v = \int_D f v = \int_D a \nabla u \cdot \nabla v \quad \forall v \in H_0^1(D),$$

we have for any $v \in V$

$$\begin{aligned} \int_D a_0 \nabla(u - u_{0,h}) \cdot \nabla v &= \int_D a_0 \nabla(u - u_0) \cdot \nabla v + \int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v \\ &= - \int_D (a - a_0) \nabla u \cdot \nabla v + \int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v \\ &\leq \left[\left(\int_D \frac{(a_0 - a)^2}{a_0} |\nabla u|^2 \right)^{\frac{1}{2}} + \left(\int_D a_0 |\nabla(u_0 - u_{0,h})|^2 \right)^{\frac{1}{2}} \right] \cdot \left(\int_D a_0 |\nabla v|^2 \right)^{\frac{1}{2}}. \end{aligned} \quad (1.33)$$

Thanks to the inequality $(a + b)^2 \leq 2(a^2 + b^2)$, $v = u(\cdot, \mathbf{Y}(\omega)) - u_{0,h} \in V$ a.s. in Ω in the last inequality yields

$$\left(\int_D a_0 |\nabla(u - u_{0,h})|^2 \right)^{\frac{1}{2}} \leq \sqrt{2} \left[\frac{1}{a_{0,min}} \int_D (a - a_0)^2 |\nabla u|^2 + \int_D a_0 |\nabla(u_0 - u_{0,h})|^2 \right]^{\frac{1}{2}}. \quad (1.34)$$

The second term of the right-hand side of (1.34) can be bounded in a standard manner as follows. Using the Galerkin orthogonality property

$$\int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v_h = 0 \quad \forall v_h \in V_h,$$

we easily get

$$\int_D a_0 |\nabla(u_0 - u_{0,h})|^2 \leq a_{0,max} \|\nabla(u_0 - \mathcal{I}_h u_0)\|_{L^2(D)}^2.$$

Since $u_0 \in H^2(D)$ by assumption, thanks to the interpolation result (1.25) we get

$$\int_D a_0 |\nabla(u_0 - u_{0,h})|^2 \leq a_{0,max} C^2 h^2 |u_0|_{H^2(D)}^2. \quad (1.35)$$

1.3. Error analysis for the first order approximation

Therefore, using this last relation and the lower bound for a_0 in (1.34) yields a.s. in Ω

$$\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \leq 2 \left[\frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|_{H^2(D)}^2 + \frac{1}{a_{0,min}^2} \int_D (a - a_0)^2 |\nabla u|^2 \right].$$

Then, we take the expected value on both sides of the last inequality to get

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right] \leq 2 \left[\frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|_{H^2(D)}^2 + \frac{1}{a_{0,min}^2} \mathbb{E} \left[\int_D (a - a_0)^2 |\nabla u|^2 \right] \right]. \quad (1.36)$$

To complete the proof, we finally bound the expected value that appears on the right-hand side of (1.36). First, using the relation $(\sum_{j=1}^L x_j)^2 \leq L \sum_{j=1}^L x_j^2$, we easily get

$$\mathbb{E} \left[\int_D (a - a_0)^2 |\nabla u|^2 \right] \leq L \frac{\varepsilon^2 C_P^2}{a_{min}^2} \|f\|_{L^2(D)}^2 \sum_{j=1}^L \|a_j^2\|_{L^\infty(D)}$$

which proves (1.31). For (1.32), we use the additional assumption and the relation $\sum_i a_i b_i \leq (\sum_i a_i^2)^{\frac{1}{2}} (\sum_i b_i^2)^{\frac{1}{2}}$ to obtain

$$(a - a_0)^2 = \varepsilon^2 \left(\sum_{j=1}^L a_j j^\alpha j^{-\alpha} Y_j \right)^2 \leq \varepsilon^2 \left(\sum_{j=1}^L a_j^2 j^{2\alpha} \right) \left(\sum_{j=1}^L Y_j^2 j^{-2\alpha} \right) \leq M_\alpha \varepsilon^2 \sum_{j=1}^L Y_j^2 j^{-2\alpha}.$$

Therefore, thanks to (1.18) and the fact that $\mathbb{E}[Y_j^2] = 1$, we obtain

$$\mathbb{E} \left[\int_D (a - a_0)^2 |\nabla u|^2 \right] \leq M_\alpha \frac{\varepsilon^2 C_P^2}{a_{min}^2} \|f\|_{L^2(D)}^2 \sum_{j=1}^L j^{-2\alpha} \leq M_\alpha \frac{\varepsilon^2 C_P^2}{a_{min}^2} \|f\|_{L^2(D)}^2 \sum_{j=1}^\infty j^{-2\alpha}.$$

Since $\alpha > \frac{1}{2}$, the series $\sum_{j=1}^\infty j^{-2\alpha}$ converges which concludes the proof. \square

Mean of the error estimate

We are now interested in the error on the law of u . We restrict ourselves, in particular, to the $H_0^1(D)$ norm of the expected value of $u - u_{0,h}$. In this case, the statistical error is of order 2, to be compared to the order 1 of the strong error. Under the same regularity condition on u_0 , we can show the following *a priori* error estimate.

Proposition 1.3.3. *Let u and u_0 be the (weak) solutions of problems (1.11) and (1.20), respectively, and let $u_{0,h}$ be the solution of problem (1.30). If $u_0 \in H^2(D)$, then we have the *a priori* error estimate*

$$\|\mathbb{E}[u - u_{0,h}]\|_{H_0^1(D)} \leq \sqrt{\frac{a_{0,max}}{a_{0,min}}} C_1 h |u_0|_{H^2(D)} + \frac{\varepsilon^2 C_P}{a_{0,min}^3} \|f\|_{L^2(D)} \sum_{j=1}^L \|a_j\|_{L^\infty(D)}^2 + C_2 \varepsilon^3, \quad (1.37)$$

where $C_1 > 0$ is the constant in (1.25) and C_2 is a constant independent of u , h and ε . Therefore,

there exists a constant $\tilde{C} > 0$ independent of h and ε such that

$$\|\mathbb{E}[u - u_{0,h}]\|_{H_0^1(D)} \leq \tilde{C}(h + \varepsilon^2).$$

Proof. Let us define $u_1 = \sum_{j=1}^L U_j Y_j$, where U_j is the solution of problem (1.21) for $j = 1, \dots, L$. First, the expected value of the error $u(\cdot, \mathbf{Y}) - u_{0,h}$ naturally splits into two parts

$$\mathbb{E}[u - u_{0,h}] = \mathbb{E}[u - u_0] + (u_0 - u_{0,h})$$

and thus, thanks to the triangle inequality, we get

$$\|\mathbb{E}[u - u_{0,h}]\|_{H_0^1(D)} \leq \|\mathbb{E}[u - u_0]\|_{H_0^1(D)} + \|u_0 - u_{0,h}\|_{H_0^1(D)}.$$

From (1.35), we deduce a bound for the second term given by

$$\|u_0 - u_{0,h}\|_{H_0^1(D)} \leq \sqrt{\frac{a_{0,max}}{a_{0,min}}} C_1 h |u_0|_{H^2(D)},$$

where C_1 is the constant that appears in (1.25). Let us bound the term $\|\mathbb{E}[u - u_0]\|_{H_0^1(D)}$, which is due to the uncertainty in the diffusion coefficient. Proceeding as in (1.29) and using the fact that $\int_D (a_j \nabla u_0 + a_0 \nabla U_j) \cdot \nabla v = 0$ for all $v \in V$, the following equalities hold for any $v \in V$ and a.s. in Ω

$$\begin{aligned} \int_D a \nabla(u - u^1) \cdot \nabla v &= -\varepsilon \int_D a_0 \nabla u_1 \cdot \nabla v - \int_D (a - a_0) \nabla u^1 \cdot \nabla v \\ &= -\varepsilon \sum_{j=1}^L Y_j \int_D (a_0 \nabla U_j + a_j \nabla u_0) \cdot \nabla v - \varepsilon^2 \int_D \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_i \cdot \nabla v \\ &= -\varepsilon^2 \int_D \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_i \cdot \nabla v. \end{aligned} \tag{1.38}$$

Therefore, we have

$$\int_D a_0 \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v = - \int_D (a - a_0) \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v - \varepsilon^2 \sum_{i,j=1}^L Y_i Y_j \int_D a_i \nabla U_j \cdot \nabla v.$$

Since $\mathbb{E}[u_1] = 0$ and $\mathbb{E}[Y_i Y_j] = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta, taking the expected value on both sides of last equality yields

$$\int_D a_0 \nabla \mathbb{E}[u - u_0] \cdot \nabla v = \mathbb{E} \left[- \int_D (a - a_0) \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v \right] - \varepsilon^2 \sum_{j=1}^L \int_D a_j \nabla U_j \cdot \nabla v.$$

1.3. Error analysis for the first order approximation

Thanks to Jensen's inequality (see e.g. [89]), we obtain

$$\begin{aligned} \int_D a_0 \nabla \mathbb{E}[u - u_0] \cdot \nabla v &\leq \mathbb{E} \left[\|a - a_0\|_{L^\infty(D)} \|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)} \right] \|\nabla v\|_{L^2(D)} \\ &\quad + \varepsilon^2 \|\nabla v\|_{L^2(D)} \sum_{j=1}^L \|a_j\|_{L^\infty(D)} \|\nabla U_j\|_{L^2(D)}. \end{aligned}$$

If we take $v = \mathbb{E}[u - u_0]$ in the last inequality, we get

$$\|\mathbb{E}[u - u_0]\|_{H_0^1(D)} \leq \frac{1}{a_{0,min}} \left\{ \mathbb{E} \left[\|a - a_0\|_{L^\infty(D)} \|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)} \right] + \varepsilon^2 \sum_{j=1}^L \|a_j\|_{L^\infty(D)} \|\nabla U_j\|_{L^2(D)} \right\}. \quad (1.39)$$

We now give a bound on $\|\nabla U_j\|_{L^2(D)}$, $j = 1, \dots, L$. First, using standard techniques (Cauchy-Schwarz, Poincaré inequalities, lower bound for a_0), we get the following bound on the solution of problem (1.20)

$$\|\nabla u_0\|_{L^2(D)} \leq \frac{C_P}{a_{0,min}} \|f\|_{L^2(D)}.$$

Then, taking $v = U_j$ as test function in the weak formulation of problem (1.21) yields

$$a_{0,min} \|\nabla U_j\|_{L^2(D)}^2 \leq \int_D a_0 |\nabla U_j|^2 = - \int_D a_j \nabla u_0 \cdot \nabla U_j \leq \|a_j\|_{L^\infty(D)} \|\nabla u_0\|_{L^2(D)} \|\nabla U_j\|_{L^2(D)}$$

and thus

$$\|\nabla U_j\|_{L^2(D)} \leq \frac{C_P}{a_{0,min}^2} \|f\|_{L^2(D)} \|a_j\|_{L^\infty(D)}.$$

Inserting this result in (1.39), we get

$$\|\mathbb{E}[u - u_0]\|_{H_0^1(D)} \leq \frac{1}{a_{0,min}} \left\{ \mathbb{E} \left[\|a - a_0\|_{L^\infty(D)} \|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)} \right] + \frac{\varepsilon^2 C_P}{a_{0,min}^2} \|f\|_{L^2(D)} \sum_{j=1}^L \|a_j\|_{L^\infty(D)}^2 \right\}. \quad (1.40)$$

To conclude the proof, we show that the first term of the right-hand side of the last inequality is of higher order in ε , namely of order ε^3 . Indeed, we have

$$\|a - a_0\|_{L^\infty(D)} = \varepsilon \sum_{j=1}^L |Y_j| \|a_j\|_{L^\infty(D)} \leq c_1 \varepsilon$$

and, taking $v = u - (u_0 + \varepsilon u_1)$ in (1.38),

$$\|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)} \leq \frac{1}{a_{min}} \varepsilon^2 \sum_{i,j=1}^L |Y_i Y_j| \|a_i\|_{L^\infty(D)} \|\nabla U_j\|_{L^2(D)} \leq c_2 \varepsilon^2 \quad (1.41)$$

with c_1, c_2 two (deterministic) constants independent of u , h and ε . Therefore, we have

$$\mathbb{E} \left[\|a - a_0\|_{L^\infty(D)} \|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)} \right] \leq C_2 \varepsilon^3$$

with $C_2 = c_1 c_2$. □

Remark 1.3.4. A bound for $\|\mathbb{E}[u - u_0]\|_{H_0^1(D)}$ can also be obtained using Jensen's inequality, the fact that the term u_1 is mean-free and (1.41) as follows

$$\begin{aligned} \|\mathbb{E}[u - u_0]\|_{H_0^1(D)} &= \|\mathbb{E}[u - u_0 - \varepsilon u_1]\|_{H_0^1(D)} \\ &\leq \mathbb{E}[\|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)}] \\ &\leq \frac{\varepsilon^2 C_P}{a_{\min} a_{0,\min}^2} \|f\|_{L^2(D)} \left(\sum_{j=1}^L \|a_j\|_{L^\infty(D)} \right)^2. \end{aligned}$$

Compared to (1.37), there is no additional higher order term here but the constant for the term of order ε^2 is larger since the cross terms do not vanish and $a_{0,\min}^{-1}$ is replaced by a_{\min}^{-1} .

1.3.2 A posteriori error analysis

A posteriori error estimate in the $L_p^2(\Omega; H_0^1(D))$ norm

The goal is now to obtain an estimate of the error between u and $u_{0,h}$ which does not depend on the exact (unknown) solution. Let us define the jump of a function φ across an edge $e \in \mathcal{T}_h$ in the direction of \mathbf{n}_e by

$$[\varphi]_{\mathbf{n}_e}(\mathbf{x}) := \begin{cases} \lim_{t \rightarrow 0^+} (\varphi(\mathbf{x} + t\mathbf{n}_e) - \varphi(\mathbf{x} - t\mathbf{n}_e)) & \text{if } e \not\subset \partial D \\ 0 & \text{if } e \subset \partial D, \end{cases}$$

where \mathbf{n}_e denotes a normal vector to e of arbitrary (but fixed) direction for internal edges and the outwards normal to ∂D if $e \in \partial D$. Notice that the quantity $[\nabla \varphi \cdot \mathbf{n}_e]_{\mathbf{n}_e}$ is independent of the choice of the direction of the normal vector \mathbf{n}_e . We obtain the following residual type error upper bound, proceeding as in [118], which is based on the relation

$$\mathcal{A}(u - u_{0,h}, v; \mathbf{y}) = \mathcal{R}(v; \mathbf{y}_0) + [\mathcal{R}(v; \mathbf{y}) - \mathcal{R}(v; \mathbf{y}_0)] \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma$$

with

$$\mathcal{R}(v; \mathbf{y}) := F(v) - \mathcal{A}(u_{0,h}, v; \mathbf{y}),$$

where \mathcal{A} and F are defined in (1.15) and (1.16), respectively, and $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$.

Proposition 1.3.5. Let u be the weak solution of problem (1.11) and let $u_{0,h}$ be the solution of problem (1.30), respectively. There exists a constant $C > 0$ depending only on the constants in (1.26) and (1.28) such that

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{2}}{a_{\min}} [C\eta_1^2 + \eta_2^2]^{\frac{1}{2}}, \quad (1.42)$$

with

$$\eta_1^2 := \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{F}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \quad (1.43)$$

$$\eta_2^2 := \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 |\nabla u_{0,h}|^2. \quad (1.44)$$

Remark 1.3.6. We mention that the analysis is similar to the one given below if we consider the error in the energy norm $\|a_0^{1/2} \nabla(u - u_{0,h})\|_{L_p^2(\Omega; L^2(D))}$ instead of $\|\nabla(u - u_{0,h})\|_{L_p^2(\Omega; L^2(D))}$. The former should be preferred if the deterministic part a_0 of the diffusion coefficient a varies widely over D .

Proof. In the sequel, C will denote a constant whose value might change from one line to another. Let v be any function in $H_0^1(D)$. We have a.s. in Ω

$$\begin{aligned} \int_D a \nabla(u - u_{0,h}) \cdot \nabla v &= \int_D a \nabla u \cdot \nabla v - \int_D a \nabla u_{0,h} \cdot \nabla v \\ &= \underbrace{\int_D (f v - a_0 \nabla u_{0,h} \cdot \nabla v)}_{=: A_1} + \underbrace{\int_D (a_0 - a) \nabla u_{0,h} \cdot \nabla v}_{=: A_2}, \end{aligned} \quad (1.45)$$

where A_1 and A_2 correspond respectively to the error due to the finite element approximation of u_0 , solution to problem (1.20), and the error due to the truncation in the expansion (1.19) of u . We bound now each term separately, starting with A_2 . Using the expansion of a given by (1.12), we have

$$A_2 \leq \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \left(\int_D |\nabla v|^2 \right)^{\frac{1}{2}} = \varepsilon \left(\int_D \left(\sum_{j=1}^L a_j Y_j \right)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}. \quad (1.46)$$

For the first term A_1 , we use the relation $\int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D f v_h$ for all $v_h \in V_h$ with v_h the Clément interpolant of v together with interpolation results (1.26) to get

$$\begin{aligned} A_1 &\leq \sum_{K \in \mathcal{T}_h} \left(\int_K |f + \nabla \cdot (a_0 \nabla u_{0,h})|^2 \right)^{\frac{1}{2}} C h_K |v|_{H^1(N(K))} \\ &\quad + \sum_{e \in \mathcal{F}_h} \left(\int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right)^{\frac{1}{2}} C h_e^{\frac{1}{2}} |v|_{H^1(N(K_e))} \\ &\leq \sqrt{2} C \left[\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |f + \nabla \cdot (a_0 \nabla u_{0,h})|^2 + \sum_{e \in \mathcal{F}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right]^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}. \end{aligned} \quad (1.47)$$

We have used the fact that

$$\sum_{K \in \mathcal{T}_h} \|\nabla v\|_{L^2(N(K))}^2 \leq C_0 \|\nabla v\|_{L^2(D)}^2 \quad \text{and} \quad \sum_{e \in \mathcal{T}_h} \|\nabla v\|_{L^2(N(K_e))}^2 \leq C_0 \|\nabla v\|_{L^2(D)}^2$$

where C_0 depends on the maximum number of neighbours of each element in \mathcal{T}_h , which in turn depends on the constant in (1.23). Since a_{\min} is a lower bound for a , we deduce from (1.45) with $v = u(\cdot, \mathbf{Y}(\omega)) - u_{0,h} \in H_0^1(D)$ that a.s. in Ω we have

$$\int_D |\nabla(u - u_{0,h})|^2 \leq \frac{1}{a_{\min}} [A_1 + A_2].$$

Combining this last inequality with the bounds for A_1 and A_2 given by (1.47) and (1.46) respectively, we obtain a.s. in Ω

$$\begin{aligned} \|\nabla(u - u_{0,h})\|_{L^2(D)} &\leq \frac{1}{a_{\min}} \left\{ \sqrt{2}C \left[\sum_{K \in \mathcal{T}_h} h_K^2 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 \right. \right. \\ &\quad \left. \left. + \sum_{e \in \mathcal{T}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right]^{\frac{1}{2}} + \varepsilon \left(\int_D \left(\sum_{j=1}^L a_j Y_j \right)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \right\} \end{aligned} \quad (1.48)$$

and thus, taking the square of this last equation and using again $(a + b)^2 \leq 2(a^2 + b^2)$ yields

$$\begin{aligned} \|\nabla(u - u_{0,h})\|_{L^2(D)}^2 &\leq \frac{2}{a_{\min}^2} \left\{ 2C^2 \left(\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |f + \nabla \cdot (a_0 \nabla u_{0,h})|^2 \right. \right. \\ &\quad \left. \left. + \sum_{e \in \mathcal{T}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right) + \varepsilon^2 \int_D \left(\sum_{j=1}^L a_j Y_j \right)^2 |\nabla u_{0,h}|^2 \right\}. \end{aligned}$$

The *a posteriori* error estimate (1.42) is obtained taking the square root of the expected value on both sides of the last inequality and exploiting the independence of the random variables, namely that $\mathbb{E}[Y_i Y_j] = \delta_{ij}$ for $i, j = 1, \dots, L$. \square

Remark 1.3.7. In the one-dimensional case, we can take $v_h = r_h v$ the Lagrange interpolant of v and the sum over the edges (the discrete nodes here) vanishes. Indeed, any function and its Lagrange interpolant coincide at each node x_i , $i = 0, \dots, N_h$, of the considered discretization, or more precisely $v(x_i) - r_h v(x_i) = 0$ for all $i = 0, \dots, N_h$. Since (1.24) holds for e.g. $C = 2$, we can show that we have the following *a posteriori* error estimate

$$\mathbb{E} \left[\|u' - u'_{0,h}\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{2}}{a_{\min}} \left(4 \sum_{i=0}^{N_h-1} h_i^2 \int_{x_i}^{x_{i+1}} (f + (a_0 u'_{0,h})')^2 + \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 (u'_{0,h})^2 \right)^{\frac{1}{2}}, \quad (1.49)$$

where u' denotes the spatial derivative $\frac{\partial u(x, \omega)}{\partial x}$.

Remark 1.3.8. The computable quantity $\eta = (\eta_1^2 + \eta_2^2)^{\frac{1}{2}}$ can be used as an *a posteriori* error

estimator, which is reliable thanks to (1.42). It can be used to determine a mesh yielding comparable accuracy in h and ε , i.e. for balancing the error due to physical space discretization and the error due to the uncertainty. The spatial error estimator η_1 is efficient in the sense that it provides (up to a multiplicative constant depending only on a_{\max} and the regularity of the mesh) a lower bound for the error plus the other contribution η_2 and oscillation terms, the proof being similar to the one given in Appendix 1.B. Even though we have not been able to prove that η_2 in (1.44) also provides a similar lower bound, the estimator η appears to be efficient for all the numerical experiments we have considered.

We give below an *a posteriori* error estimator for the error $\|u - u_{0,h}\|_{L^2_p(\Omega; H^1_0(D))}$ for which both upper and lower bounds can be shown. The spatial error estimator is the same, namely η_1 given in (1.43), while the stochastic error estimator is obtained by computing (approximately) the dual norm of the residual $r(v; \mathbf{y}) := \mathcal{R}(v; \mathbf{y}) - \mathcal{R}(v; \mathbf{y}_0)$. Here, we only give the statement of the error estimator and we refer to Appendix 1.B for more details including the proof of the bounds. Let $W_{j,h} \in V_h$ be the solution of the problem

$$\int_D \nabla W_{j,h} \cdot \nabla v_h = - \int_D a_j \nabla u_{0,h} \cdot \nabla v_h \quad \forall v_h \in V_h.$$

The error estimator can then be defined as

$$\hat{\eta}^2 = (\eta_1^2 + \hat{\eta}_2^2)^{\frac{1}{2}} \quad \text{with} \quad \hat{\eta}_2^2 := \varepsilon^2 \sum_{j=1}^L \|\nabla W_{j,h}\|_{L^2(D)}^2. \quad (1.50)$$

Notice that the computation of $\hat{\eta}$ in (1.50) requires the solution of L additional Poisson problems compared to the error estimator η based on (1.42), and a strategy to reduce the computational cost could be to introduce auxiliary local problems defined on an element or a small subdomain, see e.g. [15, 107] and references therein. We mention that the extra computational effort to get $\hat{\eta}_2$ instead of η_2 is apparently not worth to pay in the present case, since the *a posteriori* error estimator based on Proposition 1.3.5 is efficient, at least for all the numerical experiments we have performed.

A posteriori error estimate in the $L^2_p(\Omega; L^2(D))$ norm

We now give an *a posteriori* error estimate of the error between u and $u_{0,h}$ in the L^2 norm in space, which leads to a gain of one order in h . To do so, we use a duality argument (often called the *Aubin-Nitsche* trick). We thus consider the dual problem of problem (1.11) given by:

find $\phi : D \times \Omega \rightarrow \mathbb{R}$ such that P -almost everywhere:

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega) \nabla \phi(\mathbf{x}, \omega)) &= u(\mathbf{x}, \omega) - u_{0,h}(\mathbf{x}) & \mathbf{x} \in D \\ \phi(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D, \end{cases} \quad (1.51)$$

whose pointwise in $\mathbf{y} \in \Gamma$ weak form reads:

find $\phi(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla \phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D (u(\mathbf{x}, \mathbf{y}) - u_{0,h}(\mathbf{x})) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (1.52)$$

Under regularity conditions on D , we have the following *a posteriori* error upper bound, which implies that the convergence rate of the error is $\mathcal{O}(h^2 + \varepsilon)$ in that case. That is that we gain one order in h compared to the error in the $L_P^2(\Omega; H_0^1(D))$ norm. However, the order of the statistical error is not improved.

Proposition 1.3.9. *Let u and u_0 be the (weak) solutions of problems (1.11) and (1.20), respectively, and let $u_{0,h}$ be the solution of problem (1.30). If $\phi(\cdot, \mathbf{Y}(\omega)) \in H^2(D)$ and $\|\phi\|_{H^2(D)} \leq C\|u - u_{0,h}\|_{L^2(D)}$ a.s. in Ω , then there exist constants $C_1, C_2 > 0$ independent of u, h and ε such that*

$$\mathbb{E} \left[\|u - u_{0,h}\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \sqrt{2} [C_1 \eta_1^2 + C_2 \eta_2^2]^{\frac{1}{2}} \quad (1.53)$$

with

$$\eta_1^2 := \sum_{K \in \mathcal{T}_h} h_K^4 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{E}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]^2_{\mathbf{n}_e} \quad (1.54)$$

$$\eta_2^2 := \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 |\nabla u_{0,h}|^2. \quad (1.55)$$

Remark 1.3.10. *Since we assumed $a_j \in W^{1,\infty}(D)$, $j = 0, \dots, L$, the assumptions of Proposition 1.3.9 on the regularity of the dual solution ϕ are satisfied if, for instance, D is a convex polygon (see [84]). The constant C in $\|\phi\|_{H^2(D)} \leq C\|u - u_{0,h}\|_{L^2(D)}$ may depend on the uniform bounds of Y_j, a_j and ∇a_j and on ε_{\max} but is independent of ε .*

Proof. First note that if we take $v = u(\cdot, \mathbf{y}) - u_{0,h}$, ρ -a.e. in Γ , in (1.52), we directly get the L^2 norm in space of the error at the right-hand side. We thus only need to estimate the left-hand side by a quantity which does not depend on the exact solutions $u = u(\mathbf{x}, \mathbf{Y}(\omega))$ and $\phi = \phi(\mathbf{x}, \mathbf{Y}(\omega))$ of respectively the primal and dual problems. In what follows, all equations hold a.s. in Ω without specifically mentioning it. Since

$$\int_D a \nabla (u - u_{0,h}) \cdot \nabla v_h + \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v_h = 0 \quad \forall v_h \in V_h,$$

we have for any $v_h \in V_h$

$$\begin{aligned} \|u - u_{0,h}\|_{L^2(D)}^2 &= \int_D a \nabla (u - u_{0,h}) \cdot \nabla \phi \\ &= \int_D a \nabla (u - u_{0,h}) \cdot \nabla (\phi - v_h) - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v_h \\ &= \underbrace{\int_D f(\phi - v_h) - \int_D a_0 \nabla u_{0,h} \cdot \nabla (\phi - v_h)}_{=: A_1} - \underbrace{\int_D (a - a_0) \nabla u_{0,h} \cdot \nabla \phi}_{=: A_2}. \end{aligned} \quad (1.56)$$

1.3. Error analysis for the first order approximation

We now treat each term separately. For the first one, we follow the usual procedure. For any $v_h \in V_h$, we have

$$\begin{aligned} A_1 &= \sum_{K \in \mathcal{T}_h} \int_K f(\phi - v_h) - \sum_{K \in \mathcal{T}_h} \int_K a_0 \nabla(\phi - v_h) \nabla u_{0,h} \\ &\leq \sum_{K \in \mathcal{T}_h} \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)} \|\phi - v_h\|_{L^2(K)} + \sum_{e \in \mathcal{E}_h} \| [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \|_{L^2(e)} \|\phi - v_h\|_{L^2(e)}. \end{aligned}$$

If we take $v_h = r_h \phi$, the Lagrange interpolant of ϕ , thanks to the interpolation error estimate (1.25), the trace inequality and the standard elliptic regularity result $\|\phi\|_{H^2(D)} \leq C \|u - u_{0,h}\|_{L^2(D)}$ (see [31, 49] for instance), we obtain

$$\begin{aligned} A_1 &\leq C_1 \left[\left(\sum_{K \in \mathcal{T}_h} h_K^4 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 \right)^{\frac{1}{2}} + \left(\sum_{e \in \mathcal{E}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right)^{\frac{1}{2}} \right] \|\phi\|_{H^2(D)} \\ &\leq \sqrt{2} C_1 \left(\sum_{K \in \mathcal{T}_h} h_K^4 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{E}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right)^{\frac{1}{2}} \|u - u_{0,h}\|_{L^2(D)}, \end{aligned} \quad (1.57)$$

where C_1 is a constant whose value might change from one line to another. Consider now the second term A_2 of (1.56). We have

$$A_2 = - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla \phi \leq \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \|\nabla \phi\|_{L^2(D)},$$

and thus, it only remains to obtain an upper bound for $\|\nabla \phi\|_{L^2(D)}$. Taking $v = \phi$ in the weak form (1.52) of the dual problem yields

$$\int_D a \nabla \phi \cdot \nabla \phi = \int_D (u - u_{0,h}) \phi \leq \|u - u_{0,h}\|_{L^2(D)} \|\phi\|_{L^2(D)}.$$

Since a is bounded from below by a_{min} , thanks to the Poincaré inequality we get

$$a_{min} \|\nabla \phi\|_{L^2(D)}^2 \leq C_P \|u - u_{0,h}\|_{L^2(D)} \|\nabla \phi\|_{L^2(D)},$$

and thus

$$\|\nabla \phi\|_{L^2(D)} \leq \frac{C_P}{a_{min}} \|u - u_{0,h}\|_{L^2(D)}.$$

Therefore, A_2 can be bounded by

$$A_2 \leq \frac{C_P}{a_{min}} \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \|u - u_{0,h}\|_{L^2(D)}. \quad (1.58)$$

Inserting (1.57) and (1.58) into (1.56) yields

$$\begin{aligned} \|u - u_{0,h}\|_{L^2(D)} &\leq \sqrt{2}C_1 \left(\sum_{K \in \mathcal{T}_h} h_K^4 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{E}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right)^{\frac{1}{2}} \\ &\quad + \frac{C_P}{a_{\min}} \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}}, \end{aligned}$$

and thus

$$\begin{aligned} \|u - u_{0,h}\|_{L^2(D)}^2 &\leq 2 \left[2C_1^2 \left(\sum_{K \in \mathcal{T}_h} h_K^4 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{E}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right) \right. \\ &\quad \left. + \frac{C_P^2}{a_{\min}^2} \int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right]. \end{aligned} \quad (1.59)$$

Since $\mathbb{E}[(a - a_0)^2] = \varepsilon^2 \sum_{j=1}^L a_j^2$, the result follows from taking first the expected value and then the square root on both sides of (1.59). \square

Goal-oriented error estimate

The *a posteriori* error estimates obtained so far yield upper bounds on the error in global norms. In the case where we are interested in a particular quantity of interest, e.g. point values or contour integrals, these estimates may not be appropriate. Goal-oriented error estimation has thus been developed (see [13, 22, 100] and [4, 33, 35, 92] and the references therein for the deterministic and stochastic framework, respectively) to bound a given functional using optimal control techniques (based on a duality-argument). In this section we only sketch the derivation of a goal-oriented error upper bound for the first-order FEM approximation $u_{0,h}$. Assume that we are interested in computing $Q(u)$ with Q a linear functional on $H_0^1(D)$ representing a quantity of interest which depends on the random vector \mathbf{Y} only through the random solution $u(\cdot, \mathbf{Y})$ itself. We introduce the dual problem:

$$\text{find } \varphi(\cdot, \mathbf{y}) \in H_0^1(D) \text{ such that } \mathcal{A}(v, \varphi(\cdot, \mathbf{y}); \mathbf{y}) = Q(v), \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma, \quad (1.60)$$

where \mathcal{A} is defined by (1.15). Let $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$ denotes the nominal value for \mathbf{Y} , for which $a(\mathbf{x}, \mathbf{y}_0) = a_0(\mathbf{x})$, and let φ_0 be the deterministic solution of (1.60) with $\mathbf{y} = \mathbf{y}_0$ and $\varphi_{0,h}$ its FE approximation. Using the fact that Q does not depend on \mathbf{Y} explicitly, we can easily show that

a.s. in Ω

$$\begin{aligned}
 Q(u(\cdot, \mathbf{Y}(\omega))) - Q(u_{0,h}) &= \underbrace{\int_D f \varphi_0}_{=:A_1} - \underbrace{\int_D a_0 \nabla u_{0,h} \cdot \nabla \varphi_0}_{=:A_2} - \underbrace{\int_D (a - a_0) \nabla u_{0,h} \cdot \nabla \varphi_{0,h}}_{=:A_2} \\
 &\quad - \underbrace{\int_D (a - a_0) \nabla u_{0,h} \cdot \nabla (\varphi_0 - \varphi_{0,h})}_{=:A_3} - \underbrace{\int_D (a - a_0) \nabla (u - u_{0,h}) \cdot \nabla \varphi_{0,h}}_{=:A_4} \\
 &\quad - \underbrace{\int_D (a - a_0) \nabla (u - u_{0,h}) \cdot \nabla (\varphi_0 - \varphi_{0,h})}_{=:A_5}.
 \end{aligned}$$

The first term A_1 , which is deterministic and of order h^2 , can be bounded using standard techniques such as the Dual-weighted residual (DWR) method (see e.g. [13, 22]) or using the parallelogram identity as proposed by Oden and Prudhomme in [100]. In the DWR method, the upper bound depends on the unknown influence function φ_0 , either through $|\varphi_0|_{H^2(K)}$ or $\|\nabla(\varphi_0 - \varphi_{0,h})\|_{L^2(K)}$, K being an element of the mesh. In the former case, the H^2 semi-norm can be estimated by a discrete analogue and in the latter case, the influence function might be replaced by a discrete solution computed on a space richer than V_h or by post-processing. All the other terms can be bounded provided we can obtain an upper bound for $\|\nabla(u - u_{0,h})\|_{L^2(D)}$, which is given by (1.48), as well as an upper bound for $\|\nabla(\varphi_0 - \varphi_{0,h})\|_{L^2(D)}$ which can be done as in the previous sections. Moreover, based on the results obtained in the previous sections we have

$$A_1 = \mathcal{O}(h^2), A_2 = \mathcal{O}(\varepsilon), A_3 = \mathcal{O}(h\varepsilon), A_4 = \mathcal{O}(h\varepsilon + \varepsilon^2) \text{ and } A_5 = \mathcal{O}(h^2\varepsilon + \varepsilon^2h).$$

We might be interested in estimating the expectation or the variance of $Q(u(\cdot, \mathbf{Y})) - Q(u_{0,h})$. In the former case, notice that $\mathbb{E}[A_2] = \mathbb{E}[A_3] = 0$ and since A_1 is a deterministic quantity, we have

$$\mathbb{E}[Q(u) - Q(u_{0,h})] = A_1 + \mathbb{E}[A_4] + \mathbb{E}[A_5].$$

Moreover, the term $\mathbb{E}[A_5]$ is of higher order than $\mathbb{E}[A_4]$ and can thus be neglected, so that we have $\mathbb{E}[Q(u) - Q(u_{0,h})] = \mathcal{O}(h^2 + h\varepsilon + \varepsilon^2)$. In the latter case, we have

$$\mathbb{E}[|Q(u) - Q(u_{0,h})|^2] \leq 5(A_1^2 + \mathbb{E}[A_2^2] + \mathbb{E}[A_3^2] + \mathbb{E}[A_4^2] + \mathbb{E}[A_5^2]).$$

As before, the term $\mathbb{E}[A_5^2]$ can be neglected and we have $\mathbb{E}[|Q(u) - Q(u_{0,h})|^2]^{\frac{1}{2}} = \mathcal{O}(h^2 + \varepsilon + h\varepsilon)$. Moreover, if the mesh space h is chosen such that $h^2 \sim \varepsilon$, then both terms $\mathbb{E}[A_3^2]$ and $\mathbb{E}[A_4^2]$ can also be omitted in the estimation of the variance and $\mathbb{E}[|Q(u) - Q(u_{0,h})|^2]^{\frac{1}{2}} = \mathcal{O}(h^2 + \varepsilon)$.

Finally, we mention that the estimate on the variance of $Q(u) - Q(u_{0,h})$ can be used to have a rough estimate on the failure probability $P(Q(u) > Q_{crit})$ with some critical value Q_{crit}

sufficiently far from $Q(u_{0,h})$. Indeed, using the Bienaymé-Tchebychev inequality we have

$$P(Q(u) > Q_{crit}) \leq \frac{\mathbb{E}[(Q(u) - Q(u_{0,h}))^2]}{(Q(u_{0,h}) - Q_{crit})^2}.$$

1.4 Error analysis for higher order approximations

In this section, we generalize the *a posteriori* error estimate of Proposition 1.3.5 to higher order approximation, that is when more terms in the expansion (1.19) of u are taken into account. We start by giving the result for the second order approximation before generalizing to any order of approximation.

1.4.1 Second order approximation

In this section, instead of considering the error between u and $u_{0,h}$, we will give an estimation of the error between u and u_h^1 , the FE approximation of $u^1 := u_0 + \varepsilon u_1 = u_0 + \varepsilon \sum_{j=1}^L U_j Y_j$, where U_j is the solution of problem (1.21). Since the random variables Y_j , $j = 1, \dots, L$, are assumed to be bounded, the error due to the stochastic approximation of u is of order ε^2 in this case. Indeed, if we do not take the finite element approximation error into account, we have a.s. in Ω (see (1.38) for details)

$$\int_D a \nabla(u - u^1) \cdot \nabla v = -\varepsilon^2 \int_D \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_i \cdot \nabla v, \quad (1.61)$$

and only the term of order ε^2 remains. Let us now take the error due to the approximation of u^1 by $u_h^1 := u_{0,h} + \varepsilon u_{1,h}$ into account, where $u_{1,h} = \sum_{j=1}^L Y_j U_{j,h}$ and, for $j = 1, \dots, L$, $U_{j,h}$ is the solution of

$$\int_D a_0 \nabla U_{j,h} \cdot \nabla v_h = - \int_D a_j \nabla u_{0,h} \cdot \nabla v_h \quad \forall v_h \in V_h. \quad (1.62)$$

To simplify the notation, we define

$$w_{j,h} := a_0 \nabla U_{j,h} + a_j \nabla u_{0,h}.$$

We can show that, if the solution is regular enough in physical space, the convergence of the error is in $\mathcal{O}(h + \varepsilon h + \varepsilon^2)$, i.e., that for a mesh size h of order ε^2 , the error is divided by 4 when ε is halved. The following proposition provides an *a posteriori* error estimate.

Proposition 1.4.1. *Let u be the weak solution of problem (1.11) and let $u_{0,h}$ and $U_{j,h}$, $j = 1, \dots, L$, be the solutions of problems (1.30) and (1.62), respectively. There exist two constants $C_1, C_2 > 0$ depending only on the constants in (1.26) and (1.28) such that*

$$\mathbb{E} \left[\|\nabla(u - u_h^1)\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{3}}{a_{min}} [C_1 \eta_1^2 + C_2 \eta_2^2 + \eta_3^2]^{\frac{1}{2}}, \quad (1.63)$$

with

$$\eta_1^2 = \sum_K h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_e h_e \| [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \|_{L^2(e)}^2, \quad (1.64)$$

$$\eta_2^2 = \varepsilon^2 \left(\sum_K h_K^2 \int_K \sum_{j=1}^L (\nabla \cdot w_{j,h})^2 + \sum_e h_e \int_e \sum_{j=1}^L [w_{j,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \right), \quad (1.65)$$

$$\eta_3^2 = \varepsilon^4 \left(\int_D \sum_{i=1}^L a_i^2 |\nabla U_{i,h}|^2 \mathbb{E}[Y_i^4] + \int_D \sum_{\substack{i,j=1 \\ i \neq j}}^L [a_i^2 |\nabla U_{j,h}|^2 + 2a_i a_j \nabla U_{i,h} \cdot \nabla U_{j,h}] \right). \quad (1.66)$$

From (1.63), we see that the error splits into three parts, namely the error due to the FE approximation of u_0 , the FE approximation of the U_j , $j = 1, \dots, L$ and the truncation in the expansion of u with respect to ε .

Proof. For any $v \in H_0^1(D)$ and a.s. in Ω we have

$$\begin{aligned} \int_D a \nabla(u - u_h^1) \cdot \nabla v &= \underbrace{\int_D f v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v}_{=: A_1} - \varepsilon \underbrace{\int_D \sum_{j=1}^L Y_j (a_0 \nabla U_{j,h} + a_j \nabla u_{0,h}) \cdot \nabla v}_{=: A_2} \\ &\quad - \varepsilon \underbrace{\int_D (a - a_0) \nabla u_{1,h} \cdot \nabla v}_{=: A_3}. \end{aligned} \quad (1.67)$$

where A_1 and A_2 are respectively the residual for $u_{0,h}$ and for $U_{j,h}$, for $j = 1, \dots, L$, while A_3 is due to the truncation in the expansion (1.19) of u . Let us treat each term separately. The first term A_1 is bounded by (see Section 1.3)

$$A_1 \leq C_1 \left[\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{T}_h} h_e \| [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \|_{L^2(e)}^2 \right]^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}. \quad (1.68)$$

Let us consider now the term A_2 . Since $\int_D w_{j,h} \cdot \nabla v_h = 0$ for all $v_h \in V_h$, we have

$$\begin{aligned} A_2 &= -\varepsilon \int_D \sum_{j=1}^L Y_j w_{j,h} \cdot \nabla(v - \mathcal{I}_h v) \\ &= \varepsilon \sum_{K \in \mathcal{T}_h} \int_K \left(\sum_{j=1}^L Y_j \nabla \cdot w_{j,h} \right) (v - \mathcal{I}_h v) + \varepsilon \sum_{e \in \mathcal{T}_h} \int_e \left[\sum_{j=1}^L Y_j w_{j,h} \cdot \mathbf{n}_e \right]_{\mathbf{n}_e} (v - \mathcal{I}_h v) \\ &\leq C_2 \left(\sum_{K \in \mathcal{T}_h} \varepsilon^2 h_K^2 \left\| \sum_{j=1}^L Y_j \nabla \cdot w_{j,h} \right\|_{L^2(K)}^2 + \sum_{e \in \mathcal{T}_h} \varepsilon^2 h_e \left\| \left[\sum_{j=1}^L Y_j w_{j,h} \cdot \mathbf{n}_e \right]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \right)^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}, \end{aligned} \quad (1.69)$$

Chapter 1. Elliptic model problems with random diffusion coefficient

where C_2 depends only on the interpolation constants that appear in (1.26) and (1.28). Finally, we estimate the last term A_3 . We have

$$\begin{aligned} A_3 &= -\varepsilon \int_D \left(\varepsilon \sum_{j=1}^L Y_j a_j \right) \nabla \left(\sum_{i=1}^L Y_i U_{i,h} \right) \cdot \nabla v = -\varepsilon^2 \int_D \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_{i,h} \cdot \nabla v \\ &\leq \varepsilon^2 \left\| \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_{i,h} \right\|_{L^2(D)} \|\nabla v\|_{L^2(D)}. \end{aligned} \quad (1.70)$$

Since a is bounded from below by a_{\min} , combining (1.67) with (1.68), (1.69) and (1.70) with $v = u(\cdot, \mathbf{Y}(\omega)) - u_h^1(\cdot, \mathbf{Y}(\omega)) \in H_0^1(D)$ yields a.s. in Ω

$$\begin{aligned} \|\nabla(u - u_h^1)\|_{L^2(D)} &\leq \frac{\sqrt{3}}{a_{\min}} \left[C_1^2 \left(\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{T}_h} h_e \|[a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}\|_{L^2(e)}^2 \right) \right. \\ &\quad + C_2^2 \left(\sum_{K \in \mathcal{T}_h} \varepsilon^2 h_K^2 \left\| \sum_{j=1}^L Y_j \nabla \cdot w_{j,h} \right\|_{L^2(K)}^2 + \sum_{e \in \mathcal{T}_h} \varepsilon^2 h_e \left\| \left[\sum_{j=1}^L Y_j w_{j,h} \cdot \mathbf{n}_e \right]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \right) \\ &\quad \left. + \varepsilon^4 \left\| \sum_{i,j=1}^L Y_i Y_j a_j \nabla U_{i,h} \right\|_{L^2(D)}^2 \right]^{\frac{1}{2}}, \end{aligned}$$

using the inequality $(a + b + c) \leq \sqrt{3}(a^2 + b^2 + c^2)^{\frac{1}{2}}$. To conclude the proof, it only remains to take the expected value on both sides of the square of this last inequality. By linearity of the expected value, we can consider the three terms of the right-hand side separately. The first term is a deterministic quantity and thus, taking the expected value on it has no effect. For the two other terms, we just have to evaluate $\mathbb{E}[Y_i Y_j]$ for $1 \leq i, j \leq L$ and $\mathbb{E}[Y_i Y_j Y_k Y_l]$ for $1 \leq i, j, k, l \leq L$. Since the random variables are assumed to be independent, with zero mean and unit variance, we have $\mathbb{E}[Y_i Y_j] = \delta_{ij}$ and

$$\mathbb{E}[Y_i Y_j Y_k Y_l] = \begin{cases} \mathbb{E}[Y_j^4] & \text{if } i = j = k = l \\ 1 & \text{if the indices are pairwise equal} \\ 0 & \text{otherwise.} \end{cases}$$

Let us write

$$B := \sum_{i,j,k,l=1}^L Y_i Y_j Y_k Y_l a_j a_k \nabla U_{i,h} \cdot \nabla U_{l,h},$$

which we split into three parts B_1 (all indices are equal), B_2 (two pairs of indices) and B_3 (remaining indices). Thanks to the linearity of expectation, we have $\mathbb{E}[B] = \mathbb{E}[B_1] + \mathbb{E}[B_2] + \mathbb{E}[B_3]$. First, we can notice that $\mathbb{E}[B_3] = 0$. Moreover, the contribution to $\mathbb{E}[B]$ when $i = j = k = l$ is

$$\mathbb{E}[B_1] = \sum_{i=1}^L a_i^2 |\nabla U_{i,h}|^2 \mathbb{E}[Y_i^4].$$

Let us consider now all the cases when we have pairwise equal pairs of indices. Out of 4 indices, there are three different ways to form two pairs of indices, namely $(j = k, i = l)$, $(j = i, k = l)$

and $(j = l, k = i)$. Since the two last cases lead to the same result, we get

$$\mathbb{E}[B_2] = \sum_{\substack{i,j=1 \\ i \neq j}}^L a_j^2 |\nabla U_{i,h}|^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^L a_i a_j \nabla U_{i,h} \cdot \nabla U_{j,h}.$$

Altogether, we finally get

$$\mathbb{E}[B] = \sum_{i=1}^L a_i^2 |\nabla U_{i,h}|^2 \mathbb{E}[Y_i^4] + \sum_{\substack{i,j=1 \\ i \neq j}}^L [a_i^2 |\nabla U_{j,h}|^2 + 2a_i a_j \nabla U_{i,h} \cdot \nabla U_{j,h}],$$

which concludes the proof. \square

1.4.2 Generalization

Suppose now that the random solution u of problem (1.11) is expanded with respect to ε up to order $N \in \mathbb{N}$, see (1.19). For $1 \leq n \leq N$, let us write

$$u_n(\mathbf{x}, \mathbf{Y}(\omega)) = \sum_{j_1, j_2, \dots, j_n=1}^L U_{j_1 j_2 \dots j_n}(\mathbf{x}) Y_{j_1}(\omega) Y_{j_2}(\omega) \dots Y_{j_n}(\omega) \quad (1.71)$$

the n^{th} term in the expansion. The L^n functions $U_{j_1 j_2 \dots j_n}$ are obtained by solving for $j_1, j_2, \dots, j_n = 1, \dots, L$ the deterministic problem

$$\begin{cases} -\operatorname{div}(a_{j_1}(\mathbf{x}) \nabla U_{j_2 \dots j_n}(\mathbf{x}) + a_0(\mathbf{x}) \nabla U_{j_1 \dots j_n}(\mathbf{x})) &= 0 & \mathbf{x} \in D \\ U_{j_1 \dots j_n}(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad (1.72)$$

using the solutions $U_{j_2 \dots j_n}$, $j_2, \dots, j_n = 1, \dots, L$, obtained for the $(n-1)^{th}$ order term. Proceeding as in Sections 1.3 and 1.4.1, it is easy to show that the error due to the truncation in the expansion of u is of order ε^{N+1} . More precisely, we have for any $v \in H_0^1(D)$ and almost surely

$$\int_D a \nabla \left(u - \sum_{n=0}^N \varepsilon^n u_n \right) \cdot \nabla v = -\varepsilon^{N+1} \sum_{j_0, j_1, \dots, j_N=1}^L Y_{j_0} Y_{j_1} \dots Y_{j_N} \int_D a_{j_0} \nabla U_{j_1 j_2 \dots j_N} \cdot \nabla v. \quad (1.73)$$

Since Y_j , $j = 1, \dots, L$ are bounded, in particular they have bounded $2(N+1)^{th}$ moment. When the various deterministic functions are approximated using finite elements, if the solution is regular enough in physical space then the error $u - \sum_{n=0}^N \varepsilon^n u_{n,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm is of order

$$h + \varepsilon h + \varepsilon^2 h + \dots + \varepsilon^N h + \varepsilon^{N+1}.$$

The error in $\mathcal{O}(\varepsilon^n h)$, $0 \leq n \leq N$, corresponds to the error made when the functions $U_{j_1 \dots j_n}$ (u_0 for $n = 0$) are replaced by their FE approximation $U_{j_1 \dots j_n, h}$ (resp. $u_{0,h}$). An *a posteriori* error estimate can thus easily be obtained as follows. First, the term in $\mathcal{O}(h)$, which corresponds to the residual for $u_{0,h}$, is obtained by estimating $\int_D (f v - a_0 \nabla u_{0,h} \cdot \nabla v)$, see (1.47). For the term in $\mathcal{O}(h \varepsilon^n)$, $n = 1, \dots, N$, it suffices to estimate for $j_1, \dots, j_n = 1, \dots, L$ the residual defined for any

$v \in H_0^1(D)$ by

$$\langle \mathcal{R}(U_{j_1 \dots j_n, h}), v \rangle := \int_D (a_{j_1} \nabla U_{j_2 \dots j_n, h} + a_0 \nabla U_{j_1 \dots j_n, h}) \cdot \nabla v,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing bracket. For an explicit error estimate, computable up to multiplicative interpolation constants, we finally need to express the expectation of the product of n random variables $\mathbb{E}[Y_{j_1} \cdots Y_{j_n}]$ for all combinations of indices and for $n = 1, \dots, 2(N+1)$. More precisely, we can show the following result.

Proposition 1.4.2. *Let u be the weak solution of problem (1.11) and $u_h^N = \sum_{n=0}^N \varepsilon^n u_{n,h}$, where $u_{n,h}$ is the FE approximation of u_n given by (1.71). There exist $N+1$ constants $C_n > 0$, $n = 0, 1, \dots, N$, depending only on the constants in (1.26) and (1.28) such that*

$$\mathbb{E} \left[\|\nabla(u - u_h^N)\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{N+2}}{a_{\min}} \left[C_0 \eta_0^2 + \sum_{n=1}^N C_n \eta_n^2 + \eta_{N+1}^2 \right]^{\frac{1}{2}}, \quad (1.74)$$

with

$$\begin{aligned} \eta_0^2 &= \sum_K h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_e h_e \| [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \|_{L^2(e)}^2, \\ \eta_n^2 &= \varepsilon^{2n} \mathbb{E} \left[\sum_K h_K^2 \left\| \sum_{j_1, \dots, j_n=1}^L Y_{j_1} \cdots Y_{j_n} \nabla \cdot w_{j_1 \dots j_n, h} \right\|_{L^2(K)}^2 \right. \\ &\quad \left. + \sum_e h_e \left\| \sum_{j_1, \dots, j_n=1}^L Y_{j_1} \cdots Y_{j_n} w_{j_1 \dots j_n, h} \cdot \mathbf{n}_e \right\|_{L^2(e)}^2 \right] \\ \eta_{N+1}^2 &= \varepsilon^{2(N+1)} \mathbb{E} \left[\left\| \sum_{j_0, j_1, \dots, j_N=1}^L Y_{j_0} Y_{j_1} \cdots Y_{j_N} a_{j_0} \nabla U_{j_1 \dots j_N, h} \right\|_{L^2(D)}^2 \right], \end{aligned}$$

where

$$w_{j_1 \dots j_n, h} := a_{j_1} \nabla U_{j_2 \dots j_n, h} + a_0 \nabla U_{j_1 \dots j_n, h} \quad j_1, \dots, j_n = 1, \dots, L.$$

Proceeding similarly, this generalization can also be applied to the other error estimates we obtained in Section 1.3. Finally, notice that the constant $\sqrt{N+2}$ that appears in (1.74) can be avoided thanks to the triangle inequality for the $L_p^2(\Omega)$ norm, yielding an upper bound of the form $a_{\min}^{-1} (C_0 \eta_0 + \dots + C_N \eta_N + \eta_{N+1})$. The same holds for all the error estimates obtained in Sections 1.3 and 1.4.1.

1.5 Extension to nonlinear problems

Keeping the same notations as in the previous sections, we are now interested in solving problems of the form:

find $u : D \times \Omega \rightarrow \mathbb{R}$ such that almost surely:

$$\begin{cases} F(a, u) = 0 & \text{in } D \\ u = 0 & \text{on } \partial D, \end{cases} \quad (1.75)$$

where F is a smooth nonlinear mapping that depends on the uncertain input a given by (1.12). Again, the random solution u is expanded with respect to ε up to a certain order

$$u(\mathbf{x}, \mathbf{Y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{Y}(\omega)) + \mathcal{O}(\varepsilon^2).$$

Formally, we have

$$F(a, u) = F(a_0, u_0) + D_a F(a_0, u_0)(a - a_0) + D_u F(a_0, u_0)(u - u_0) + \mathcal{O}(\varepsilon^2),$$

where D_a and D_u denote the Fréchet derivatives with respect to a and u respectively, the deterministic part u_0 of u is the solution of the (nonlinear) problem

$$\begin{cases} F(a_0, u_0) = 0 & \text{in } D \\ u_0 = 0 & \text{on } \partial D, \end{cases} \quad (1.76)$$

while the U_j in $u_1 = \sum_{j=1}^L Y_j U_j$ can be found by solving the (linear) problems

$$\begin{cases} D_a F(a_0, u_0)(a_j) + D_u F(a_0, u_0)(U_j) = 0 & \text{in } D \\ U_j = 0 & \text{on } \partial D, \end{cases} \quad j = 1, \dots, L. \quad (1.77)$$

We can directly see one of the advantages of expanding the solution as proposed here, namely that a single nonlinear problem must be solved to find u_0 , the other problems being linear. A new FE solver corresponding to (1.77) has to be implemented to approximate the U_j , $j = 1, \dots, L$.

In the case of quasi-linear problems, the error analysis is very similar to the linear case considered in Section 1.1. Indeed, under certain conditions such as well-posedness of the problem, only the part of the estimate corresponding to the residual error in the physical space has to be changed in the *a posteriori* estimate of the error between u and $u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm. For instance, let us consider problem (1.75) with

$$F(a(\mathbf{x}, \omega), u(\mathbf{x}, \omega)) := -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) + u^3(\mathbf{x}, \omega) - f(\mathbf{x}). \quad (1.78)$$

This well-posed problem has a unique solution in $L_P^2(\Omega; H_0^1(D))$ and we can show the following *a posteriori* error estimate for $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$, where $u_{0,h} \in V_h$ is the deterministic solution of

$$\int_D a_0 \nabla u_{0,h} \cdot \nabla v_h + \int_D u_{0,h}^3 v_h = \int_D f v_h \quad \forall v_h \in V_h. \quad (1.79)$$

Proposition 1.5.1. *Let u be the weak solution of problem (1.75) with F given by (1.78), and let $u_{0,h}$ be the solution of (1.79). There exists a constant $C > 0$ depending only on the constants in*

(1.26) and (1.28) such that

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{C}{a_{\min}} [\eta_1^2 + \eta_2^2]^{\frac{1}{2}},$$

with

$$\begin{aligned} \eta_1^2 &:= \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (f - u_{0,h}^3 + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{E}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \\ \eta_2^2 &:= \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 |\nabla u_{0,h}|^2. \end{aligned}$$

Proof. Since the proof is very similar to the one of Proposition 1.3.5, we only give the key ingredients here. First, for any $v \in V$ we have almost surely

$$\int_D a \nabla(u - u_{0,h}) \cdot \nabla v = \int_D (f - u_{0,h}^3) v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v - \int_D (u^3 - u_{0,h}^3) v.$$

Then, for $v = u - u_{0,h}$ the last term in the above equality is non-positive. Indeed, using that

$$u^3 - u_{0,h}^3 = \int_0^1 3(u_{0,h} + t(u - u_{0,h}))^2 (u - u_{0,h}) dt,$$

we get

$$- \int_D (u^3 - u_{0,h}^3)(u - u_{0,h}) = - \int_D \int_0^1 3(u_{0,h} + t(u - u_{0,h}))^2 (u - u_{0,h})^2 dt \leq 0.$$

Therefore, this term can be omitted since we are looking for an upper bound of the error. \square

Another example is the following. Let $k > 0$ be such that $\frac{kC_p^2}{a_{\min}} < 1$, or in other words $\frac{kC_p^2}{a_{\min}} \leq 1 - \delta$ for any $\delta \in (0, 1)$. If we take

$$F(a(\mathbf{x}, \omega), u(\mathbf{x}, \omega)) := -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) - g(u(\mathbf{x}, \omega)) \quad (1.80)$$

in problem (1.75), where g is a Lipschitz function with Lipschitz constant k , then we can show the well-posedness of the problem and the following *a posteriori* error estimate for the error $u - u_{0,h}$, where $u_{0,h} \in V_h$ is the deterministic solution of

$$\int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D g(u_{0,h}) v_h \quad \forall v_h \in V_h. \quad (1.81)$$

Proposition 1.5.2. *Let u be the weak solution of problem (1.75) with F given by (1.80), and let $u_{0,h}$ be the solution of (1.81). There exists a constant $C > 0$, depending only on δ and the constants in (1.26) and (1.28), such that*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{C}{a_{\min}} [\eta_1^2 + \eta_2^2]^{\frac{1}{2}},$$

with

$$\begin{aligned}\eta_1^2 &:= \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (g(u_{0,h}) + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{F}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \\ \eta_2^2 &:= \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 |\nabla u_{0,h}|^2.\end{aligned}$$

Proof. Again, we only give the key ingredients of the proof. First, for any $v \in V$ we have almost surely

$$\begin{aligned}\int_D a \nabla(u - u_{0,h}) \cdot \nabla v &= \underbrace{\int_D g(u_{0,h}) v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v}_{=: A(v)} \\ &\quad - \int_D (g(u) - g(u_{0,h})) v.\end{aligned}\tag{1.82}$$

With $v = u - u_{0,h}$, the last term is bounded by

$$- \int_D (g(u) - g(u_{0,h}))(u - u_{0,h}) \leq k C_P^2 \|\nabla(u - u_{0,h})\|_{L^2(D)}^2.\tag{1.83}$$

Since

$$a_{\min} \|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \leq \int_D a |\nabla(u - u_{0,h})|^2,$$

taking (1.83) to the left-hand side of (1.82) and using $k C_P^2 \leq a_{\min}(1 - \delta)$ yield

$$a_{\min} \delta \|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \leq A(u - u_{0,h}).$$

A bound on $A(u - u_{0,h})$, which contains the residual for u_0 and a term of order ε , is found proceeding exactly as in the proof of Proposition 1.3.5. \square

The constant C that appears in the error estimate of Proposition 1.5.2 is of order δ^{-1} , and thus explodes when δ tends to zero, i.e. when $\frac{k C_P^2}{a_{\min}}$ is close to one. In practise, it is usual to restrict the analysis to Lipschitz function with Lipschitz constant k such that $k \leq \frac{a_{\min}}{2 C_P^2}$, so that $\delta \geq \frac{1}{2}$.

Finally, let us consider an example where the uncertain coefficient is associated to the nonlinear term, namely the problem (1.75) with

$$F(a(\mathbf{x}, \omega), u(\mathbf{x}, \omega)) = -\Delta u(\mathbf{x}, \omega) + a(\mathbf{x}, \omega) u^3(\mathbf{x}, \omega) - f(\mathbf{x}).\tag{1.84}$$

In this case, we can show the well-posedness of the problem and the following *a posteriori* error estimate in $H_0^1(D)$ -norm in physical space for the first order approximation $u \approx u_{0,h}$, where $u_{0,h}$ is the solution of

$$\int_D \nabla u_{0,h} \cdot \nabla v_h + \int_D a_0 u_{0,h}^3 v_h = \int_D f v_h \quad \forall v_h \in V_h.\tag{1.85}$$

Proposition 1.5.3. *Let u be the weak solution of problem (1.75) with F given by (1.84), and let $u_{0,h}$ be the solution of (1.85). There exists a constant $C > 0$ depending only on the constants in (1.26) and (1.28) such that*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq C [\eta_1^2 + \eta_2^2]^{\frac{1}{2}},$$

with

$$\begin{aligned} \eta_1^2 &:= \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (f + \Delta u_{0,h} - a_0 u_{0,h}^3)^2 + \sum_{e \in \mathcal{T}_h} h_e \int_e [\nabla u_{0,h} \cdot \mathbf{n}_e]^2_{\mathbf{n}_e} \\ \eta_2^2 &:= \varepsilon^2 \int_D \sum_{j=1}^L a_j^2 u_{0,h}^6. \end{aligned}$$

Proof. The proof is based on the relations

$$\int_D \nabla(u - u_{0,h}) \cdot \nabla v = \int_D f v - \int_D a_0 u_{0,h}^3 v - \int_D \nabla u_{0,h} \cdot \nabla v - \int_D (a u^3 - a_0 u_{0,h}^3) v$$

and

$$- \int_D (a u^3 - a_0 u_{0,h}^3) v = - \int_D a \int_0^1 3(u_{0,h} + t(u - u_{0,h}))^2 (u - u_{0,h}) dt v - \int_D (a - a_0) u_{0,h}^3 v.$$

Since a is positive, the first term of the right-hand side of the last equality is less or equal to zero for $v = u - u_{0,h}$. \square

1.6 Computational costs

We perform here a comparison of the computational costs between the SC-FEM method [7, 124] and the one presented here, called *perturbation method* in the sequel, when comparable accuracy is reached. Briefly, the SC-FEM applied to the model problem (1.11) consists, given a set of (collocation) points $\{\mathbf{y}_k \in \Gamma, k = 1, \dots, N_c\}$, in finding $u_h(\cdot, \mathbf{y}_k) \in V_h$ such that

$$\int_D a(\mathbf{x}, \mathbf{y}_k) \nabla u_h(\mathbf{x}, \mathbf{y}_k) \cdot \nabla v_h(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v_h(\mathbf{x}) d\mathbf{x} \quad \forall v_h \in V_h$$

for $k = 1, \dots, N_c$ and building a global polynomial approximation

$$u_{h,N_c}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N_c} u_h(\mathbf{x}, \mathbf{y}_k) \psi_k(\mathbf{y}),$$

for appropriate multivariate polynomials $\{\psi_k\}_{k=1}^{N_c}$. Since the FEM is used to approximate the physical space in both methods (stochastic collocation and *perturbation*), we use the same mesh for the discretization of D . For a comparable statistical error, say an error with convergence rate of order ε^2 , we take $N = 1$ in the expansion (1.19) of u for the *perturbation*

method and use a sparse grid of level 1 for the SC method, based either on Clenshaw-Curtis (see [51]) or Gaussian abscissas. The construction of the sparse grid interpolant of level 1 is briefly described in the following. We refer to [65, 97, 124] for more details and the general construction of sparse grid of arbitrary level. First, the sparse grid interpolant of level 0 of a function $f(\mathbf{y})$, denoted $S_0 f$, is simply the evaluation of the function at (y_1^0, \dots, y_L^0) , where y_j^0 is the unique interpolation point in direction j . Next, for each variable y_j , we define the sequence of interpolation points at level $i \geq 1$ by $\{y_{j,k}^i, k = 1, \dots, m(i)\}$, where the number of collocation points $m(i)$ can be taken for instance as

$$m(i) = i + 1 \quad \text{or} \quad m(i) = \begin{cases} 1 & \text{if } i = 0 \\ 2^i + 1 & \text{if } i \geq 1. \end{cases}$$

The former choice for m corresponds to a total degree (TD) approximation space while the latter corresponds to a Smolyak one (see [11]). Notice that compared to the articles mentioned above, the level index i starts here at 0 instead of 1. We define then the one dimensional (Lagrange) interpolation operator in direction j at level $i = 1$ by

$$\mathcal{U}_j^1 f(y_1, \dots, y_L) := \sum_{k=1}^{m(1)} f(y_1^0, \dots, y_{j-1}^0, y_{j,k}^1, y_{j+1}^0, \dots, y_L^0) \left(\prod_{l=1, l \neq k}^{m(1)} \frac{y_j - y_{j,l}^1}{y_{j,k}^1 - y_{j,l}^1} \right),$$

which is a polynomial of degree $m(1) - 1$ in the direction j and constant in all other directions. Finally, the level 1 sparse grid interpolant is defined as

$$S_1 f := S_0 f + \sum_{j=1}^L (\mathcal{U}_j^1 f - S_0 f) = (1 - L) S_0 f + \sum_{j=1}^L \mathcal{U}_j^1 f$$

which is nothing else than the sum of the level 0 sparse grid interpolant and the details in each direction.

Remark 1.6.1. *It can be proved that the SC approximation computed with a sparse grid of level 1 indeed yields an error of order ε^2 , using for instance a scaling argument together with the fact that S_1 is exact for any polynomial of (total) degree at most 1 (see [18]). More generally, we can show that a sparse grid of level l yields an error of order ε^{l+1} for the choice $m(i) = i$, while for the second choice of m it is of order ε^{l+k+1} , where $k = 0$ if $l < L$ and $k = l - L + 1$ otherwise.*

The type of points in each direction is chosen according to the distribution of the random variables. Note that the use of Clenshaw-Curtis points, which are the extrema of Chebyshev polynomials and which are suitable for uniformly distributed random variables, and Smolyak sparse grid leads to nested set of abscissas. However, since only sparse grids of level 1 are considered, there is no real advantage to consider hierarchical sparse grids. In both cases $m(1) = 2$ and Gauss-Legendre abscissas and $m(1) = 3$ and Clenshaw-Curtis abscissas, referred to as SC1 and SC2 in the following, the sparse grid of level 1 consists of $2L + 1$ collocation points (due to the use of nested set of abscissas in each direction for SC2).

Let W_l , respectively W_{nl} , denote the work to solve once a given linear, respectively nonlinear, problem. Moreover, let $W_{\tilde{l}}$ denote the work to solve the linear problem for U_j associated with the nonlinear one, see (1.77). Table 1.1 contains the computational costs for the SC-FEM and the *perturbation method*. Notice that the work to construct the sparse grid is not taken into account.

	linear problem	nonlinear problem
SC-FEM	$(2L + 1) \cdot W_l$	$(2L + 1) \cdot W_{nl}$
<i>perturbation method</i>	$(L + 1) \cdot W_l$	$W_{nl} + L \cdot W_{\tilde{l}}$

Table 1.1: Computational costs for the SC-FEM and the *perturbation method*.

The *perturbation method* presents no real advantage for solving linear problems since the costs for both methods differ only by a factor 2. The situation is different when a nonlinear problem is considered. Indeed, when using the SC method, we need to solve as many nonlinear problems as collocation points, i.e. $2L + 1$ problems, whereas only one nonlinear problem needs to be solved for the *perturbation method*. The L remaining problems, to compute the U_j , $j = 1, \dots, L$, are linear and so usually much cheaper to solve. However, one should invest extra effort to derive by hand the Fréchet derivatives and implement the problems solved by the U_j , $j = 1, \dots, L$.

1.7 Numerical results

This section is devoted to illustration and validation of the theoretical results obtained in the previous sections. We start with the analysis of 1D problems, analysing first the convergence rate for various errors and norms and presenting, next, algorithms which adaptively refine the (physical) mesh to balance the two sources of error: the physical space discretization and the uncertainty. We present then two 2D examples and conclude this section with a comparison with the stochastic collocation method in term of computational costs when solving linear and nonlinear problems.

1.7.1 1D problems

Let $D = (0, 1)$. In what follows, the true errors in the $L_P^2(\Omega; H_0^1(D))$ and $L_P^2(\Omega; L^2(D))$ norms have been accurately approximated with the standard Monte Carlo method, with a sample of size $K = 10000$, i.e. for $V = H_0^1(D)$ or $L^2(D)$ we approximate

$$\|v\|_{L_P^2(\Omega; V)} \approx \left(\frac{1}{K} \sum_{k=1}^K \|v(\cdot, \mathbf{y}_k)\|_V^2 \right)^{\frac{1}{2}} \quad \forall v \in L_P^2(\Omega; V),$$

where $\{\mathbf{y}_k\} \in \Gamma$ are i.i.d realizations of the random vector \mathbf{Y} . With this choice for the sample size, the variance of the estimation of the error for all the considered values of h and ε is at most 10^{-5} the estimated error. In what follows, whenever we refer to *error* it should be understood that the true error has been accurately computed by the Monte Carlo procedure. Since the exact random solution of the problems considered below is not known, the error is computed with respect to a reference solution computed on a fine uniform mesh for D , namely with a mesh-grid of length $h_{ref} = 2^{-12}$. Notice that if we take a FE space of mesh size $h = h_{ref}$, then only the statistical error is considered. Finally, all the involved integrals are evaluated numerically with sufficiently accurate quadrature formulas that permit to neglect the effect of quadrature.

Let us first consider $L = 50$ random variables Y_j , $j = 1, \dots, L$, which can take the values ± 1 with probability $\frac{1}{2}$. Such discrete random variables have zero mean, unit variance and unit fourth moment. Similarly to what is done in [124], we take a diffusion coefficient of the form

$$a(x, \mathbf{Y}(\omega)) = 1 + \varepsilon \sum_{j=1}^L \frac{\cos(2\pi j x)}{(\pi j)^2} Y_j(\omega), \quad (1.86)$$

which is similar to a (truncated) Karhunen-Loève expansion with eigenvalues of order $\frac{1}{j^4}$. With this choice of random diffusion coefficient, we have $1 - \frac{\varepsilon}{6} \leq a(x, \mathbf{y}) \leq 1 + \frac{\varepsilon}{6}$. We take $\varepsilon \in [0, 4]$ which guarantee property (A1) with $a_{min} = \frac{1}{3}$ and $a_{max} = \frac{5}{3}$. Finally, we consider two different right-hand sides, namely

$$f_1(x) = 1 \quad \text{and} \quad f_2(x) = 72(1 - 72(x - 0.5)^2) e^{-36(x-0.5)^2}. \quad (1.87)$$

The latter corresponds to the exact solution $u_0(x) = e^{-36(x-0.5)^2} - e^{-9}$ for problem (1.20) while it is $u_0(x) = 0.5x(1-x)$ for the case $f = f_1$.

Error in $L_P^2(\Omega; H_0^1(D))$ -norm

We consider first the error measured in $L_P^2(\Omega; H_0^1(D))$ -norm. We show in Figure 1.1 the convergence rate of the error $u - u_{0,h}$ with respect to $2^{-9} \leq h \leq 2^{-3}$ for $\varepsilon = 32h$, along with the *a posteriori* estimator based on (1.43) and (1.44). Based on this result, we can see that a division of h and ε by two halves the error, which is in agreement with the convergence of $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$ in $\mathcal{O}(h + \varepsilon)$ predicted by the foregoing error analysis. Moreover, for the two cases f_1 and f_2 , the gap between the error and the estimator is of about 1.6 and 2.8, respectively, which is comprised between the effectivity index of the stochastic error estimator (1) and the spatial error estimator (3.46), see below for details. Concerning the convergence rate of the second order approximation, we present in Figure 1.2 the error between u and u_h^1 with respect to $2^{-3} \leq \varepsilon \leq 2$ for $h = \varepsilon^2/32$. This result confirms the convergence in $\mathcal{O}(\varepsilon^2)$ of the stochastic truncation predicted by (1.63), when the exact solution is approximated by $u_0 + \varepsilon u_1$.

The error estimators depicted in Figures 1.1 and 1.2 do not take into account the unknown

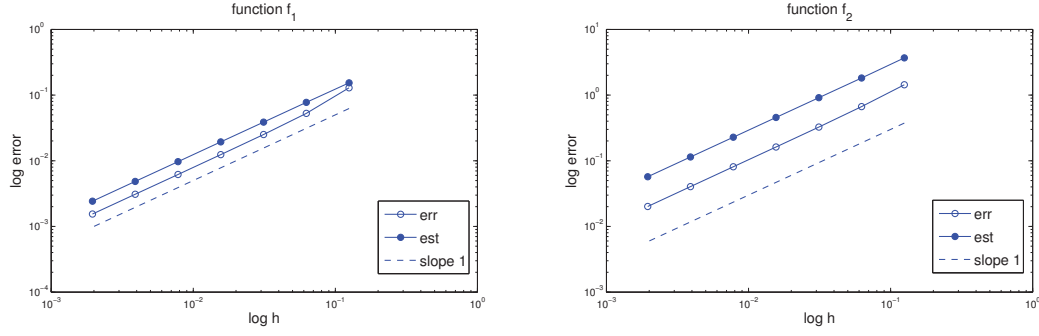


Figure 1.1: Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L^2_P(\Omega; H_0^1(D))$ -norm w.r.t h with $\varepsilon = 32h$.

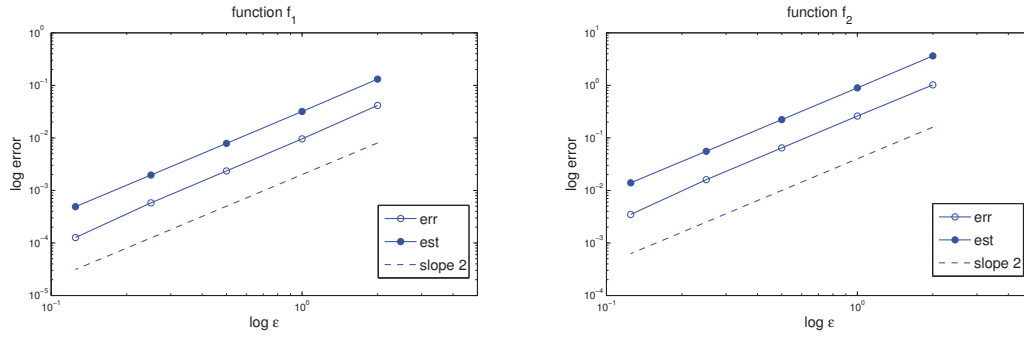


Figure 1.2: Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and u_h^1 in $L^2_P(\Omega; H_0^1(D))$ -norm w.r.t ε with $h = \varepsilon^2/32$.

constants due to interpolation error that appear in (1.42) and (1.63). These constants can be estimated numerically as follows to obtain a sharp error estimator: consider the problem $-u_0'' = f$ with f such that the exact solution is known, for instance $f = f_1$ or $f = f_2$, and define $1/C_{H_0^1} := 3.46 \approx \eta_1 / \|u_0 - u_{0,h}\|_{H_0^1(D)}$ for h small enough. This estimation can be done once for all since $C_{H_0^1}$ does not depend on the input data. We define then $\tilde{\eta} := \left(C_{H_0^1}^2 \eta_1^2 + \eta_2^2\right)^{\frac{1}{2}}$ as an estimator for the error $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$. We will say that $\tilde{\eta}$ is a good approximation of the error if the ratio $\tilde{\eta} / \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$ remains between a_{min} and a_{max} . Since in the considered case the ratio a_{max}/a_{min} tends to 1 as ε goes to 0, we expect the effectivity index of the estimator $\tilde{\eta}$ to approach 1 as ε gets smaller. We give in Tables 1.2 and 1.3 the results obtained when the constant $C_{H_0^1}$ is considered. In Table 1.2, the mesh size is fixed to $h = 2^{-7}$ while in Table 1.3 we fix $\varepsilon = 0.25$. In both cases, the ratio of the estimator $\tilde{\eta}$, which contains the estimated constant $C_{H_0^1}$, over the error is close to one.

	ε	<i>error</i>	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
$f = f_1$	4	1.2167e-1	2.2579e-3	9.1996e-2	9.2024e-2	0.75632
	2	4.9276e-2	2.2579e-3	4.5998e-2	4.6054e-2	0.93461
	1	2.3460e-2	2.2579e-3	2.2999e-2	2.3110e-2	0.98505
	0.5	1.1760e-2	2.2579e-3	1.1500e-2	1.1719e-2	0.99652
	0.25	6.1805e-3	2.2579e-3	5.7498e-3	6.1772e-3	0.99947
	0.125	3.6545e-3	2.2579e-3	2.8749e-3	3.6556e-3	1.00031

	ε	<i>error</i>	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
$f = f_2$	4	9.5591e-1	6.4347e-2	7.8646e-1	7.8909e-1	0.82548
	2	4.1806e-1	6.4347e-2	3.9323e-1	3.9846e-1	0.95312
	1	2.0916e-1	6.4347e-2	1.9661e-1	2.0688e-1	0.98910
	0.5	1.1782e-1	6.4347e-2	9.8307e-2	1.1749e-1	0.99720
	0.25	8.0974e-2	6.4347e-2	4.9154e-2	8.0973e-2	0.99999
	0.125	6.8769e-2	6.4347e-2	2.4577e-2	6.8881e-2	1.00163

Table 1.2: Error $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta} / \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$ for $h = 2^{-7}$ and various ε for both cases f_1 and f_2 .

The same observation holds for the approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$ taking $C_1 = C_2 = C_{H_0^1}^2$ in (1.63) and for the generalization (1.74) with $C_i = C_{H_0^1}^2$ for $i = 0, \dots, N$, see Table 1.4 where the case $u \approx u_{0,h} + \varepsilon u_{1,h}$ is presented for the case $f = f_2$. Recall that η_1 , η_2 and η_3 are given in (1.64), (1.65) and (1.66), respectively, and here $\tilde{\eta} := \left(C_{H_0^1}^2 \eta_1^2 + C_{H_0^1}^2 \eta_2^2 + \eta_3^2\right)^{\frac{1}{2}}$.

Error in $L_P^2(\Omega; L^2(D))$ -norm

We consider now the error $u - u_{0,h}$ in $L_P^2(\Omega; L^2(D))$ -norm. According to the theoretical result, we should get a convergence of order h^2 for $\varepsilon = Ch^2$. Figure 1.3, which contains the plot of the error and estimator based on (1.54) and (1.55) for $C = 32$ and $2^{-6} \leq h \leq 2^{-2}$, confirms that this

	N	$error$	$C_{H_0^1}\eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
$f = f_1$	8	3.6575e-2	3.6127e-2	5.5801e-3	3.6556e-2	0.99946
	16	1.8951e-2	1.8064e-2	5.7030e-3	1.8942e-2	0.99955
	32	1.0712e-2	9.0318e-3	5.7384e-3	1.0701e-2	0.99890
	64	7.3076e-3	4.5159e-3	5.7475e-3	7.3094e-3	1.00024
	128	6.1765e-3	2.2580e-3	5.7498e-3	6.1772e-3	1.00011
	256	5.8822e-3	1.1290e-3	5.7503e-3	5.8601e-3	0.99625

	N	$error$	$C_{H_0^1}\eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
$f = f_2$	8	9.7697e-1	1.0441e-0	4.6189e-2	1.0451e-0	1.06977
	16	5.1089e-1	5.1478e-1	4.8261e-2	5.1704e-1	1.01204
	32	2.6109e-1	2.5739e-1	4.8942e-2	2.6200e-1	1.00349
	64	1.3766e-1	1.2869e-1	4.9112e-2	1.3775e-1	1.00066
	128	8.0919e-2	6.4347e-2	4.9154e-2	8.0973e-2	1.00066
	256	5.8787e-2	3.2174e-2	4.9164e-2	5.8756e-2	0.99946

Table 1.3: Error $\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$ for $\varepsilon = 0.25$ and various $h = 1/N$ for both cases f_1 and f_2 .

	ε	$error$	$C_{H_0^1}\eta_1$	$C_{H_0^1}\eta_2$	η_3	$\tilde{\eta}$	$\tilde{\eta}/error$
$h = 2^{-10}$	4	3.5236e-1	8.0434e-3	1.8607e-3	2.7488e-1	2.7500e-1	0.78044
	2	7.3380e-2	8.0434e-3	9.3037e-4	6.8719e-2	6.9194e-2	0.94295
	1	1.9054e-2	8.0434e-3	4.6519e-4	1.7180e-2	1.8975e-2	0.99586
	0.5	8.9126e-3	8.0434e-3	2.3259e-4	4.2949e-3	9.1213e-3	1.02341
	0.25	7.8616e-3	8.0434e-3	1.1630e-4	1.0737e-3	8.1156e-3	1.03231
	0.125	7.7840e-3	8.0434e-3	5.8148e-5	2.6843e-4	8.0481e-3	1.03393

	N	$error$	$C_{H_0^1}\eta_1$	$C_{H_0^1}\eta_2$	η_3	$\tilde{\eta}$	$\tilde{\eta}/error$
$\varepsilon = 1$	32	2.6010e-1	2.5739e-1	1.4792e-2	1.6950e-2	2.5837e-1	0.99337
	64	1.3107e-1	1.2869e-1	7.4313e-3	1.7121e-2	1.3004e-1	0.99219
	128	6.7384e-2	6.4347e-2	3.7201e-3	1.7165e-2	6.6701e-2	0.98987
	256	3.6796e-2	3.2174e-2	1.8606e-3	1.7176e-2	3.6519e-2	0.99246
	512	2.3761e-2	1.6087e-2	9.3036e-4	1.7179e-2	2.3554e-2	0.99128
	1024	1.9131e-2	8.0434e-3	4.6519e-4	1.7180e-2	1.8975e-2	0.99188

Table 1.4: Error and estimators for the approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$ with h fixed (top) and ε fixed (bottom) for the case $f = f_2$.

is the case. Similarly to the error in $H_0^1(D)$ -norm, the constant due to interpolation error could

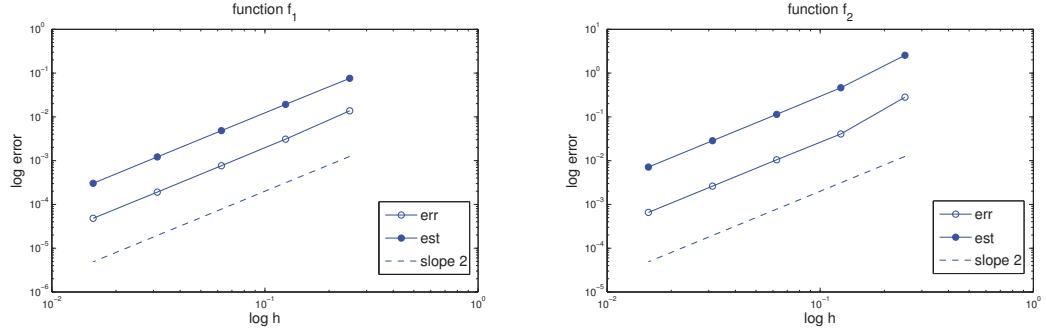


Figure 1.3: Convergence orders for problem (1.11) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L_P^2(\Omega; L^2(D))$ -norm w.r.t h with ε fixed to $32h^2$.

be estimated numerically once for all following the same procedure as above. However, even with a sharp estimation of such constant, there is no guarantee that the estimator is efficient though it has the correct convergence rate. We see two reasons for that. First of all, there are no proofs, to our knowledge, that the part of the estimator due to the uncertainty (η_2) is a lower bound for the error in $L^2(D)$ -norm, mainly due to the use of the Poincaré inequality. Considering $h = h_{ref}$, the estimator over estimates the error by a factor of about 4.2 for $f = f_1$ and 9 for $f = f_2$, showing that the constant multiplying η_2 does depend on f . Moreover, the constant C_1 in (1.53) depends in an implicit way on the uniform bound for a and ∇a (see Remark 1.3.10).

Different setup

Similar results are obtained when other input data are considered. For instance, let us consider independent uniformly distributed random variables in $[-\sqrt{3}, \sqrt{3}]$. In this case, the random variables still have zero mean and unit variance but $\mathbb{E}[Y_j^4] = \frac{9}{5}$. This only modifies the part η_3 in the *a posteriori* error estimate (1.63) for $\|u - u_h^1\|_{L_P^2(\Omega; H_0^1(D))}$. Moreover, we also modify the functions a_j considering here

$$a(x, \mathbf{Y}(\omega)) = 1 + \varepsilon \sum_{j=1}^{50} \frac{\cos(8\pi j x) \sin(2\pi j x)}{(\pi j)^2} Y_j(\omega) \quad (1.88)$$

for the random diffusion coefficient. Notice that this choice satisfies $1 - \frac{\sqrt{3}\varepsilon}{6} \leq a(x, \mathbf{y}) \leq 1 + \frac{\sqrt{3}\varepsilon}{6}$. We give in Figure 1.4 some realizations of a and the corresponding solution for the case $\varepsilon = 1$ and $f = f_2$ defined in (1.87).

The results obtained when the constant $C_{H_0^1} = 1/3.46$ is taken into account are given in Table 1.5. First, the mesh size is fixed to $h = 1/N = 2^{-8}$ and ε varies and then, we set $\varepsilon = 0.5$ and consider various partitions of $[0, 1]$. When h is fixed, the error decreases linearly with respect

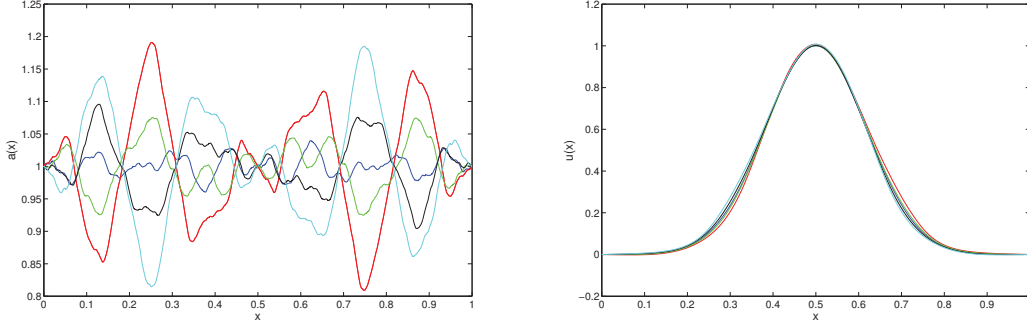


Figure 1.4: Five realizations of the random diffusion coefficient a given in (1.88) with $\varepsilon = 1$ (left) and the corresponding solution for $f = f_2$ (right).

to ε until the FE error is no longer negligible. The same observation holds when ε is fixed and h varies. In both cases, the effectivity index of the error estimator $\tilde{\eta} = (C_{H_0^1}^2 \eta_1^2 + \eta_2^2)^{\frac{1}{2}}$ is close to one.

ε	$error$	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
2	3.2152e-1	3.2174e-2	3.0331e-1	3.0501e-1	0.94866
1	1.5541e-1	3.2174e-2	1.5165e-1	1.5503e-1	0.99754
0.5	8.1168e-2	3.2174e-2	7.5827e-2	8.2371e-2	1.01482
0.25	4.9399e-2	3.2174e-2	3.7914e-2	4.9725e-2	1.00659
0.125	3.7192e-2	3.2174e-2	1.8957e-2	3.7343e-2	1.00406
0.0625	3.3432e-2	3.2174e-2	9.4784e-3	3.3541e-2	1.00325

N	$error$	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
8	9.7920e-1	1.0441e-0	9.7528e-2	1.0487e-0	1.07093
16	5.1403e-1	5.1478e-1	7.6937e-2	5.2050e-1	1.01258
32	2.6726e-1	2.5739e-1	7.5506e-2	2.6824e-1	1.00365
64	1.4900e-1	1.2869e-1	7.5726e-2	1.4932e-1	1.00217
128	9.8399e-2	6.4347e-2	7.5805e-2	9.9433e-2	1.01051
256	8.1817e-2	3.2174e-2	7.5827e-2	8.2371e-2	1.00676

Table 1.5: Error $\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$ for $h = 2^{-8}$ (top) and $\varepsilon = 0.5$ (bottom).

Adaptive algorithm

We propose here adaptive algorithms to determine, for a given ε , a mesh for D that balances the two sources of error. The convergence rate of the error in the $L_p^2(\Omega; H_0^1(D))$ norm with respect to h for uniform refinements and for the first, second and third order approximation for several given (fixed) values of ε is depicted in Figure 1.5 in the case $f = 1$ and a given in

(1.86). First, we can notice that a better accuracy is reached when u is approximated by u_h^2 than

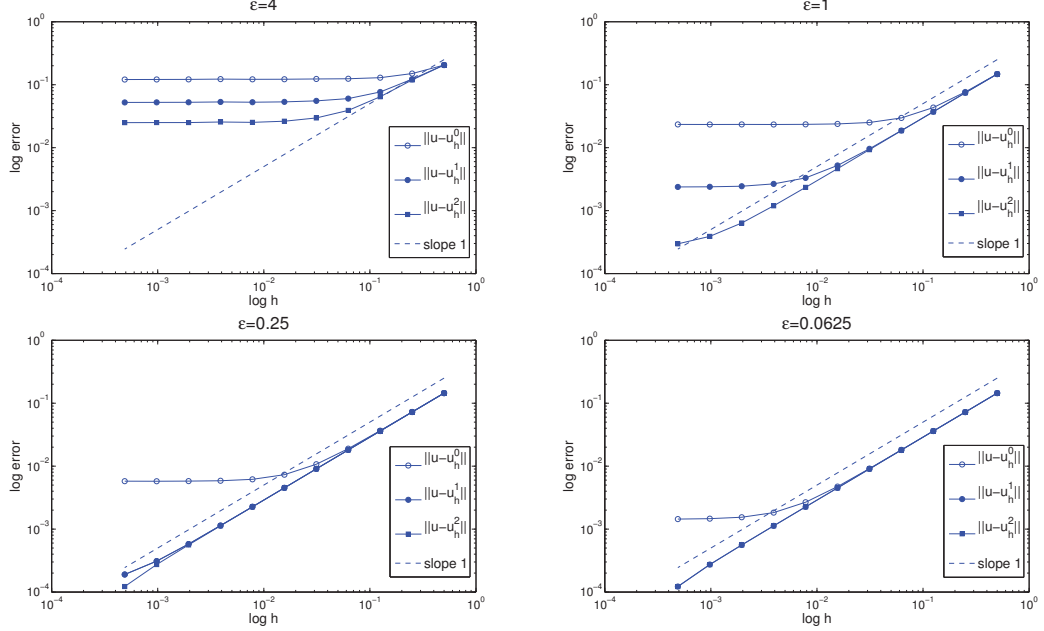


Figure 1.5: Convergence rate for problem (1.11) with $f = f_1$ for $\varepsilon = 4, 1, 0.25, 0.0625$. Log log scale plot of the error in $L^2_P(\Omega; H^1_0(D))$ -norm w.r.t h .

with u_h^1 , which in turn provides a better approximation than only the deterministic part $u_{0,h}$. This observation holds except for coarse meshes where the FE error is dominating yielding comparable accuracy in all cases. Moreover, the global approximation error remains constant for mesh sizes smaller than a critical value h_0 of the mesh-size. Any further refinement of the mesh below this value should thus be avoided since it would not improve the global approximation error, being dominated by the stochastic error.

Based on this observation, it is interesting to determine how fine the mesh should be to get a comparable error in h and ε . More precisely, for a given ε and for the approximation $u \approx u_{0,h}$, we would like to find a mesh for D such that

$$\frac{T-1}{T}\eta_2 \leq \eta_1 \leq \frac{T+1}{T}\eta_2 \quad (1.89)$$

for a given preset tolerance $T > 1$, where η_1 and η_2 are given by (1.43) and (1.44), respectively. Notice that in all what follows, η_1 can be replaced by $C_{H^1_0}\eta_1$ if the estimated constant $C_{H^1_0}$ is at disposal, so that the *correct* balance of the two sources of error is considered. Moreover, we mention that the choice of the law of the Y_j , $j = 1, \dots, L$, is irrelevant here as long as $\mathbb{E}[Y_j] = 0$ and $\text{Var}(Y_j) = 1$. Indeed, the error estimator η_2 given in (1.44) is valid under these conditions irrespectively of the law of Y_j and only the solution $u_{0,h}$ is computed.

Uniform refinement

The adaptation can be done in 1D using Algorithm 1 given below, where $N_h + 1$ denotes the number of discretization points in $[0, 1]$.

Algorithm 1 find $h = N_h^{-1}$ such that (1.89) holds

Require: N_{init} and T

Ensure: mesh-size h which yield comparable accuracy in h and ε

```

1:  $N_h = N_{init}$ 
2: Compute  $u_{0,h}$  on the uniform partition  $x_i = ih$ ,  $h = N_h^{-1}$ ,  $i = 0, 1, \dots, N_h$ 
3: Compute  $\eta_1$  and  $\eta_2$  according to (1.43) and (1.44)
4: if  $\frac{T-1}{T} \leq \frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$  then
5:   stop
6: else
7:   if  $\frac{\eta_1}{\eta_2} < \frac{T-1}{T}$  then
8:      $N_h \leftarrow \lfloor \frac{N_h}{2} \rfloor$  (mesh too fine)
9:   else
10:     $N_h \leftarrow 2N_h$  (mesh too coarse)
11:   end if
12:   go to 2.
13: end if
```

Applying Algorithm 1 to our problem for $T = 2$ and various given ε , we get the results presented in Table 1.6.

ε	f_1			f_2		
	N_h	η_1	η_2	N_h	η_1	η_2
1	32	0.03125	0.02295	128	0.22264	0.19661
0.5	64	0.01563	0.01149	256	0.11132	0.09833
0.25	128	0.00781	0.00575	512	0.05566	0.04917
0.125	256	0.00391	0.00288	1024	0.02783	0.02458
0.0625	512	0.00195	0.00144	2048	0.01392	0.01229

Table 1.6: Value of $h = N_h^{-1}$ with respect to ε such that (1.89) holds with $T = 2$.

We mention that if T is large, i.e. $\frac{T-1}{T}$ is close to $\frac{T+1}{T}$, the algorithm might not converge due to an oscillation of the ratio $\frac{\eta_1}{\eta_2}$ below the lower bound $\frac{T-1}{T}$ and above the upper bound $\frac{T+1}{T}$ in two consecutive steps. Such behaviour will be observed if no uniform partition of D satisfies (1.89). Moreover, notice that with Algorithm 1, only refinement or only coarsening is performed, depending on the initial number N_{init} of subintervals.

Non-uniform refinement

Algorithm 1 given above only uses uniform refinement or coarsening. Of course, adaptive refinements can be considered as well exploiting the local nature of the estimator η_1 , which can indeed be written as

$$\eta_1^2 = \sum_{K \in \mathcal{T}_h} \eta_K^2 \quad \text{with} \quad \eta_K^2 = h_K^2 \int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \frac{1}{2} \sum_{e \subset \partial K} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}^2 \quad (1.90)$$

taking into account that each edge is then counted twice.

Remark 1.7.1. *The factor $\frac{1}{2}$ could in fact be replaced by $\frac{1}{4}$ if we do not split the summation over the elements and the edges in the derivation of the error estimate in (1.47), namely if we consider an element point of view. Indeed, we can use the fact that for any $v \in H_0^1(D)$ and any $v_h \in V_h$ we have*

$$\int_D f v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v = \sum_{K \in \mathcal{T}_h} \left[\int_K (f + \nabla \cdot (a_0 \nabla u_{0,h}))(v - v_h) + \int_{\partial K} \frac{1}{2} [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} (v - v_h) \right].$$

Recall that in 1D, for a partition $0 = x_0 < x_1 < \dots < x_{N_h} = 1$, the error estimator η_1 reads

$$\eta_1^2 = \sum_{i=0}^{N_h-1} \eta_{1,i}^2 \quad \text{with} \quad \eta_{1,i}^2 = h_i^2 \|f + (a_0 u'_{0,h})'\|_{L^2(x_i, x_{i+1})}^2.$$

The goal being to satisfy (1.89), a first possibility is to require that

$$B_{\inf} := \left(\frac{T-1}{T} \right)^2 \eta_2^2 \frac{1}{N_h} \leq \eta_{1,i}^2 \leq \left(\frac{T+1}{T} \right)^2 \eta_2^2 \frac{1}{N_h} =: B_{\sup} \quad \forall i = 0, \dots, N_h - 1. \quad (1.91)$$

Another sufficient condition for (1.89) to hold is to impose that

$$B_{\inf} := \left(\frac{T-1}{T} \right)^2 \eta_2^2 \frac{h_i}{|D|} \leq \eta_{1,i}^2 \leq \left(\frac{T+1}{T} \right)^2 \eta_2^2 \frac{h_i}{|D|} =: B_{\sup} \quad \forall i = 0, \dots, N_h - 1 \quad (1.92)$$

using the fact that $\sum_{i=0}^{N_h-1} h_i = |D|$. The criterion (1.91) imposes an equidistribution of the error, enforcing a comparable value of the local error estimator on each subinterval regardless of its length. In the second strategy (1.92), the repartition of the error is weighted by h_i . This is commonly used in a time-adaptivity framework so that the solution does not need to be computed until the final time before adapting the time step.

We give in Algorithm 2 an adaptive procedure which find a (non-uniform) partition of D for which (1.89) holds. The idea is to check for each subinterval $[x_i, x_{i+1}]$, $i = 0, \dots, N_h - 1$, of the current partition of D if the local error estimator $\eta_{1,i}$ satisfies the criterion (1.91) or (1.92). If it is too large, then we should refine the interval $[x_i, x_{i+1}]$, for instance by adding its midpoint, while a coarsening should be done if it is too small.

To better appreciate the behaviour of the non-uniform adaptation, we test Algorithm 2 with a

Algorithm 2 adaptive algorithm with non-uniform partition

Require: T and initial partition $\mathcal{T}_h = \{x_i, i = 0, \dots, N_h - 1\}$

Ensure: partition of D such that (1.89) holds

```

1: Compute  $u_{0,h}$  on  $\mathcal{T}_h$ 
2: Compute  $\eta_1$  and  $\eta_2$  according to (1.43) and (1.44)
3: if  $\frac{T-1}{T} \leq \frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$  then
4:   stop
5: else
6:   for  $i = 0, \dots, N_h - 1$  do
7:     if  $\eta_{1,i}^2 > B_{\sup}$  then
8:       add the midpoint  $\frac{x_i + x_{i+1}}{2}$  to  $\mathcal{T}_h$ 
9:     else if  $\eta_{1,i}^2 < B_{\inf}$  then
10:      remove the endpoint  $x_{i+1}$  from  $\mathcal{T}_h$  ( $x_i$  if  $i = N_h - 1$ )
11:    end if
12:  end for
13: end if
14: go to 1.

```

different forcing term than in the previous sections, keeping the diffusion coefficient a as in (1.86) and all other input data being unchanged. We consider the source term f for which the corresponding solution u_0 of problem (1.20) is given by¹

$$u_0(x) = x - \frac{1 - e^{x\tau^{-1}}}{1 - e^{\tau^{-1}}}. \quad (1.93)$$

The solution presents a boundary layer near $x = 1$ of width proportional to τ , see Figure 1.6. It is linear on the remaining part of the interval, where only few points are thus sufficient to obtain a good approximation. In the numerical results below, we choose $\tau = 0.05$.

We give in Tables 1.7 and 1.8 the results obtained for various values of ε when using the two adaptive criterion (1.91) and (1.92), respectively. We have denoted by N_h the number of subintervals of D (i.e. $N_h + 1$ is the number of nodes), $h_{\min} = \min_i h_i$ and $h_{\max} = \max_i h_i$ are the minimum and maximum mesh sizes, respectively, and iter stands for the number of iterations of the adaptive algorithm. In all cases, we have started the adaptation with the initial partition $\{0, 0.5, 1\}$.

First, we can see that the number of iterations is similar in both cases and the same holds for the values of the error estimators η_1 and η_2 . Moreover, the number of nodes is smaller when criterion (1.91) is used while the maximum subinterval length h_{\max} is in general larger with (1.92). The latter strategy indeed allows to have large subintervals if the corresponding local error estimator is small. This can be seen in Figure 1.6 where the repartition of the nodes is given for various values of ε and for both criteria (1.91) and (1.91). The continuous line

¹The function u_0 in (1.93) is the solution of the problem $-\tau u_0'' + u_0' = 1$ in $(0, 1)$ with homogeneous Dirichlet boundary conditions.

ε	N_h	h_{\min}	h_{\max}	η_1	η_2	iter
1	28	3.91e-3	6.56e-1	3.6424e-1	3.0474e-1	8
0.5	53	1.95e-3	6.25e-1	1.9848e-1	1.5272e-1	9
0.1	231	4.88e-4	4.69e-1	3.8504e-2	3.0727e-2	11
0.05	461	2.44e-4	2.50e-1	1.9233e-2	1.5164e-2	12
0.01	2056	6.10e-5	1.88e-1	4.4334e-3	3.0339e-3	14
0.005	4119	3.05e-5	2.81e-1	2.2138e-3	1.5178e-3	15
0.001	25646	3.81e-6	1.05e-1	3.3770e-4	3.0304e-4	18
0.0005	51292	1.91e-6	1.05e-1	1.6884e-4	1.5150e-4	19
0.0001	233216	4.77e-7	5.27e-2	3.7686e-5	3.0301e-5	21

Table 1.7: Adaptive partition of D such that (1.89) holds with $T = 2$ when criterion (1.91) is used.

ε	N_h	h_{\min}	h_{\max}	η_1	η_2	iter
1	64	9.77e-04	5.63e-01	2.4702e-1	3.0720e-1	10
0.5	70	1.95e-03	6.25e-01	1.7704e-1	1.5273e-1	9
0.1	293	4.88e-04	4.69e-01	3.6756e-2	3.0704e-2	11
0.05	581	2.44e-04	5.47e-01	2.2340e-2	1.5356e-2	12
0.01	3880	3.05e-05	2.50e-01	3.2449e-3	3.0329e-3	15
0.005	7741	1.53e-05	4.38e-01	1.7937e-3	1.5338e-3	16
0.001	33949	3.81e-06	3.75e-01	4.0887e-4	3.0531e-4	18
0.0005	99606	9.54e-07	1.88e-01	1.6707e-4	1.5170e-4	20
0.0001	295692	4.77e-07	2.50e-01	4.0904e-5	3.0320e-5	21

Table 1.8: Adaptive partition of D such that (1.89) holds with $T = 2$ when criterion (1.92) is used.

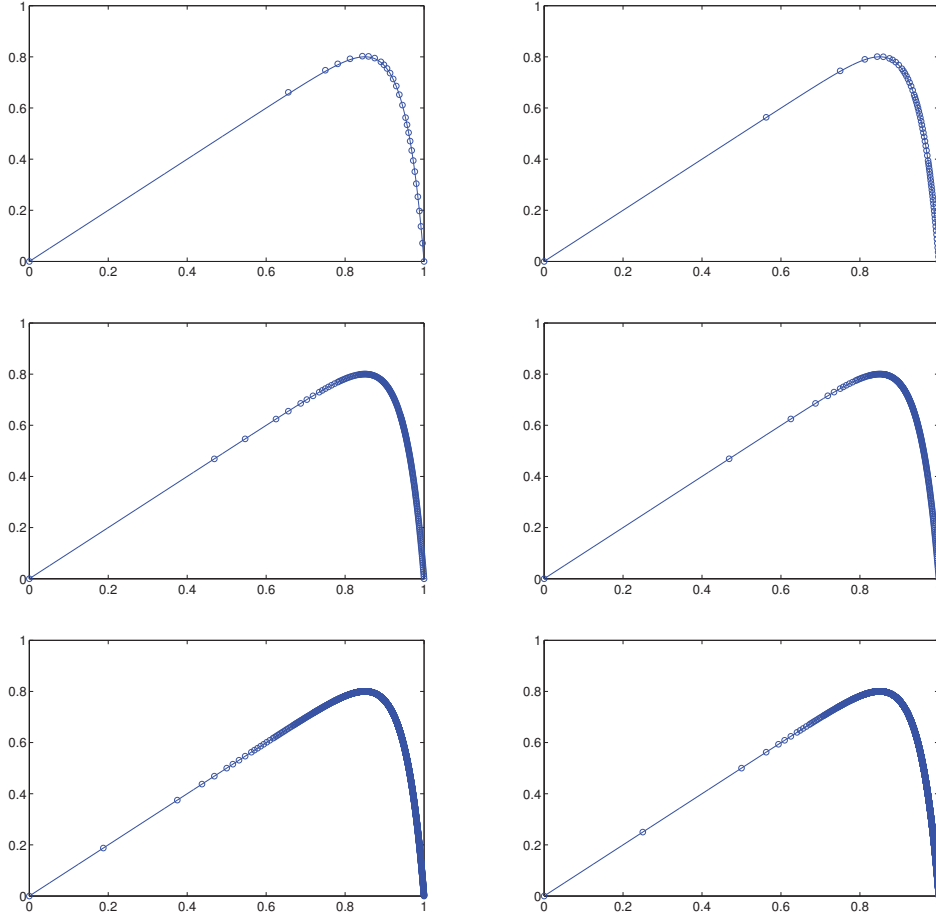


Figure 1.6: Repartition of the nodes for $\varepsilon = 1$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.01$ (bottom) in the case $T = 2$. Left: strategy (1.91), right: strategy (1.92).

represents the exact solution u_0 given in (1.93).

As we have seen in Tables 1.7 and 1.8, the two methods yield comparable results. The number of nodes for criterion (1.92) is larger but it allows, in general, larger maximum mesh size h_{\max} .

Finally, we compare the results of Tables 1.7 and 1.8 with those obtained using a Dörfler [57] bulk-chasing marking commonly used in adaptive finite element method (AFEM), see for instance [42, 114]. To reach the target $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$, a suitable fraction of the subintervals with highest local error estimator is selected for refinement at each iteration. More precisely, for a given parameter $\theta \in (0, 1]$, we select an index set $J \subseteq \{0, 1, \dots, N_h - 1\}$ of minimal cardinality such that

$$\left(\sum_{j \in J} \eta_{1,j}^2 \right)^{\frac{1}{2}} \geq \theta \left(\sum_{i=0}^{N_h-1} \eta_{1,i}^2 \right)^{\frac{1}{2}} = \theta \eta_1.$$

This marking strategy is often referred to as *equilibration strategy* and yields comparable

results than the so-called *maximum strategy*, see [119]. Notice that if θ is closed to 0, then only few subintervals will be refined at each iteration while choosing θ close to 1 will generate a set J of large cardinality. In particular, the case $\theta = 1$ gives similar results than Algorithm 1 without coarsening, namely all the subintervals are refined at each iteration, except those for which² $\eta_{1,i} = 0$. The procedure based on Dörfler marking is described in Algorithm 3. The search for the index $i \in \{0, \dots, N_h - 1\} \setminus J$ with largest $\eta_{1,i}^2$ (see line 8) can be achieved by sorting the local estimators $\eta_{1,i}$ in decreasing order before the *while* loop.

Algorithm 3 adaptive algorithm with Dörfler marking

Require: T, θ and initial partition $\mathcal{T}_h = \{x_i, i = 0, \dots, N_h - 1\}$

Ensure: partition of D such that $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$

- 1: Compute $u_{0,h}$ on \mathcal{T}_h
 - 2: Compute η_1 and η_2 according to (1.43) and (1.44)
 - 3: **if** $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ **then**
 - 4: stop
 - 5: **else**
 - 6: $J = \emptyset$ and $\vartheta = 0$
 - 7: **while** $\vartheta < \theta\eta_1$ **do**
 - 8: $J \leftarrow J \cup \{j\}$ with $j = \arg \max_{i \in \{0, \dots, N_h - 1\} \setminus J} \eta_{1,i}^2$
 - 9: $\vartheta \leftarrow \vartheta + \eta_{1,j}^2$
 - 10: add the midpoint $\frac{x_j + x_{j+1}}{2}$ to \mathcal{T}_h
 - 11: **end while**
 - 12: **end if**
 - 13: go to 1.
-

We give in Table 1.9 the results obtained using the Dörfler strategy of Algorithm 3 for the same values of ε than in Tables 1.7 and 1.8.

ε	N_h	h_{\min}	h_{\max}	η_1	η_2	iter
1	23	3.91e-03	5.00e-01	4.3240e-1	3.0736e-1	16
0.5	41	1.95e-03	5.00e-01	2.2862e-1	1.5371e-1	21
0.1	201	4.88e-04	5.00e-01	4.3679e-2	3.0756e-2	36
0.05	419	2.44e-04	2.50e-01	2.0873e-2	1.5164e-2	43
0.01	2017	3.05e-05	2.50e-01	4.3602e-3	3.0330e-3	58
0.005	4177	1.53e-05	2.50e-01	2.1044e-3	1.5165e-3	65
0.001	19715	3.81e-06	1.25e-01	4.4296e-4	3.0306e-4	80
0.0005	40705	1.91e-06	1.25e-01	2.1412e-4	1.5147e-4	87
0.0001	191790	4.77e-07	6.25e-02	4.5111e-5	3.0300e-5	102

Table 1.9: Dörfler strategy such that $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ holds with $T = 2$ and $\theta = 0.5$.

Compared to the results obtained with the two previous adaptive strategies, the Dörfler

²From a numerical point of view, any element which does not contribute to the sum for η_1 will not be refined, i.e. any element which is *numerically zero* due to machine precision.

Chapter 1. Elliptic model problems with random diffusion coefficient

marking procedure requires more iterations but produces a partition of D satisfying $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ with fewer nodes. Moreover, this last inequality is tight here which is an expected feature for moderate θ , or when few local error estimators are large compared to the others, since only few subintervals are refined at each step. It is therefore more likely to stop the refinement process when the tolerance is just satisfied. We give in Table 1.10 the results obtained when changing the value of θ .

θ	N_h	h_{\min}	h_{\max}	η_1	η_2	iter
0.1	1934	3.05e-5	2.50e-1	4.5434e-3	3.0330e-3	704
0.4	2034	3.05e-5	2.50e-1	4.3241e-3	3.0330e-3	86
0.7	2202	3.05e-5	2.50e-1	3.9900e-3	3.0330e-3	31
0.95	2356	3.05e-5	2.50e-1	3.7186e-3	3.0330e-3	16
1	15872	6.10e-5	1.22e-4	3.8602e-3	3.0303e-3	14

Table 1.10: Dörfler strategy such that $\frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ holds with $T = 2$ in the case $\varepsilon = 0.01$.

We see that when θ is small, the number of nodes is small but it requires many iterations of the adaptive process. On the contrary, a large value of θ yields a partition of D with many nodes obtained with few iterations. Notice that here, all cases but $\theta = 1$ yield comparable results in terms of number of nodes, minimal and maximal mesh sizes and estimators. As mentioned above, the case $\theta = 1$ yields similar results to those obtained with uniform refinement of the mesh. The only difference lies in the fact that here, the midpoint of a subinterval $[x_i, x_{i+1}]$ is not added if $\eta_{1,i}$ is (numerically) zero. This explain why in Table 1.10 we get $h_{\min} \neq h_{\max}$. If we consider f_1 or f_2 as forcing term and $\varepsilon = 0.0625$, in which cases no local error estimator $\eta_{1,i}$ vanishes, we get $N_h = 512$ and $N_h = 2048$ for $f = f_1$ and $f = f_2$, respectively, as in Table 1.6.

Adaptation for higher-order approximation in ε

Here, we give only a sketch of a possible adaptive scheme to achieve an approximate solution with a prescribed accuracy, but we do not provide numerical experiments. As mentioned previously, further mesh refinement should be avoided once the two error estimators η_1 and η_2 are balanced since it would not decrease the total error. The latter can be decreased only by adding more terms in the expansion of u . Based on this observation, we can think of a strategy to adaptively increase the degree N in the expansion (1.19) of u together with adaptive mesh refinements for each deterministic term in this expansion. Recall that the estimator for the error $u - u_h^N = u - \sum_{n=0}^N \varepsilon^n u_{h,n}$ in the $L_P^2(\Omega; H_0^1(D))$ norm reads $a_{\min}^{-1}(C_{H_0^1} \sum_{n=0}^N \eta_n + \eta_{N+1})$, see Section 1.4.2. Starting with $N = 0$, we find a mesh of D (using Algorithm 2 for instance) such that $C_{H_0^1} \eta_0 \approx \eta_1$. If the error estimate does not reach the given tolerance, we increase N by one and find a mesh such that $C_{H_0^1}(\eta_0 + \eta_1) \approx \eta_2$ and proceed then iteratively. Notice that different meshes could be used for the FE approximation of each deterministic part of the solution (u_0, U_1, U_2, \dots) .

1.7.2 2D problems

The numerical results obtained for the one-dimensional case generalize to problems of higher dimensions. To motivate this statement, we present two numerical examples in 2D. In both cases, the physical domain is $D = (0, 1)^2$ that we partition using uniform meshes of size $h \sim 1/n$ for different values of n . The true error in the norm $L_p^2(\Omega; H_0^1(D))$ is computed via the Monte-Carlo method with sample size $K = 1000$ and a reference solution computed on the finest mesh considered which corresponds here to $n_{ref} = 2^8$.

First example

We consider first the problem (1.11) with $f(\mathbf{x}) = 32(x_1(1 - x_1) + x_2(1 - x_2))$ and

$$a(\mathbf{x}, \mathbf{Y}(\omega)) = 1 + \varepsilon \sum_{j=1}^5 \frac{\cos(2\pi j x_1) + \cos(2\pi j x_2)}{(\pi j)^2} Y_j(\omega)$$

for $\mathbf{x} = (x_1, x_2) \in D$, where Y_j , $j = 1, \dots, 5$, are uniform random variables in $[-\sqrt{3}, \sqrt{3}]$. In this setting, the exact solution u_0 for the deterministic case $\varepsilon = 0$ is given by $u_0(\mathbf{x}) = x_1 x_2 (1 - x_1)(1 - x_2)$. The expected value and the standard deviation of u for the case $\varepsilon = 0.5$ is given in Figure 1.7.

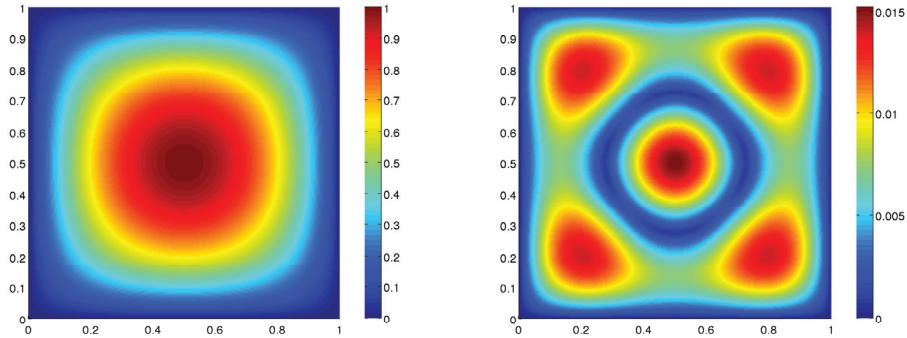


Figure 1.7: Expected value (left) and standard deviation (right) of the solution with $\varepsilon = 0.5$ for the first example.

Similarly to the 1D case, the constant due to interpolation can be estimated numerically, yielding³ $C_{H_0^1} := 1/5.7$. We define then $\tilde{\eta} = (C_{H_0^1}^2 \eta_1^2 + \eta_2^2)^{\frac{1}{2}}$ with η_1 and η_2 given by (1.43) and (1.44), respectively. We report in Table 1.11 the results obtained for $\varepsilon = 0.5$ fixed and uniform meshes of various sizes $h \sim 1/n$ while in Table 1.12, we fix $n = 64$ and vary ε .

In Table 1.12, where ε is fixed and n varies, the error decreases linearly with respect to $h \sim 1/n$

³If the factor $\frac{1}{2}$ is replaced by $\frac{1}{4}$ for the jump contribution, see Remark 1.7.1, then we should take $C_{H_0^1} := 1/5$. See Appendix 1.C for more details.

ε	error	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/\text{error}$
1	0.1749	0.0604	0.1842	0.1939	1.108
0.5	0.0974	0.0604	0.0921	0.1101	1.131
0.25	0.0703	0.0604	0.0461	0.0759	1.081
0.125	0.0622	0.0604	0.0230	0.0646	1.039
0.0625	0.0597	0.0604	0.0115	0.0615	1.029

Table 1.11: Error $\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$ with $n = 64$ for the first example.

n	error	$C_{H_0^1} \eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/\text{error}$
8	0.4891	0.4649	0.0927	0.4741	0.969
16	0.2551	0.2381	0.0923	0.2554	1.001
32	0.1439	0.1202	0.0922	0.1515	1.053
64	0.0974	0.0604	0.0921	0.1101	1.131
128	0.0833	0.0303	0.0921	0.0969	1.164

Table 1.12: Error $\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_p^2(\Omega; H_0^1(D))}$ with $\varepsilon = 0.5$ for the first example.

when η_2 is negligible compared to $C_{H_0^1} \eta_1$. When it is no longer the case, the error continues diminishing with refinement of the mesh but with a smaller rate. The same observation holds for the results of Table 1.11 switching the role of h and ε . Finally, we observe in both cases that the effectivity index of the error estimator $\tilde{\eta}$ that contains the estimated constant $C_{H_0^1}$ is close to 1.

Second example

Let $\{\lambda_i, \varphi_i\}$ be the eigenpairs of the Karhunen-Loève expansion of a (1D) Gaussian random field with exponential covariance function $C : D \times D \rightarrow \mathbb{R}$ given by

$$C(x, x') = \sigma^2 e^{-\frac{|x-x'|}{L_c}}$$

for which the analytical expression is known, see for instance [67] or [90]. We set the variance σ^2 and the correlation length L_c to $\sigma^2 = L_c = 1$ and we consider the random diffusion coefficient a obtained by tensorization

$$a(\mathbf{x}, \mathbf{Y}(\omega)) = 1 + \varepsilon \sum_{i=1}^3 \sum_{k=1}^3 \sqrt{\lambda_i \lambda_k} \varphi_i(x_1) \varphi_k(x_2) Y_{ik}(\omega) = 1 + \varepsilon \sum_{j=1}^9 a_j(\mathbf{x}) Y_j(\omega),$$

where Y_j , $j = 1, \dots, 9$, are uniform random variables in $[-\sqrt{3}, \sqrt{3}]$. Finally, we choose here $f(\mathbf{x}) = 10 \sin(2\pi(x_1 + x_2))$ for the forcing term. We give in Figure 1.8 the functions $\sqrt{\lambda_i \lambda_k} \varphi_i(x_1) \varphi_k(x_2)$

for $i, k = 1, 2, 3$. Notice that we can choose the global index j so that $\lambda_j = \lambda_i \lambda_k$ is non-decreasing but it is irrelevant here. Indeed, we do not perform a truncation on j and so an ordering to keep the more relevant functions is not required.

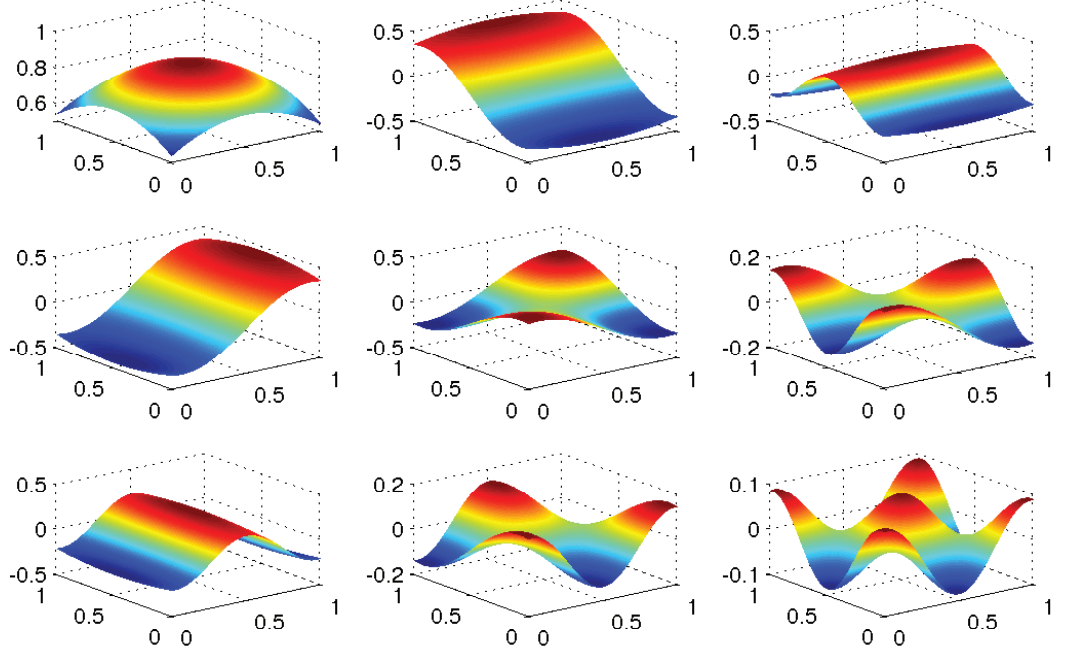


Figure 1.8: Plot of the functions a_j , $j = 1, \dots, 9$, constructed by tensorization of one-dimensional KL functions.

The expected value and the standard deviation of u for the case $\varepsilon = 0.5$ is given in Figure 1.9.

Finally, the results for a fixed $n = 128$ and a fixed $\varepsilon = 0.05$ are given in Tables 1.13 and 1.14, respectively.

The conclusions for this second example are the same as in the previous example.

1.7.3 Comparison with the stochastic collocation method

We finally illustrate the findings of Section 1.6 concerning the computation costs for the SC-FEM and the *perturbation method*. We consider the linear problem (1.11) and the nonlinear problem (1.75) with F given by (1.78). In both cases, homogeneous Dirichlet boundary condition are considered and we assume that the random variables Y_j , $j = 1, \dots, L$, that appear in the characterization (1.86) of a are uniform random variables in $[-\sqrt{3}, \sqrt{3}]$. We compare the computation time to solve the two problems with accuracy of order 2 in ε . Such accuracy is reached when we consider a sparse grid of level 1 for the SC-FEM method and the second

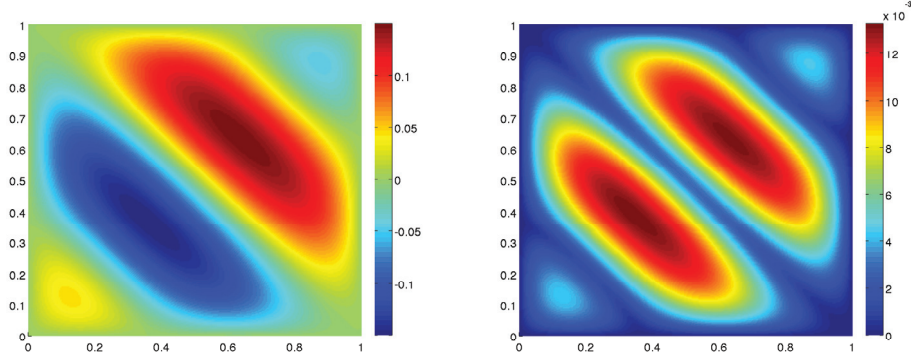


Figure 1.9: Expected value (left) and standard deviation (right) of the solution with $\varepsilon = 0.1$ for the second example.

ε	$error$	$C_{H_0^1}\eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
0.1	0.0623	0.0201	0.0605	0.0637	1.0227
0.05	0.0356	0.0201	0.0302	0.0363	1.0195
0.025	0.0245	0.0201	0.0151	0.0252	1.0269
0.0125	0.0210	0.0201	0.0076	0.0215	1.0263
0.00625	0.0200	0.0201	0.0038	0.0205	1.0274

Table 1.13: Error $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$ with $n = 128$ for the second example.

n	$error$	$C_{H_0^1}\eta_1$	η_2	$\tilde{\eta}$	$\tilde{\eta}/error$
8	0.3397	0.2762	0.0260	0.2774	0.8167
16	0.1804	0.1527	0.0291	0.1555	0.8616
32	0.0941	0.0791	0.0300	0.0848	0.9007
64	0.0527	0.0401	0.0302	0.0505	0.9577
128	0.0358	0.0201	0.0302	0.0367	1.0259

Table 1.14: Error $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$, estimators η_1 , η_2 and $\tilde{\eta}$ and ratio $\tilde{\eta}/\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$ with $\varepsilon = 0.05$ for the second example.

order approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$ for the *perturbation method*. Note that $u_{1,h} = \sum_{j=1}^L U_{j,h} Y_j$ where $U_{j,h}$ for $j = 1, \dots, L$ is the solution of

$$\int_D a_0 \nabla U_{j,h} \cdot \nabla v_h + \int_D 3u_{0,h}^2 U_{j,h} v_h = - \int_D a_j \nabla u_{0,h} \cdot v_h \quad \forall v_h \in V_h.$$

when problem (1.75) is considered. Finally, we use the same physical space discretization for both methods, namely a uniform partition with $h = 2^{-12}$. With this choice of mesh size, the work to solve the $(2L + 1)$ problems dominates the one needed to construct the grid. The computational time to solve both problems with respect to the number of random variables L is given in Figure 1.10.

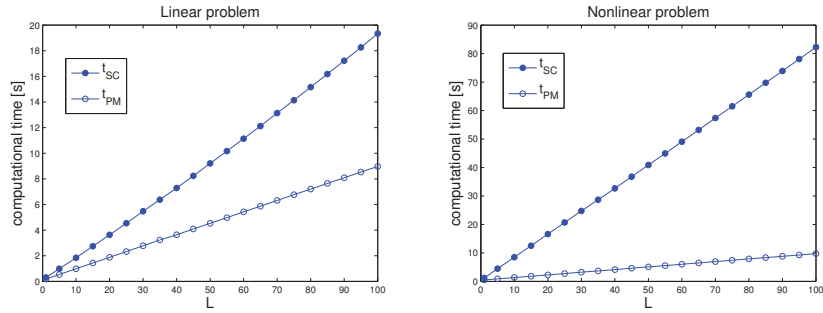


Figure 1.10: Time to solve the linear problem (1.11) and the nonlinear problem (1.78) with accuracy of order 2 in ε using the SC-FEM and the *perturbation method*.

As predicted in Section 1.6, the *perturbation method* presents no real advantage in terms of computation time over the stochastic collocation one, since it is only twice faster. This factor 2 comes from the fact that the *perturbation method* requires the solution of $L + 1$ problems, while $2L + 1$ problems need to be solve in the stochastic collocation method. The situation is different for nonlinear problems. In this case, the *perturbation method* is significantly faster than the stochastic collocation one. Indeed, only one nonlinear problem and L linear problems need to be solve for the former, to obtain respectively the deterministic part u_0 of u and the U_j , $j = 1, \dots, L$. For the SC method, we need to solve as many nonlinear problems as collocation points. Even for the nonlinear problem considered here, where the nonlinearity comes from the term u^3 and which is quite cheap to solve, the *perturbation method* is about 8 times faster.

To conclude, we can mention that for $h = h_{ref}$, i.e. without error due to FE approximation and a convergence of the error in $\mathcal{O}(\varepsilon^2)$, the error for the *perturbation method* is about 1.4 and 3.5 times larger than the error obtained using respectively SC1 and SC2. Again, the error for the *perturbation method* and the SC method has been accurately computed using the Monte Carlo method. However, for a given problem, that is for fixed value of ε and L , the *perturbation method* perform better than the SC method in terms of CPU time versus error for $h > h_{ref}$, especially for nonlinear problems. We plot in Figure 1.11 the computation time with respect to the error for problems (1.11) and (1.78) with $f = f_2$, $\varepsilon = 0.5$, $L = 10$ and $2^{-10} \leq h \leq 2^{-3}$. Notice

that the results for SC1 are not depicted on this figure since they are indistinguishable from those of SC2. Finally, we mention that it would be better, in terms of computational costs, to adapt the level l of the sparse grid for the SC-FEM, respectively the order in the approximation $u \approx \sum_{n=0}^l \varepsilon^n u_{n,h}$ for the *perturbation method*, with respect to h . Indeed, for the value of h for which the total error is not too small, namely of order ε or larger, it is more suitable to take $l = 0$ than $l = 1$ since comparable accuracy is reached at lower computational costs. However, the error due to the uncertainty, which is of order ε for $l = 0$, will be dominating at some point (see also Figure 1.5) and the value of l must be increased to be able to further reduce the error.

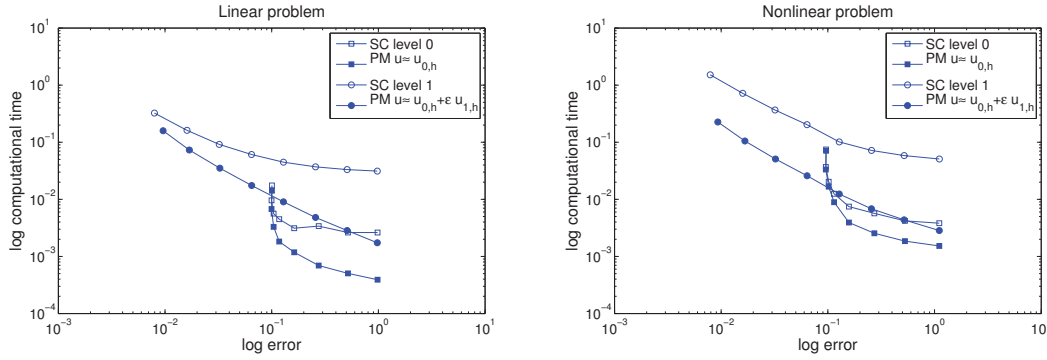


Figure 1.11: Log log scale plot of the computational time w.r.t. the error in $L^2_p(\Omega; H_0^1(D))$ -norm using the SC-FEM with Smolyak and Clenshaw-Curtis abscissas and the *perturbation method*.

Conclusions

In this chapter, we have performed error analyses for elliptic PDEs with coefficients affected by small uncertainties, characterized through random variables. The exact random solution has been approximated using a perturbation approach combined with the finite element method for the physical space discretization.

For the first order approximation $u \approx u_{0,h}$, we derived strong and weak *a priori* error estimates as well as *a posteriori* error estimates in the $L^2_p(\Omega; H_0^1(D))$ and $L^2_p(\Omega; L^2(D))$ norms. These estimates naturally split into two parts, namely the error in h due to the physical discretization and the error in ε due to the model. In the *a priori* error estimation, we have shown that the order of the weak error in the model is twice the order of the strong error, the order of the error due to FE approximation being the same in both cases. The *a posteriori* error estimator in the $L^2_p(\Omega; H_0^1(D))$ norm that we have obtained is a computable quantity of order $h + \varepsilon$ if the solution is regular enough in physical space. Given $u_{0,h}$, this estimator is cheap to compute and does not require any other FE solution. It can be used for mesh adaptation so that comparable accuracy in h and ε is reached. We have shown that taking the L^2 norm in physical space leads to a gain of one order in h but no improvement in the error due to the model. Finally, we gave a sketch of the derivation of a goal-oriented error estimate, which is more suitable than an estimate in global norm when a particular quantity of interest is

considered.

The *a posteriori* error estimation procedure for the error in the $L^2(\Omega; H_0^1(D))$ norm has been applied to the second-order approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$, before giving a generalization for approximations of any order. This reliable error upper bound can be used to adaptively determine the order of approximation and partitions of D such that the total error is below a given tolerance.

A posteriori error estimates have then been derived for a class of nonlinear problems through three different examples. A comparison in terms of computational costs with the stochastic collocation method has been performed, considering an error of order 2 in the model. The *perturbation method* presents only mild advantages for solving linear problems, the computational cost being halved with respect to the SC method. The situation is different for nonlinear problems. Indeed, the SC method requires the resolution of as many nonlinear problems as collocation points while for the *perturbation method*, only one nonlinear problem has to be solved for $u_{0,h}$, the remaining problems being linear.

1.A Derivation of problems (1.20), (1.21) and (1.22)

We make here some remarks about the derivation of the problems (1.20), (1.21) and (1.22) that we need to solve to build the approximate solution $u_0 + \varepsilon u_1 + \varepsilon^2 u_2$. In particular, we will see that the deterministic problems for the terms u_0 and u_1 are uniquely determined while those for u_2 are not. We thus discuss the various ways to build the term u_2 . Moreover, we will make a more precise link between each term and the derivatives of $u = u(\mathbf{x}, \mathbf{y})$ with respect to the y_j , $j = 1, \dots, L$.

Let us first give some details about the derivation of the problems. Recall that we assume that the diffusion coefficient a has the form

$$a(\mathbf{x}, \omega) = a(\mathbf{x}, \mathbf{Y}(\omega)) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L a_j(\mathbf{x}) Y_j(\omega).$$

Moreover, the random solution u is expanded as

$$u(\mathbf{x}, \mathbf{Y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{Y}(\omega)) + \varepsilon^2 u_2(\mathbf{x}, \mathbf{Y}(\omega)) + \dots$$

with $u_1 = \sum_{j=1}^L U_j Y_j$ and $u_2 = \sum_{j,k=1}^L U_{jk} Y_j Y_k$. Similar expansion can be used for the higher order terms, see (1.71) where the general case is treated or [126, 127]. If we substitute the expansions of a and u in the first equation of problem (1.13), we get

$$-\nabla \cdot \left((a_0 + \varepsilon \sum_{j=1}^L a_j Y_j) \nabla (u_0 + \varepsilon \sum_{j=1}^L U_j Y_j + \varepsilon^2 \sum_{j,k=1}^L U_{jk} Y_j Y_k + \dots) \right) = f.$$

After recalling that f is deterministic by assumption, we separate then the terms of different

order in ε . The equation for the $\mathcal{O}(1)$ term is

$$-\nabla \cdot (a_0 \nabla u_0) = f$$

which yields problem (1.20) after adding suitable boundary conditions. Next, the equation for the $\mathcal{O}(\varepsilon)$ term is

$$-\varepsilon \sum_{j=1}^L Y_j \nabla \cdot (a_0 \nabla U_j + a_j \nabla u_0) = 0. \quad (1.94)$$

Since the set $\{Y_j : j = 1, \dots, L\}$ is orthonormal, it is in particular linearly independent. Therefore, equation (1.94) holds if and only if each term is zero, i.e.

$$\nabla \cdot (a_0 \nabla U_j + a_j \nabla u_0) = 0 \quad \forall j = 1, \dots, L, \quad (1.95)$$

which is nothing else than the first equation of problem (1.21). Notice that the relation (1.95) can also be obtained by multiplying (1.94) by Y_k and taking the ensemble mean, see [127], thanks again to the fact that $\mathbb{E}[Y_j Y_k] = \delta_{jk}$. Finally, we collect the terms in $\mathcal{O}(\varepsilon^2)$ to obtain

$$-\varepsilon^2 \sum_{j,k=1}^L Y_j Y_k \nabla \cdot (a_0 \nabla U_{jk} + a_j \nabla U_k) = 0. \quad (1.96)$$

A sufficient condition for (1.96) to hold is that

$$\nabla \cdot (a_0 \nabla U_{jk} + a_j \nabla U_k) = 0 \quad \forall j, k = 1, \dots, L, \quad (1.97)$$

which corresponds to the set of PDEs in (1.22). However, it is not necessary to have (1.97) to verify (1.96) since the set $\{Y_j Y_k : j, k = 1, \dots, L\}$ is not linearly independent. Using the fact that $Y_j Y_k = Y_k Y_j$, we can rewrite (1.96) as

$$-\varepsilon^2 \sum_{1 \leq j \leq k \leq L} Y_j Y_k \nabla \cdot (a_0 \nabla (U_{jk} + U_{kj}) + a_j \nabla U_k + a_k \nabla U_j) \beta_{jk} = 0 \quad (1.98)$$

where $\beta_{jk} = 1 - \frac{1}{2} \delta_{jk}$ is introduced to allow to keep the cases $j < k$ and $j = k$ under the same summation sign. Now, the set $\{Y_j Y_k : 1 \leq j \leq k \leq L\}$ is linearly independent [127] and thus (1.98) holds if and only if

$$\nabla \cdot (a_0 \nabla (U_{jk} + U_{kj}) + a_j \nabla U_k + a_k \nabla U_j) = 0 \quad \forall 1 \leq j \leq k \leq L.$$

If we write $\tilde{U}_{jk} := \frac{U_{jk} + U_{kj}}{2}$ for $j, k = 1, \dots, L$ we have then

$$u_2 = \sum_{j,k=1}^L U_{jk} Y_j Y_k = \sum_{1 \leq j \leq k \leq L} \beta_{jk} (U_{jk} + U_{kj}) Y_j Y_k = \sum_{j,k=1}^L \tilde{U}_{jk} Y_j Y_k.$$

Notice that \tilde{U}_{jk} solves

$$-\nabla \cdot \left(a_0 \nabla \tilde{U}_{jk} + \frac{a_j \nabla U_k + a_k \nabla U_j}{2} \right) \quad \forall j, k = 1, \dots, L$$

and $\tilde{U}_{jk} + \tilde{U}_{kj} = U_{jk} + U_{kj}$. The advantage of building u_2 with the \tilde{U}_{jk} instead of the U_{jk} relies in the fact that $\tilde{U}_{jk} = \tilde{U}_{kj}$ while U_{jk} is not necessarily equal to U_{kj} . Therefore, the construction of u_2 with the \tilde{U}_{jk} requires the resolution of $\frac{L(L+1)}{2}$ whereas L^2 problems need to be solved when the U_{jk} are used.

Notice that the problems we obtain for u_0 , U_j , U_{jk} and $U_{j_1 j_2 \dots j_n}$, given by (1.20), (1.21), (1.22) and (1.71), respectively, are equivalent to those derived in [6]. In that paper, the authors apply what they called the method of successive approximations which uses the Karhunen-Loève expansion to represent the stochastic diffusion coefficient combined with the Neumann series expansion method. In fact, applied to the specific linear elliptic diffusion model problem (1.11), the (*generalized* or standard) Neumann expansion method and the perturbation method are equivalent [121].

In the remaining part of this section, we clarify the link between the various terms u_0 , U_j , U_{jk} and \tilde{U}_{jk} defined above and the derivatives of $u = u(\mathbf{x}, \mathbf{y})$ with respect to the y_j . In other words, we compare the expansion (1.19) of u with its Taylor expansion around $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$. Recall that it has been proved (see for instance [7]) that the weak solution $u = u(\mathbf{x}, \mathbf{y})$ of problem (1.13), i.e. the solution of (1.14), is analytic with respect to each variable y_j , $j = 1, \dots, L$. First of all, we have

$$a(\mathbf{x}, \mathbf{y}_0) = a_0(\mathbf{x}), \quad \frac{\partial a}{\partial y_j}(\mathbf{x}, \mathbf{y}_0) = \varepsilon a_j(\mathbf{x}) \quad \text{and} \quad \frac{\partial^2 a}{\partial y_k \partial y_j}(\mathbf{x}, \mathbf{y}_0) = 0 \quad \forall j, k = 1, \dots, L.$$

Then, we recall that for each $\mathbf{y} \in \Gamma$ the solution $u(\cdot, \mathbf{y}) \in H_0^1(D)$ of problem (1.14) satisfies

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (1.99)$$

The evaluation of equation (1.99) at \mathbf{y}_0 yields

$$\int_D a_0(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}_0) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}. \quad (1.100)$$

We can formally differentiate equation (1.99) with respect to y_j to get

$$\int_D \left(\frac{\partial a}{\partial y_j} \nabla u + a \nabla \frac{\partial u}{\partial y_j} \right) (\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0, \quad j = 1, \dots, L, \quad (1.101)$$

and thus for $\mathbf{y} = \mathbf{y}_0$ we have

$$\int_D \left(\varepsilon a_j(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}_0) + a_0(\mathbf{x}) \nabla \frac{\partial u}{\partial y_j}(\mathbf{x}, \mathbf{y}_0) \right) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0, \quad j = 1, \dots, L. \quad (1.102)$$

Taking then the derivative of (1.102) with respect to y_k , or equivalently the second derivative of (1.99), we obtain for $j, k = 1, \dots, L$ the relation

$$\int_D \left(\frac{\partial^2 a}{\partial y_k \partial y_j} \nabla u + \frac{\partial a}{\partial y_j} \nabla \frac{\partial u}{\partial y_k} + \frac{\partial a}{\partial y_k} \nabla \frac{\partial u}{\partial y_j} + a \nabla \frac{\partial^2 u}{\partial y_k \partial y_j} \right) (\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0.$$

Since $\frac{\partial^2 a}{\partial y_k \partial y_j} = 0$, the evaluation of last relation at \mathbf{y}_0 gives us

$$\int_D \left(\varepsilon a_j(\mathbf{x}) \nabla \frac{\partial u}{\partial y_k}(\mathbf{x}, \mathbf{y}_0) + \varepsilon a_k(\mathbf{x}) \nabla \frac{\partial u}{\partial y_j}(\mathbf{x}, \mathbf{y}_0) + a_0(\mathbf{x}) \nabla \frac{\partial^2 u}{\partial y_k \partial y_j}(\mathbf{x}, \mathbf{y}_0) \right) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0, \quad j, k = 1, \dots, L. \quad (1.103)$$

Finally, based on equations (1.100), (1.102) and (1.103) we conclude that

$$u_0 = u(\cdot, \mathbf{y}_0), \quad \varepsilon U_j = \frac{\partial u}{\partial y_j}(\cdot, \mathbf{y}_0), \quad \varepsilon^2(U_{jk} + U_{kj}) = \frac{\partial^2 u}{\partial y_k \partial y_j}(\cdot, \mathbf{y}_0) \quad \text{and} \quad \varepsilon^2 \tilde{U}_{jk} = \frac{1}{2} \frac{\partial^2 u}{\partial y_k \partial y_j}(\cdot, \mathbf{y}_0)$$

for $j, k = 1, \dots, L$.

1.B Upper and lower bounds for the error $u - u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm

The goal here is to prove that the error estimator introduced in (1.50) provides both lower and upper bounds for the error $\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))}$. We assume here that $D \subset \mathbb{R}^d$ with $d = 2$, mentioning that the case $d = 1$ can be treated easily since no jump terms occur while the extension to the case $d = 3$ is straightforward. We first introduce the estimator in more details, starting from the relation

$$\begin{aligned} \mathcal{A}(u - u_{0,h}, v; \mathbf{y}) &= \int_D f v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v \\ &= \mathcal{R}(v; \mathbf{y}_0) + [\mathcal{R}(v; \mathbf{y}) - \mathcal{R}(v; \mathbf{y}_0)] \end{aligned}$$

for all $v \in H_0^1(D)$ and ρ -a.e. in Γ , where $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$ and

$$\mathcal{R}(v; \mathbf{y}) := F(v) - \mathcal{A}(u_{0,h}, v; \mathbf{y}) = \int_D f v - \int_D a(\cdot, \mathbf{y}) \nabla u_{0,h} \cdot \nabla v.$$

For any $\mathbf{y} \in \Gamma$, let $r(\cdot; \mathbf{y}) : H_0^1(D) \rightarrow \mathbb{R}$ be defined by

$$r(v; \mathbf{y}) := \mathcal{R}(v; \mathbf{y}) - \mathcal{R}(v; \mathbf{y}_0) = - \int_D (a(\cdot, \mathbf{y}) - a_0) \nabla u_{0,h} \cdot \nabla v. \quad (1.104)$$

The dual norm of r is then given by $\|r(\cdot; \mathbf{y})\|_{H^{-1}(D)} = \|\nabla w(\cdot, \mathbf{y})\|_{L^2(D)}$ with $w(\cdot, \mathbf{y})$ the solution of

$$\int_D \nabla w(\cdot, \mathbf{y}) \cdot \nabla v = r(v; \mathbf{y}) \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (1.105)$$

1.B. Upper and lower bounds for the error $u - u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm

We write then $w(\mathbf{x}, \mathbf{Y}(\omega)) = \varepsilon \sum_{j=1}^L W_j(\mathbf{x}) Y_j(\omega)$ with $W_j \in H_0^1(D)$ such that

$$\int_D \nabla W_j \cdot \nabla v = - \int_D a_j \nabla u_{0,h} \cdot \nabla v \quad \forall v \in H_0^1(D).$$

Let $w_h(\mathbf{x}, \mathbf{Y}(\omega)) = \varepsilon \sum_{j=1}^L W_{j,h}(\mathbf{x}) Y_j(\omega)$, where $W_{j,h} \in V_h$ is the FE approximation of W_j , and let R and J denote the interior element residual and the jump defined on an element K and an internal edge e by respectively

$$R|_K = (f + \nabla \cdot (a_0 \nabla u_{0,h}))|_K \quad \text{and} \quad J|_e = [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e}.$$

The spatial and stochastic *a posteriori* error estimators η_1 and $\hat{\eta}_2$ are given by (1.43) and (1.50), respectively, definitions that we recall here for clarity

$$\eta_1^2 := \sum_{K \in \mathcal{T}_h} \eta_K^2 \quad \text{with} \quad \eta_K^2 = h_K^2 \|R\|_{L^2(K)}^2 + \frac{1}{2} \sum_{e \subset \partial K} h_e \|J\|_{L^2(e)}^2, \quad (1.106)$$

$$\hat{\eta}_2^2 := \varepsilon^2 \sum_{j=1}^L \|\nabla W_{j,h}\|_{L^2(D)}^2. \quad (1.107)$$

To prove the *spatial lower bound*, see (1.112), we will need some definitions and notation that we introduce now.

For any element $K \in \mathcal{T}_h$, using the notation given in Figure 1.12-left, we define the so-called element bubble function ψ_K and edge bubble function ψ_{e_i} , see for instance [118], by

$$\psi_K = 27\lambda_1\lambda_2\lambda_3 \quad \text{and} \quad \psi_{e_i} = 4\lambda_{i+1}\lambda_{i+2}, \quad i = 1, 2, 3,$$

where the indices are taken modulo 3 and $\lambda_1, \lambda_2, \lambda_3$ are the (linear) barycentric coordinates on K . Using the notation used in [118], we denote by w_K the union of all the elements sharing an edge with K and, for an internal edge e , we write w_e the union of the two elements sharing e as an edge, see Figure 1.12 for an illustration.

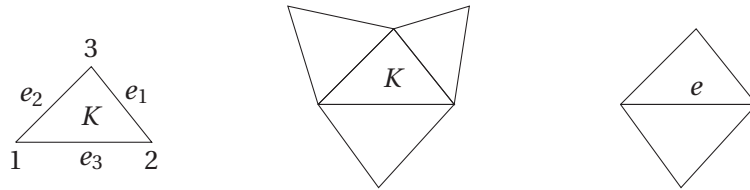


Figure 1.12: Notation for an element K in \mathcal{T}_h (left) and illustration of the domains w_K (middle) and w_e (right).

The bubble functions satisfy the following properties: for any polynomial φ of degree less or equal to k we have

$$\|\varphi\|_{L^2(K)} \leq c_1 \|\psi_K^{\frac{1}{2}} \varphi\|_{L^2(K)}, \quad \|\nabla(\psi_K \varphi)\|_{L^2(K)} \leq c_2 h_K^{-1} \|\varphi\|_{L^2(K)} \quad (1.108)$$

and

$$\|\varphi\|_{L^2(e)} \leq c_3 \|\psi_e^{\frac{1}{2}} \varphi\|_{L^2(e)}, \quad \|\nabla(\psi_e \varphi)\|_{L^2(w_e)} \leq c_4 h_e^{-\frac{1}{2}} \|\varphi\|_{L^2(e)}, \quad \|\psi_e \varphi\|_{L^2(w_e)} \leq c_5 h_e^{\frac{1}{2}} \|\varphi\|_{L^2(e)}, \quad (1.109)$$

where the constants C_i , $i = 1, \dots, 5$, depend only on k and on the shape regularity parameter of \mathcal{T}_h given in (1.23). Moreover, we have

$$0 \leq \psi_K(\mathbf{x}) \leq 1 \quad \forall \mathbf{x} \in K, \quad \psi_K(\mathbf{x}) = 0 \quad \forall \mathbf{x} \notin K, \quad \max_{\mathbf{x} \in K} \psi_K(\mathbf{x}) = 1$$

and

$$0 \leq \psi_e(\mathbf{x}) \leq 1 \quad \forall \mathbf{x} \in w_e, \quad \psi_e(\mathbf{x}) = 0 \quad \forall \mathbf{x} \notin w_e, \quad \max_{\mathbf{x} \in w_e} \psi_e(\mathbf{x}) = 1.$$

For any element K , we denote by \bar{g}_K the mean value of g on K and similarly we denote by \bar{g}_e the mean value of g on any internal edge e , i.e.

$$\bar{g}_K = \frac{1}{|K|} \int_K g \quad \text{and} \quad \bar{g}_e = \frac{1}{|e|} \int_e g.$$

Finally, we introduce the *oscillation term* θ_K defined by

$$\theta_K^2 := \sum_{T \subset w_K} h_T^2 \|R - \bar{R}_T\|_{L^2(T)}^2 + \sum_{e \subset \partial K} h_e \|J - \bar{J}_e\|_{L^2(e)}^2. \quad (1.110)$$

We can now state the upper and lower bounds, given in the following proposition.

Proposition 1.B.1. *Let u be the weak solution of problem (1.11) and let $u_{0,h}$ be the solution of problem (1.30), respectively. There exist two constants $C_1, C_2 > 0$ depending only on the mesh aspect ratio and $s \in (0, 1]$ such that*

$$\|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))} \leq \frac{1}{a_{\min}} (C_1 \eta_1 + \hat{\eta}_2) + \mathcal{O}(\varepsilon h^s), \quad (1.111)$$

$$\eta_1 \leq C_2 \left[a_{\max} \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))} + \hat{\eta}_2 + \left(\sum_{K \in \mathcal{T}_h} \theta_K^2 \right)^{\frac{1}{2}} \right] + \mathcal{O}(\varepsilon h^s) \quad (1.112)$$

and

$$\hat{\eta}_2 \leq a_{\max} \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))} + C_1 \eta_1 + \mathcal{O}(\varepsilon h^s). \quad (1.113)$$

Proof. We first derive a bound for the $L_P^2(\Omega; H_0^1(D))$ norm of w (resp. w_h) in term of the norm of w_h (resp. w) and higher order terms, where w is the solution (1.105) and w_h its FE approximation. Let us introduce $\psi(\mathbf{x}, \mathbf{Y}(\omega)) = \varepsilon \sum_{j=1}^L \psi_j(\mathbf{x}) Y_j(\omega)$, where $\psi_j \in H_0^1(D)$ is the solution of

$$\int_D \nabla \psi_j \cdot \nabla v = - \int_D a_j \nabla u_0 \cdot \nabla v \quad \forall v \in H_0^1(D),$$

1.B. Upper and lower bounds for the error $u - u_{0,h}$ in the $L_p^2(\Omega; H_0^1(D))$ norm

and let ψ_h denotes its FE approximation. Notice that $\psi(\cdot, \mathbf{Y}(\omega))$ solves

$$\int_D \nabla \psi \cdot \nabla v = - \int_D (a - a_0) \nabla u_0 \cdot \nabla v \quad \forall v \in H_0^1(D), \text{ a.s. in } \Omega,$$

which is similar to the problem (1.105) for w , except that $u_{0,h}$ is replaced by u_0 in the right-hand side. Thanks to the triangle inequality, we obtain

$$\|\nabla w\|_{L^2(D)} \leq \|\nabla w_h\|_{L^2(D)} + \|\nabla(w - \psi)\|_{L^2(D)} + \|\nabla(\psi - \psi_h)\|_{L^2(D)} + \|\nabla(\psi_h - w_h)\|_{L^2(D)}$$

from which we can deduce

$$\|\nabla w\|_{L_p^2(\Omega; L^2(D))} \leq \|\nabla w_h\|_{L_p^2(\Omega; L^2(D))} + C\varepsilon h^s,$$

where $s \in (0, 1]$ depends only on the regularity of u_0 , ψ_j , $j = 1, \dots, L$, and the domain D and C is a (deterministic) positive constant independent of h and ε but dependent on the mesh aspect ratio, $|u_0|_{H^{1+s}(D)}$ and $|\psi_j|_{H^{1+s}(D)}$, $j = 1, \dots, L$. Therefore, recalling that $w_h = \varepsilon \sum_{j=1}^L W_{j,h} Y_j$ and using $\mathbb{E}[Y_i Y_j] = \delta_{ij}$ we get

$$\|\nabla w\|_{L_p^2(\Omega; L^2(D))} \leq \hat{\eta}_2 + C\varepsilon h^s \quad (1.114)$$

with $\hat{\eta}_2$ given in (1.107). Finally, proceeding in the same way we can obtain the relation

$$\hat{\eta}_2 = \|\nabla w_h\|_{L_p^2(\Omega; L^2(D))} \leq \|\nabla w\|_{L_p^2(\Omega; L^2(D))} + C\varepsilon h^s. \quad (1.115)$$

We now prove the three bounds (1.111), (1.112) and (1.113) separately. The proof of (1.112) is inspired by what is done in [99, 118], while the idea for the proof of (1.113) is based on the proof of efficiency of the error estimator proposed in [102] for the Reduced Basis method. In the sequel, all the equations hold a.s. in Ω without specifically mentioning it.

Upper bound The proof is similar to the one of Proposition 1.3.5, only the bound of term controlling the stochastic error is different. For any $v \in H_0^1(D)$, taking $v_h = I_h$ the Clément interpolant of v we have

$$\begin{aligned} \int_D a \nabla(u - u_{0,h}) \cdot \nabla v &= \int_D f v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v \\ &= \int_D f(v - v_h) - \int_D a_0 \nabla u_{0,h} \cdot \nabla(v - v_h) + \int_D \nabla w \cdot \nabla v \\ &\leq \left[C_1 \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{\frac{1}{2}} + \|\nabla w\|_{L^2(D)} \right] \|\nabla v\|_{L^2(D)}, \end{aligned}$$

where C_1 depends only on the constants in (1.26) and (1.28). Since a_{\min} is a lower bound for a , taking $v = u - u_{0,h}$ we get

$$a_{\min} \|\nabla(u - u_{0,h})\|_{L^2(D)} \leq C_1 \eta_1 + \|\nabla w\|_{L^2(D)}$$

and thus, taking the $L_P^2(\Omega)$ norm on both sides of the last inequality we have

$$a_{min} \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))} \leq C_1 \eta_1 + \|\nabla w\|_{L_P^2(\Omega; L^2(D))}.$$

Finally, we obtain (1.111) using (1.114).

h -lower bound First of all, notice that for any $v \in H_0^1(D)$ we have

$$\begin{aligned} \int_D a \nabla(u - u_{0,h}) \cdot \nabla v &= \sum_{K \in \mathcal{T}_h} \int_K Rv + \sum_{e \in \mathcal{E}_h} \int_e Jv - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v \\ &= \sum_{K \in \mathcal{T}_h} \int_K Rv + \sum_{e \in \mathcal{E}_h} \int_e Jv + \int_D \nabla w \cdot \nabla v. \end{aligned} \quad (1.116)$$

The proof is then divided into three steps.

1. Let K be any element in \mathcal{T}_h and let $v_K = \bar{R}_K \psi_K$. We take $v = v_K$ in (1.116). Since $\text{supp} \psi_K \subset K$, we have

$$\int_K a \nabla(u - u_{0,h}) \cdot \nabla v_K = \int_K \bar{R}_K v_K + \int_K (R - \bar{R}_K) v_K + \int_K \nabla w \cdot \nabla v_K$$

and thus, using the properties of the element bubble function given in (1.108), we obtain

$$h_K \|\bar{R}_K\|_{L^2(K)} \leq c_1^2 c_2 a_{max} \|\nabla(u - u_{0,h})\|_{L^2(K)} + c_1^2 c_2 \|\nabla w\|_{L^2(K)} + c_1^2 h_K \|R - \bar{R}_K\|_{L^2(K)}.$$

Thanks to triangle's inequality, we finally obtain

$$h_K \|R\|_{L^2(K)} \leq c_1^2 c_2 a_{max} \|\nabla(u - u_{0,h})\|_{L^2(K)} + c_1^2 c_2 \|\nabla w\|_{L^2(K)} + (1 + c_1^2) h_K \|R - \bar{R}_K\|_{L^2(K)}. \quad (1.117)$$

2. Let e be any interior edge of \mathcal{T}_h , let $v_e = \bar{J}_e \psi_e$ and let K_1 and K_2 be the two elements that share e as an edge. We take $v = v_e$ in (1.116) to get

$$\int_{w_e} a \nabla(u - u_{0,h}) \cdot \nabla v_e = \sum_{K \in w_e} \int_K Rv_e + \int_e \bar{J}_e v_e + \int_e (J - \bar{J}_e) v_e + \int_{w_e} \nabla w \cdot \nabla v_e.$$

Therefore, using the properties of the edge bubble function given in (1.109), we obtain

$$\begin{aligned} h_e^{\frac{1}{2}} \|\bar{J}_e\|_{L^2(e)} &\leq c_3^2 c_4 a_{max} \|\nabla(u - u_{0,h})\|_{L^2(w_e)} + c_3^2 c_5 h_e \|R\|_{L^2(w_e)} \\ &\quad + c_3^2 h_e^{\frac{1}{2}} \|J - \bar{J}_e\|_{L^2(e)} + c_3^2 c_4 \|\nabla w\|_{L^2(w_e)} \end{aligned}$$

and thus

$$\begin{aligned}
 h_e^{\frac{1}{2}} \|J\|_{L^2(e)} &\leq c_3^2 c_4 a_{max} \|\nabla(u - u_{0,h})\|_{L^2(w_e)} + c_3^2 c_5 h_e \|R\|_{L^2(w_e)} + (1 + c_3^2) h_e^{\frac{1}{2}} \|J - \bar{J}_e\|_{L^2(e)} \\
 &\quad + c_3^2 c_4 \|\nabla w\|_{L^2(w_e)} \\
 &\leq \sum_{i=1}^2 [a_{max} c_3^2 (c_4 + c_1^2 c_2 c_5) \|\nabla(u - u_{0,h})\|_{L^2(K_i)} + (1 + c_1^2) c_3^2 c_5 h_{K_i} \|R - \bar{R}_{K_i}\|_{L^2(K_i)} \\
 &\quad + c_3^2 (c_4 + c_1^2 c_2 c_5) \|\nabla w\|_{L^2(K_i)}] + (1 + c_3^2) h_e^{\frac{1}{2}} \|J - \bar{J}_e\|_{L^2(e)}
 \end{aligned}$$

using relation (1.117).

3. Putting everything together, we obtain for any element $K \in \mathcal{T}_h$

$$\begin{aligned}
 \eta_K^2 &= h_K^2 \|R\|_{L^2(K)}^2 + \frac{1}{2} \sum_{e \in \partial K} h_e \|J\|_{L^2(e)}^2 \\
 &\leq C_2 \left(a_{max}^2 \|\nabla(u - u_{0,h})\|_{L^2(w_K)}^2 + \|\nabla w\|_{L^2(w_K)}^2 \right. \\
 &\quad \left. + \sum_{T \subset w_K} h_T^2 \|R - \bar{R}_T\|_{L^2(T)}^2 + \sum_{e \in \partial K} h_e \|J - \bar{J}_e\|_{L^2(e)}^2 \right),
 \end{aligned}$$

where C_2 depends only on the regularity of the mesh (through the constants c_i , $i = 1, \dots, 5$). Recalling the definition of θ_K in (1.110), if we sum over all $K \in \mathcal{T}_h$ and use the relation $(a^2 + b^2 + c^2) \leq (a + b + c)^2$ valid for any non-negative numbers a, b, c , we get

$$\eta_1 \leq C_2 \left[a_{max} \|\nabla(u - u_{0,h})\|_{L^2(D)} + \|\nabla w\|_{L^2(D)} + \left(\sum_{K \in \mathcal{T}_h} \theta_K^2 \right)^{\frac{1}{2}} \right]$$

where C_2 has changed but still only depends on the mesh aspect ratio. Finally, we obtain (1.112) taking the $L_P^2(\Omega)$ norm and using (1.114).

ε -lower bound For any $v \in H_0^1(D)$ we have

$$\int_D \nabla w \cdot \nabla v = - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v = \int_D a \nabla(u - u_{0,h}) \cdot \nabla v - \int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v. \quad (1.118)$$

Taking $v = w$ in (1.118) and noticing that the last term of (1.118) is nothing else than (minus) the residual for $u_{0,h}$, we can easily derive the bound

$$\|\nabla w\|_{L^2(D)}^2 \leq \left[a_{max} \|\nabla(u - u_{0,h})\|_{L^2(D)} + C_1 \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{\frac{1}{2}} \right] \|\nabla w\|_{L^2(D)}$$

where C_1 depends only on the constants in (1.26) and (1.28). From the last relation, we deduce taking the $L_P^2(\Omega)$ that

$$\|\nabla w\|_{L_P^2(\Omega; L^2(D))} \leq a_{max} \|u - u_{0,h}\|_{L_P^2(\Omega; H_0^1(D))} + C_1 \eta_1$$

which conclude the proof thanks to (1.115). \square

Remark 1.B.2. Since $u_{0,h}$ is piecewise affine, if a_0 is piecewise constant then we have $R = f$ and $J = \bar{J}_e$. Therefore, in this case θ_K reduces to $\sum_{T \in \mathcal{T}_K} h_T^2 \|f - \bar{f}_T\|_{L^2(T)}^2$ which does no longer depend on $u_{0,h}$. It is often referred to as data oscillation.

Remark 1.B.3. We deduce from the three relations (1.111), (1.112) and (1.113) that

$$a_{\min} \leq \frac{\hat{\eta}_2}{\|u - u_{0,h}\|} \leq a_{\max} \quad \text{as } h \rightarrow 0$$

and

$$C_1^{-1} a_{\min} \leq \frac{\eta_1}{\|u - u_{0,h}\|} \leq C_2 a_{\max} \quad \text{as } \varepsilon \rightarrow 0,$$

where $\|\cdot\|$ denotes the $L^2(\Omega; H_0^1(D))$ norm and C_1 and C_2 are two positive constants depending only on the mesh aspect ratio.

1.C Estimation of the interpolation constant

In this section, we briefly present the value of the interpolation constant $C_{H_0^1}$ that can be included in the error estimator to get a sharp spatial error estimator. This value depends on the degree of the finite element space as well as if we are in 1D, 2D or 3D.

In the one-dimensional case, we have already mentioned that the constant for \mathbb{P}_1 finite element can be set to $C_{H_0^1} = \frac{1}{3.46} \approx \frac{1}{2\sqrt{3}}$. The latter corresponds to the theoretical value $\left(\frac{1}{p+1}\right)^{1/p} \frac{1}{2}$ with $p = 2$ given in [9].

For the 2D case, we consider the (deterministic) Poisson problem $-\Delta u_0 = f$ with homogeneous Dirichlet boundary conditions. We set $D = (0, 1)^2$ and $u_0(x_1, x_2) = \sin(2\pi x_1) \sin(4\pi x_2)$ and compute the corresponding right-hand side given by

$$f(x_1, x_2) = 20\pi^2 \sin(2\pi x_2) \sin(4\pi x_2). \quad (1.119)$$

We give in Table 1.15 the error $\|\nabla(u_0 - u_{0,h})\|_{L^2(D)}$ and the two estimators η_1 and $\hat{\eta}_1$ defined by

$$\eta_1^2 = \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \Delta u_h\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} h_e \|\nabla u_h \cdot \mathbf{n}_e\|_{L^2(e)}^2$$

and

$$\hat{\eta}_1^2 = \sum_{K \in \mathcal{T}_h} \left[h_K^2 \|f + \Delta u_h\|_{L^2(K)}^2 + \frac{1}{4} \sum_{e \in \partial K} h_e \|\nabla u_h \cdot \mathbf{n}_e\|_{L^2(e)}^2 \right].$$

We consider both structured and Delaunay triangulations with $N = 256$ equidistant vertices on each boundary of D , see Figure 1.13 where the meshes for the case $N = 16$ are given.

The constant $1/C_{H_0^1}$ can then be set to $\eta_1 / \|\nabla(u_0 - u_{0,h})\|_{L^2(D)}$ or $\hat{\eta}_1 / \|\nabla(u_0 - u_{0,h})\|_{L^2(D)}$ depending on the definition of the estimator.

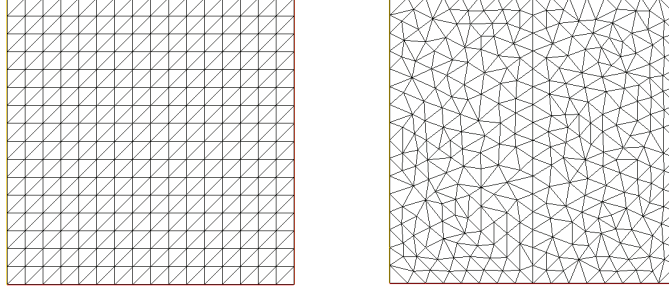


Figure 1.13: Structured (left) and Delaunay (right) triangulations of D with $N = 16$.

	Structured mesh					Delaunay mesh				
	<i>error</i>	η_1	e.i.	$\hat{\eta}_1$	e.i.	<i>error</i>	η_1	e.i.	$\hat{\eta}_1$	e.i.
\mathbb{P}_1	1.279e-1	7.352e-1	5.75	6.472e-1	5.06	1.037e-1	5.934e-1	5.72	5.296e-1	5.11
\mathbb{P}_{1b}	1.204e-1	5.225e-1	4.34	3.939e-1	3.27	9.450e-2	3.590e-1	3.80	2.668e-1	2.82
\mathbb{P}_2	9.592e-4	8.464e-3	8.82	8.195e-3	8.54	6.905e-4	6.473e-3	9.37	6.385e-3	9.25
\mathbb{P}_3	3.130e-6	7.136e-5	22.80	6.924e-5	22.12	2.017e-3	4.865e-5	24.12	4.749e-5	23.55

Table 1.15: Error, estimator and effectivity index for the Poisson problem.

Notice that we get similar values when considering other cases than (1.119). We see from the results of Table 1.15 that, as expected, the interpolation constant depends on the polynomial degree of the finite elements. Moreover, we could go further by estimating separately the efficiency of the interior residual and the contribution of the jump terms, but we will not do it in this thesis.

2 Elliptic model problems with other sources of uncertainty

Introduction

We extend here the results of Chapter 1 to include other sources of uncertainty. We first consider the case of random Neumann boundary conditions. The analysis is very similar to the one presented in Chapter 1. It is even easier in this case since the solution u depends linearly on the random input, and thus only the first two terms in the expansion are non-zero. We consider then the case where two random input data are affected by uncertainty, namely we consider a random diffusion coefficient combined with a random forcing term. Two different sets of random variables are used to describe each uncertain input data. Finally, numerical results are given to illustrate the theoretical findings.

2.1 Neumann random boundary conditions

We consider the problem:

find $u : D \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω :

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x})\nabla u(\mathbf{x},\omega)) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x},\omega) &= 0 & \mathbf{x} \in \Gamma_D \\ a_0(\mathbf{x})\frac{\partial u(\mathbf{x},\omega)}{\partial \mathbf{n}} &= g(\mathbf{x},\omega) & \mathbf{x} \in \Gamma_N, \end{cases} \quad (2.1)$$

where $\Gamma_D \cup \Gamma_N = \partial D$ with $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma_D \neq \emptyset$. We assume that a_0 is bounded from below by $a_{0,min}$ and that g is characterized by L independent random variables $\{Y_j\}_{j=1}^L$ with zero mean and unit variance as

$$g(\mathbf{x},\omega) = g(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega)) = g_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L g_j(\mathbf{x}) Y_j(\omega) \quad (2.2)$$

with $g_j \in L^2(\Gamma_N)$, $j = 0, 1, \dots, L$. Using the same notation as in the previous chapter, we can rewrite problem (2.1) in parametric form as:

find $u : D \times \Gamma \rightarrow \mathbb{R}$ such that ρ -a.e. in Γ we have:

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x})\nabla u(\mathbf{x}, \mathbf{y})) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x}, \mathbf{y}) &= 0 & \mathbf{x} \in \Gamma_D \\ a_0(\mathbf{x})\frac{\partial u(\mathbf{x}, \mathbf{y})}{\partial \mathbf{n}} &= g(\mathbf{x}, \mathbf{y}) & \mathbf{x} \in \Gamma_N, \end{cases} \quad (2.3)$$

whose weak formulation reads:

find $u(\cdot, \mathbf{y}) \in W$ such that

$$\int_D a_0 \nabla u(\cdot, \mathbf{y}) \cdot \nabla v = \int_D f v + \int_{\Gamma_N} g(\cdot, \mathbf{y}) v \quad \forall v \in W, \rho\text{-a.e. in } \Gamma \quad (2.4)$$

with $W := H_{\Gamma_D}^1(D) = \{v \in H^1(D) : v = 0 \text{ on } \Gamma_D\}$ that we endow with the gradient norm $\|\cdot\|_W := \|\nabla \cdot\|_{L^2(D)}$. This can be done thanks to the Friedrich-Poincaré inequality

$$\|v\|_{L^2(D)} \leq C_F \|\nabla v\|_{L^2(D)} \quad \forall v \in W, \quad (2.5)$$

which holds as long as $\Gamma_D \neq \emptyset$. Using again a perturbation technique, we write

$$u(\mathbf{x}, \mathbf{Y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{Y}(\omega)) + \varepsilon^2 u_2(\mathbf{x}, \mathbf{Y}(\omega)) + \dots$$

where $u_0 : D \rightarrow \mathbb{R}$ is the solution of

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x})\nabla u_0(\mathbf{x})) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u_0(\mathbf{x}) &= 0 & \mathbf{x} \in \Gamma_D \\ a_0(\mathbf{x})\frac{\partial u_0(\mathbf{x})}{\partial \mathbf{n}} &= g_0(\mathbf{x}) & \mathbf{x} \in \Gamma_N, \end{cases} \quad (2.6)$$

and $u_1 = \sum_{j=1}^L U_j Y_j$ with $U_j : D \rightarrow \mathbb{R}$, $j = 1, \dots, L$, the solution of

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x})\nabla U_j(\mathbf{x})) &= 0 & \mathbf{x} \in D \\ U_j(\mathbf{x}) &= 0 & \mathbf{x} \in \Gamma_D \\ a_0(\mathbf{x})\frac{\partial U_j(\mathbf{x})}{\partial \mathbf{n}} &= g_j(\mathbf{x}) & \mathbf{x} \in \Gamma_N. \end{cases} \quad (2.7)$$

Contrary to the problem with random diffusion coefficient a of the previous chapter, we will show that we have here $u = u_0 + \varepsilon u_1$, i.e. there is no term of order higher than one in ε . This is due to the linear dependence of u with respect to the uncertain input data g . The same holds for instance when the forcing term f is random, see also the next section. The weak formulation of problems (2.6) and (2.7) is given by, respectively,

$$\text{find } u_0 \in W : \quad \int_D a_0 \nabla u_0 \cdot \nabla v = \int_D f v + \int_{\Gamma_N} g_0 v \quad \forall v \in W \quad (2.8)$$

and

$$\text{find } U_j \in W : \quad \int_D a_0 \nabla U_j \cdot \nabla v = \int_{\Gamma_N} g_j v \quad \forall v \in W. \quad (2.9)$$

Notice that the problems for u_0 and the U_j , $j = 1, \dots, L$, are decoupled, that is the solution u_0 does not appear in the problem for U_j as it is the case when dealing with random diffusion coefficient, see problem (1.21). We first show the following three properties.

Proposition 2.1.1. *Let u be the weak solution of problem (2.1) and let u_0 and U_j , $j = 1, \dots, L$, be the solutions of problems (2.8) and (2.9), respectively. Then for $u_1 = \sum_{j=1}^L U_j Y_j$ we have*

1. $\mathbb{E}[u] = u_0$
2. $u = u_0 + \varepsilon u_1$
3. $\text{Var}[u] = \varepsilon^2 \sum_{j=1}^L U_j^2$.

Proof. First of all, if we take the expected value on both sides of equation (2.4) with $\mathbf{y} = \mathbf{Y}(\omega)$, we get

$$\int_D a_0 \nabla \mathbb{E}[u] \cdot \nabla v = \int_D f v + \int_{\Gamma_N} g_0 v \quad \forall v \in W$$

and thus, subtracting equation (2.8) we obtain

$$\int_D a_0 \nabla (\mathbb{E}[u] - u_0) \cdot \nabla v = 0 \quad \forall v \in W.$$

If we take then $v = \mathbb{E}[u] - u_0$, we have

$$0 \leq a_{0,\min} \|\nabla (\mathbb{E}[u] - u_0)\|_{L^2(D)}^2 \leq \|a_0^{\frac{1}{2}} \nabla (\mathbb{E}[u] - u_0)\|_{L^2(D)}^2 = 0$$

which implies $\mathbb{E}[u] = u_0$ a.e. in D . We proceed similarly for the second relation. Indeed, without writing the dependence of each function, we have for any $v \in W$ and a.s. in Ω

$$\begin{aligned} \int_D a_0 \nabla (u - (u_0 + \varepsilon u_1)) \cdot \nabla v &= \int_D a_0 \nabla u \cdot \nabla v - \int_D a_0 \nabla u_0 \cdot \nabla v - \varepsilon \int_D a_0 \nabla u_1 \cdot \nabla v \\ &= \int_{\Gamma_N} g v - \int_{\Gamma_N} g_0 v - \varepsilon \sum_{j=1}^L \int_{\Gamma_N} g_j v \\ &= 0. \end{aligned}$$

Taking then $v = u - u_0 - \varepsilon u_1 \in W$ a.s. in Ω , we can easily show that $\|u - (u_0 + \varepsilon u_1)\|_{L_p^2(\Omega; W)} = 0$ and thus $u = u_0 + \varepsilon u_1$ a.e. in D and a.s. in Ω . Finally, we directly get

$$\text{Var}[u] = \mathbb{E}[(u - \mathbb{E}[u])^2] = \mathbb{E}[\varepsilon^2 u_1^2] = \varepsilon^2 \sum_{j=1}^L U_j^2$$

using the fact that $\mathbb{E}[Y_i Y_j] = \delta_{ij}$. □

Remark 2.1.2. *Notice that we could also see that u does not contain any term of order $\mathcal{O}(\varepsilon^k)$ for any $k \geq 2$ by observing that the term u_k in the expansion of u would be the solution of the*

problem

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x})\nabla u_k(\mathbf{x},\omega)) &= 0 & \mathbf{x} \in D \\ u_k(\mathbf{x},\omega) &= 0 & \mathbf{x} \in \Gamma_D \\ a_0(\mathbf{x})\frac{\partial u_k(\mathbf{x},\omega)}{\partial \mathbf{n}} &= 0 & \mathbf{x} \in \Gamma_N, \end{cases}$$

for which $u_k = 0$ is the obvious solution.

To simplify the notation in the *a posteriori* error estimates given below, we introduce the *generalized jumps* across an edge e defined as

$$J_{e,0}(u_{0,h}) := \begin{cases} \frac{1}{2}[a_0\nabla u_{0,h} \cdot \mathbf{n}_e]_e & \text{if } e \subset D \\ g_0 - \lim_{t \rightarrow 0^+} (a_0\nabla u_{0,h} \cdot \mathbf{n}_e)(\mathbf{x} - t\mathbf{n}_e) & \text{if } e \subset \Gamma_N \\ 0 & \text{if } e \subset \Gamma_D \end{cases}$$

with $[\cdot]_{\mathbf{n}_e}$ the jump across an interior edge e defined by

$$[\varphi]_{\mathbf{n}_e}(\mathbf{x}) := \lim_{t \rightarrow 0^+} (\varphi(\mathbf{x} + t\mathbf{n}_e) - \varphi(\mathbf{x} - t\mathbf{n}_e)).$$

For $j = 1, \dots, L$, the quantity $J_{e,j}(U_{j,h})$ is defined analogously replacing $u_{0,h}$ and g_0 by $U_{j,h}$ and g_j , respectively. Moreover, we will need the following trace inequality (see for instance [109])

$$\|v\|_{L^2(\Gamma_N)} \leq C_T \|v\|_{H^1(D)} \quad \forall v \in H^1(D). \quad (2.10)$$

Error estimation for $u - u_{0,h}$

We consider the \mathbb{P}_1 finite element approximation of problem (2.8) given by

$$\text{find } u_{0,h} \in W_h : \quad \int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D f v_h + \int_{\Gamma_N} g_0 v_h \quad \forall v_h \in W_h \quad (2.11)$$

with $W_h = \{v \in C^0(\bar{D}) : v|_K \in \mathbb{P}_1 \ \forall K \in \mathcal{T}_h\} \cap W$ and \mathcal{T}_h a regular triangulation of D . We have the following *a posteriori* error estimate for the error $u - u_{0,h}$, yielding an error of order $\mathcal{O}(h^s + \varepsilon)$ with $s \in (0, 1]$ depending on the regularity of the solution.

Proposition 2.1.3. *Let u be the weak solution of problem (2.1) and let $u_{0,h}$ be the solution of problem (2.11). Then, there exists a constant $C > 0$ depending only on C_F in (2.5), C_T in (2.10) and the mesh aspect ratio such that*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{C}{a_{0,\min}} (\eta_h^2 + \eta_\varepsilon^2)^{\frac{1}{2}},$$

with

$$\begin{aligned}\eta_h^2 &:= \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} h_e \|J_{e,0}(u_{0,h})\|_{L^2(e)}^2 \\ \eta_\varepsilon^2 &:= \varepsilon^2 \sum_{j=1}^L \|g_j\|_{L^2(\Gamma_N)}^2.\end{aligned}$$

Proof. For any $v \in W$ and a.s. in Ω we have

$$\int_D a_0 \nabla(u - u_{0,h}) \cdot \nabla v = \underbrace{\int_D f v + \int_{\Gamma_N} g_0 v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v}_{=: I} + \underbrace{\int_{\Gamma_N} (g - g_0) v}_{=: II}.$$

We bound each term separately. The term I, which is the residual for $u_{0,h}$, can be bounded as follows

$$I \leq C_1 \left[\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} \|J_{e,0}(u_{0,h})\|_{L^2(e)}^2 \right]^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}$$

where C_1 depends only on the interpolation constants in (1.26) and (1.28). The second term is bounded by

$$II = \int_{\Gamma_N} (g - g_0) v \leq \|g - g_0\|_{L^2(\Gamma_N)} \|v\|_{L^2(\Gamma_N)} \leq C_2 \|g - g_0\|_{L^2(\Gamma_N)} \|\nabla v\|_{L^2(D)}, \quad C_2 = C_T \sqrt{1 + C_F^2}.$$

Combining these two bounds with the fact that a_0 is larger than $a_{0,min}$ we get

$$\begin{aligned}\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 &\leq \frac{1}{a_{0,min}} \left[C_1 \left(\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} \|J_{e,0}(u_{0,h})\|_{L^2(e)}^2 \right)^{\frac{1}{2}} \right. \\ &\quad \left. + C_2 \|g - g_0\|_{L^2(\Gamma_N)} \right]\end{aligned}$$

Taking the expected value of the square of last inequality and using the fact that $\mathbb{E}[Y_i Y_j] = \delta_{ij}$ allows us to conclude the proof. \square

Error estimation for $u - (u_{0,h} + \varepsilon u_{1,h})$

Let $U_{j,h}$ be the \mathbb{P}_1 finite element approximation of U_j which solves

$$\text{find } U_{j,h} \in W_h : \quad \int_D a_0 \nabla U_{j,h} \cdot \nabla v_h = \int_{\Gamma_N} g_j v_h \quad \forall v_h \in W_h. \quad (2.12)$$

We have the following *a posteriori* error estimate for the error $u - (u_{0,h} + \varepsilon u_{1,h})$, yielding an error of order $\mathcal{O}(h^s + \varepsilon h^s)$, $s \in (0, 1]$. In particular, there is no term of order $\mathcal{O}(\varepsilon^k)$, $k \geq 2$, and thus no *pure statistical error*.

Proposition 2.1.4. *Let u be the weak solution of problem (2.1) and let $u_{0,h}$ be the solution of*

Chapter 2. Elliptic model problems with other sources of uncertainty

problem (2.11). Moreover, let $u_{1,h} = \sum_{j=1}^L U_{j,h} Y_j$ with $U_{j,h}$ the solution of problem (2.12). Then, there exists a constant $C > 0$ depending only on C_F in (2.5), C_T in (2.10) and the mesh aspect ratio such that

$$\mathbb{E} \left[\|\nabla(u - (u_{0,h} + \varepsilon u_{1,h}))\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{C}{a_{0,min}} (\eta_h^2 + \eta_{\varepsilon h}^2)^{\frac{1}{2}},$$

with

$$\begin{aligned} \eta_h^2 &:= \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} h_e \|J_{e,0}(u_{0,h})\|_{L^2(e)}^2 \\ \eta_{\varepsilon h}^2 &:= \varepsilon^2 \sum_{j=1}^L \left[\sum_{K \in \mathcal{T}_h} h_K^2 \|\nabla \cdot (a_0 \nabla U_{j,h})\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h} h_e \|J_{e,j}(U_{j,h})\|_{L^2(e)}^2 \right]. \end{aligned}$$

Proof. The proof can easily be deduced from the relation

$$\int_D a_0 \nabla(u - (u_{0,h} + \varepsilon u_{1,h})) \cdot \nabla v = \underbrace{\int_D f v + \int_{\Gamma_N} g_0 v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v}_{=:I} + \varepsilon \underbrace{\sum_{j=1}^L \left(\int_{\Gamma_N} g_j v - \int_D a_0 \nabla U_{j,h} \cdot \nabla v \right)}_{=:II}$$

a.s. in Ω , where I and II are nothing else than the residual for $u_{0,h}$ and $u_{1,h}$, respectively. Each of these terms can be bounded in a standard way to conclude. \square

2.2 Two sources of uncertainty

We consider again the diffusion model problem but with two input data affected by uncertainty, namely the diffusion coefficient and the source term:

find $u : D \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω it holds:

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) &= f(\mathbf{x}, \omega) & \mathbf{x} \in D \\ u(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D, \end{cases} \quad (2.13)$$

where $f(\cdot, \omega) \in L^2(D)$ a.s. in Ω and a is uniformly bounded from below and above by a_{min} and a_{max} , respectively. We prescribe homogeneous Dirichlet boundary conditions for simplicity but we could easily extend the following results to other kinds of boundary conditions, including random boundary conditions as treated in the previous section. We assume that the two random inputs a and f are characterized through a finite number of random variables

$$a(\mathbf{x}, \omega) = a(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega)) \quad \text{and} \quad f(\mathbf{x}, \omega) = f(\mathbf{x}, Z_1(\omega), \dots, Z_M(\omega)).$$

More precisely, we assume an affine dependence of a and f with respect to the random

variables as follows

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L a_j(\mathbf{x}) Y_j(\omega), \quad (2.14)$$

$$f(\mathbf{x}, \omega) = f_0(\mathbf{x}) + \delta \sum_{j=1}^M f_j(\mathbf{x}) Z_j(\omega), \quad (2.15)$$

where $\{Y_j\}_{j=1}^L$ and $\{Z_j\}_{j=1}^M$ are two families of independent random variables with zero mean and $\text{Var}(Y_j) = (\sigma_j^y)^2 < \infty$ and $\text{Var}(Z_i) = (\sigma_i^z)^2 < \infty$ for $j = 1, \dots, L$ and $i = 1, \dots, M$. Moreover, we assume that $f_j \in L^2(D)$ for $j = 0, 1, \dots, M$. The two parameters ε and δ control the amount of randomness in a and f , respectively.

Remark 2.2.1. *The case where only the forcing term is affected by uncertainty can be easily deduced from the one considered here by setting $\varepsilon = 0$.*

Let $\mathbf{Y} = (Y_1, \dots, Y_L)$, $\mathbf{Z} = (Z_1, \dots, Z_M)$ and $\mathbf{R} = (\mathbf{Y}, \mathbf{Z})$. For $j = 1, \dots, L$, let Γ_j^y denote the bounded image in \mathbb{R} of Y_j and for $i = 1, \dots, M$ let Γ_i^z be the image in \mathbb{R} of Z_i . Moreover, we write ρ_j^y and ρ_i^z their probability density function. Let $\Gamma = \Gamma^y \times \Gamma^z = \Gamma_1^y \times \dots \times \Gamma_L^y \times \Gamma_1^z \times \dots \times \Gamma_M^z$. Thanks to the independence of the random variables, the joint density function $\rho : \Gamma \rightarrow \mathbb{R}^+$ of the random vector \mathbf{R} is given by $\rho(\mathbf{r}) = \rho^y(\mathbf{y})\rho^z(\mathbf{z}) = \prod_{j=1}^L \rho_j^y(y_j) \prod_{i=1}^M \rho_i^z(z_i)$ for all $\mathbf{r} = (\mathbf{y}, \mathbf{z}) \in \Gamma$ with $\mathbf{y} = (y_1, \dots, y_L) \in \Gamma^y$ and $\mathbf{z} = (z_1, \dots, z_M) \in \Gamma^z$. By definition, for any measurable function $g : \Gamma \rightarrow \mathbb{R}$, the expected value of the random variable $g(\mathbf{R})$ is $\mathbb{E}[g(\mathbf{R})] = \int_{\Gamma} g(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r}$. The *finite dimensional noise* assumption implies that the random solution u of problem (2.13) can be described by $L + M$ random variables

$$u(\mathbf{x}, \omega) = u(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega), Z_1(\omega), \dots, Z_M(\omega)).$$

Therefore, the solution u can be sought in the probability space (Ω, \mathcal{F}, P) or equivalently in $(\Gamma, B(\Gamma), \rho(\mathbf{r}) d\mathbf{r})$. The problem (2.13) can indeed be equivalently written in the following *deterministic* parametric form:

find $u : D \times \Gamma^y \times \Gamma^z \rightarrow \mathbb{R}$ such that ρ -a.e. in Γ it holds:

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}, \mathbf{z})) &= f(\mathbf{x}, \mathbf{z}) & \mathbf{x} \in D \\ u(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= 0 & \mathbf{x} \in \partial D. \end{cases} \quad (2.16)$$

The pointwise weak formulation of (2.16) reads:

find $u \in L^2_{\rho}(\Gamma; H_0^1(D))$ such that

$$\mathcal{A}(u(\cdot, \mathbf{y}, \mathbf{z}), v; \mathbf{y}) = F(v; \mathbf{z}) \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma, \quad (2.17)$$

where

$$\mathcal{A}(u(\cdot, \mathbf{y}, \mathbf{z}), v; \mathbf{y}) = \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}, \mathbf{z}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} \quad (2.18)$$

$$F(v; \mathbf{z}) = \int_D f(\mathbf{x}, \mathbf{z}) v(\mathbf{x}) d\mathbf{x}. \quad (2.19)$$

The well-posedness of problem (2.17) can be shown using Lax-Milgram's lemma. In particular, the assumptions on f_0 , f_i and Z_i , $i = 1, \dots, M$, ensure that $f \in L^2_\rho(\Gamma; L^2(D))$.

We assume small uncertainty and use a perturbation approach expanding u with respect to ε and δ as

$$\begin{aligned} u(\mathbf{x}, \mathbf{Y}(\omega), \mathbf{Z}(\omega)) &= u_0(\mathbf{x}) + \varepsilon u_1^y(\mathbf{x}, \mathbf{Y}(\omega)) + \delta u_1^z(\mathbf{x}, \mathbf{Z}(\omega)) \\ &\quad + \varepsilon^2 u_2^y(\mathbf{x}, \mathbf{Y}(\omega)) + \varepsilon \delta u_2^{yz}(\mathbf{x}, \mathbf{Y}(\omega), \mathbf{Z}(\omega)) + \delta^2 u_2^z(\mathbf{x}, \mathbf{Z}(\omega)) + \dots \end{aligned} \quad (2.20)$$

Notice that similarly to Section 2.1, there will be no term of higher order than 1 in δ , i.e. u_2^z vanishes, due to the linear dependence of u with respect to f .

The problem for u_0 is given by:

find $u_0 : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x}) \nabla u_0(\mathbf{x})) &= f_0(\mathbf{x}) & \mathbf{x} \in D \\ u_0(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D. \end{cases} \quad (2.21)$$

Writing then $u_1^y(\mathbf{x}, \mathbf{Y}(\omega)) = \sum_{j=1}^L U_j^y(\mathbf{x}) Y_j(\omega)$ and $u_1^z(\mathbf{x}, \mathbf{Z}(\omega)) = \sum_{j=1}^M U_j^z(\mathbf{x}) Z_j(\omega)$, the first order term in (2.20) is obtained by solving the following $L + M$ deterministic uncoupled problems:

find $U_j^y : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x}) \nabla U_j^y(\mathbf{x}) + a_j(\mathbf{x}) \nabla u_0(\mathbf{x})) &= 0 & \mathbf{x} \in D \\ U_j^y(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad j = 1, \dots, L \quad (2.22)$$

and

find $U_j^z : D \rightarrow \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x}) \nabla U_j^z(\mathbf{x})) &= f_j(\mathbf{x}) & \mathbf{x} \in D \\ U_j^z(\mathbf{x}) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad j = 1, \dots, M. \quad (2.23)$$

Notice that the solution u_0 of problem (2.21) is required in problem (2.22) but not in (2.23).

Error $u - u_{0,h}$

Let $u_{0,h}$ be the \mathbb{P}_1 finite element approximation of u_0 , i.e. the solution of

$$\text{find } u_{0,h} \in V_h : \int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D f_0 v_h \quad \forall v_h \in V_h, \quad (2.24)$$

where $V_h = \{v \in C^0(\bar{D}) : v|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h\} \cap V$ and \mathcal{T}_h is a regular triangulation of D . The following proposition gives an *a posteriori* error estimation of the error $u - u_{0,h}$ in the $L_P^2(\Omega; H_0^1(D))$ norm.

Proposition 2.2.2. *Let u be the weak solution of problem (2.13) and let $u_{0,h}$ be the solution of problem (2.24). There exists a constant $C > 0$ depending only the mesh aspect ratio such that*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{3}}{a_{\min}} [C\eta_h^2 + \eta_\varepsilon^2 + C_P^2 \eta_\delta^2]^{\frac{1}{2}}, \quad (2.25)$$

where C_P is the Poincaré constant and

$$\eta_h^2 := \sum_{K \in \mathcal{T}_h} \eta_K^2 \quad \text{with} \quad \eta_K^2 = h_K^2 \|f_0 + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \subset \partial K} h_e \left\| \frac{1}{2} [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \quad (2.26)$$

$$\eta_\varepsilon^2 := \varepsilon^2 \sum_{j=1}^L (\sigma_j^y)^2 \|a_j \nabla u_{0,h}\|_{L^2(D)}^2 \quad (2.27)$$

$$\eta_\delta^2 := \delta^2 \sum_{j=1}^M (\sigma_j^z)^2 \|f_j\|_{L^2(D)}^2. \quad (2.28)$$

Proof. For any $v \in H_0^1(D)$ and a.s. in Ω we have

$$\int_D a \nabla(u - u_{0,h}) \cdot \nabla v = \underbrace{\int_D f_0 v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v}_{=:I} + \underbrace{\int_D (f - f_0) v - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v}_{=:II}. \quad (2.29)$$

The term I is nothing else but the residual for $u_{0,h}$ and we have

$$I \leq \left(C \sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}, \quad \eta_K^2 = h_K^2 \|f + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \subset \partial K} h_e \left\| \frac{1}{2} [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \quad (2.30)$$

with C an interpolation constant which depends only on the interpolation constants in (1.26) and (1.28). For the second term, thanks to Cauchy-Schwarz and Poincaré inequalities we have the bound

$$II \leq (C_P \|f - f_0\|_{L^2(D)} + \|(a - a_0) \nabla u_{0,h}\|_{L^2(D)}) \|\nabla v\|_{L^2(D)}$$

where C_P denotes the constant in Poincaré's inequality. Using the lower bound on a , we thus

obtain

$$\|\nabla(u - u_{0,h})\|_{L^2(D)} \leq \frac{1}{a_{\min}} \left[\left(C \sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{\frac{1}{2}} + C_P \|f - f_0\|_{L^2(D)} + \|(a - a_0)\nabla u_{0,h}\|_{L^2(D)} \right].$$

The result follows from taking the expected value on the square of the last inequality. \square

As we will see in the numerical results, the loss due to the use of the Poincaré inequality for the source term is dependent on the input data. In other words, the efficiency of the estimator η_δ in (2.28), for which the Poincaré inequality has been used, will be different from one case to another. A way to skirt this drawback is to replace η_δ by an *implicit* estimator obtained by computing (approximately) the dual norm of a residual to be defined. The price to pay is that the computation of this estimator, given in the following proposition, requires the resolution of M additional (Poisson) problems.

Proposition 2.2.3. *Let u be the weak solution of problem (2.13) and let $u_{0,h}$ be the solution of problem (2.24). There exists a constant $C > 0$ depending only on the mesh aspect ratio such that*

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{3}}{a_{\min}} [C\eta_h^2 + \eta_\varepsilon^2 + \hat{\eta}_\delta^2]^{\frac{1}{2}} + h.o.t., \quad (2.31)$$

where η_h and η_ε are as in (2.26) and (2.27), respectively, and

$$\hat{\eta}_\delta^2 = \delta^2 \sum_{j=1}^M (\sigma_j^z)^2 \|\nabla W_{j,h}\|_{L^2(D)}^2 \quad (2.32)$$

with $W_{j,h} \in V_h$ the solution of

$$\int_D \nabla W_{j,h} \cdot \nabla v_h = \int_D f_j v_h \quad \forall v_h \in V_h.$$

Proof. The only difference with respect to the proof of Proposition 2.2.2 is how we bound the term II of (2.29) due to the uncertainty in the input data, more precisely the part due to the forcing term. Let us introduce for any $\mathbf{z} \in \Gamma^z$ the operator $R(\cdot; \mathbf{z}) : H_0^1(D) \rightarrow \mathbb{R}$ defined by

$$R(v; \mathbf{z}) := \int_D (f(\cdot; \mathbf{z}) - f_0) v = \delta \sum_{j=1}^M z_j \int_D f_j v.$$

The dual norm of R is then given by $\|R(\cdot; \mathbf{z})\|_{H^{-1}(D)} = \|\nabla w(\cdot; \mathbf{z})\|_{L^2(D)}$ with w the Riesz representant of R , i.e. $w(\cdot; \mathbf{z}) \in H_0^1(D)$ is such that $\int_D \nabla w \cdot \nabla v = R(v; \mathbf{z})$ for all $v \in H_0^1(D)$ and ρ^z -a.e. in Γ^z . We can write $w = w(\mathbf{x}, \mathbf{Z}(\omega)) = \delta \sum_{j=1}^M W_j(\mathbf{x}) Z_j(\omega)$ with $W_j \in H_0^1(D)$ the solution of

$$\int_D \nabla W_j \cdot \nabla v = \int_D f_j v \quad \forall v \in H_0^1(D) \quad (2.33)$$

from which we deduce

$$\mathbb{E} \left[\|R\|_{H^{-1}(D)}^2 \right] = \delta^2 \sum_{j=1}^M (\sigma_j^z)^2 \|\nabla W_j\|_{L^2(D)}^2.$$

Since the solution of (2.33) can not be computed exactly, we can replace it by its finite element approximation $W_{j,h} \in V_h$. Doing so introduce an error of higher order, the proof being similar to that of Proposition 1.B.1. \square

We mention that the computational cost to get the error estimator $\hat{\eta}_\delta$ is the same as that needed to get the finite element approximation $u_{1,h}^z$ of the term u_1^z in the expansion (2.20). Since the solution u depends linearly on the input f , there is no term of order δ^2 and it would thus be better to simply add the term $\delta u_{1,h}^z$ to $u_{0,h}$. The quantification of the error in $\mathcal{O}(\delta h)$ so introduced is made precisely in Proposition 2.2.5, see the term $\eta_{\delta h}$. As mentioned in Chapter 1, the computational cost might be reduced introducing auxiliary local problems defined on an element or a small subdomain.

Remark 2.2.4. Notice that we could use the same procedure as used in Proposition 2.2.3 for the whole term II , and not only the part due to f , by considering the residual defined for all $v \in H_0^1(D)$ and $(\mathbf{y}, \mathbf{z}) \in \Gamma$ by

$$R(v; \mathbf{y}, \mathbf{z}) = \int_D (f(\cdot, \mathbf{z}) - f_0) v - \int_D (a(\cdot, \mathbf{y}) - a_0) \nabla u_{0,h} \cdot \nabla v.$$

The dual norm of R is then given by $\|R(\cdot; \mathbf{y}, \mathbf{z})\|_{H^{-1}(D)} = \|\nabla w(\cdot; \mathbf{y}, \mathbf{z})\|_{L^2(D)}$ where $w(\cdot; \mathbf{y}, \mathbf{z}) \in H_0^1(D)$ ρ -a.e. in Γ writes

$$w(\mathbf{x}; \mathbf{Y}(\omega), \mathbf{Z}(\omega)) = \varepsilon \sum_{j=1}^L W_j^y(\mathbf{x}) Y_j(\omega) + \delta \sum_{j=1}^M W_j^z(\mathbf{x}) Z_j(\omega)$$

with W_j^y and W_j^z the solutions of

$$\int_D \nabla W_j^y \cdot \nabla v = - \int_D a_j \nabla u_{0,h} \cdot \nabla v \quad \forall v \in H_0^1(D)$$

and

$$\int_D \nabla W_j^z \cdot \nabla v = \int_D f_j v \quad \forall v \in H_0^1(D),$$

respectively. Writing $W_{j,h}^y$ and $W_{j,h}^z$ the finite element approximations of W_j^y and W_j^z , respectively, the error estimate reads then

$$\mathbb{E} \left[\|\nabla(u - u_{0,h})\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{3}}{a_{\min}} [C\eta_h^2 + \hat{\eta}_\varepsilon^2 + \hat{\eta}_\delta^2]^{\frac{1}{2}} + h.o.t., \quad (2.34)$$

with η_h defined in (2.26) and

$$\hat{\eta}_\varepsilon^2 := \varepsilon^2 \sum_{j=1}^L (\sigma_j^y)^2 \|\nabla W_{j,h}^y\|_{L^2(D)}^2 \quad \text{and} \quad \hat{\eta}_\delta^2 = \delta^2 \sum_{j=1}^M (\sigma_j^z)^2 \|\nabla W_{j,h}^z\|_{L^2(D)}^2. \quad (2.35)$$

Error $u - (u_{0,h} + \varepsilon u_{1,h}^y + \delta u_{1,h}^z)$

Let us write $u_h^1 = u_{0,h} + \varepsilon u_{1,h}^y + \delta u_{1,h}^z$, where $u_{1,h}^y = \sum_{j=1}^L U_{j,h}^y Y_j$, $u_{1,h}^z = \sum_{i=1}^M U_{i,h}^z Z_i$ and, for $j = 1, \dots, L$ and $i = 1, \dots, M$, $U_{j,h}^y$ and $U_{i,h}^z$ are the solutions of respectively

$$\int_D (a_0 \nabla U_{j,h}^y + a_j \nabla u_{0,h}) \cdot \nabla v_h = 0 \quad \forall v_h \in V_h \quad (2.36)$$

and

$$\int_D a_0 \nabla U_{i,h}^z \cdot \nabla v_h = \int_D f_i v_h \quad \forall v_h \in V_h. \quad (2.37)$$

To simplify the notation, we write $w_{j,h} = a_0 \nabla U_{j,h}^y + a_j \nabla u_{0,h}$. The following proposition gives an *a posteriori* error estimation of the error $u - u_h^1$ in the $L_p^2(\Omega; H_0^1(D))$ norm. Notice that in particular, there is no term of order δ^2 . Indeed, we deduce from Proposition 2.2.5 that $\|u - u_h^1\|_{L_p^2(\Omega; H_0^1(D))} = \mathcal{O}(h + h(\varepsilon + \delta) + \varepsilon^2 + \varepsilon\delta)$ if u is regular enough in the physical space.

Proposition 2.2.5. *Let u be the weak solution of problem (2.13) and let $u_{0,h}$, $U_{j,h}^y$, $j = 1, \dots, L$ and $U_{i,h}^z$, $i = 1, \dots, M$, be the solutions of problems (2.24), (2.36) and (2.37), respectively. There exist constants $C_1, C_2, C_3 > 0$ depending only on the mesh aspect ratio such that*

$$\mathbb{E} \left[\|\nabla(u - u_h^1)\|_{L^2(D)}^2 \right]^{\frac{1}{2}} \leq \frac{2}{a_{\min}} [C_1 \eta_h^2 + C_2 \eta_{\varepsilon h}^2 + C_3 \eta_{\delta h}^2 + 2\eta_{\varepsilon \delta}^2]^{\frac{1}{2}}, \quad (2.38)$$

where

$$\begin{aligned} \eta_h^2 &= \sum_{K \in \mathcal{T}_h} \eta_K^2 \quad \text{with} \quad \eta_K^2 = h_K^2 \|f_0 + \nabla \cdot (a_0 \nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e \subset \partial K} h_e \left\| \frac{1}{2} [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \\ \eta_{\varepsilon h}^2 &= \varepsilon^2 \sum_{K \in \mathcal{T}_h} \sum_{j=1}^L (\sigma_j^y)^2 \theta_{K,j}^2 \quad \text{with} \quad \theta_{K,j}^2 = h_K^2 \|\nabla \cdot w_{j,h}\|_{L^2(K)}^2 + \sum_{e \subset \partial K} h_e \left\| \frac{1}{2} [w_{j,h} \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \\ \eta_{\delta h}^2 &= \delta^2 \sum_{K \in \mathcal{T}_h} \sum_{j=1}^M (\sigma_j^z)^2 \vartheta_{K,j}^2 \quad \text{with} \quad \vartheta_{K,j}^2 = h_K^2 \|f_j + \nabla \cdot (a_0 \nabla U_{j,h}^z)\|_{L^2(K)}^2 + \sum_{e \subset \partial K} h_e \left\| \frac{1}{2} [a_0 \nabla U_{j,h}^z \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \\ \eta_{\varepsilon \delta}^2 &= \varepsilon^4 \left(\int_D \sum_{i=1}^L a_i^2 |\nabla U_{i,h}^y|^2 \mathbb{E}[Y_i^4] + \int_D \sum_{\substack{i,j=1 \\ i \neq j}}^L (\sigma_i^y \sigma_j^y)^2 [a_i^2 |\nabla U_{j,h}^y|^2 + 2a_i a_j \nabla U_{i,h}^y \cdot \nabla U_{j,h}^y] \right) \\ &\quad + (\varepsilon \delta)^2 \sum_{j=1}^L \sum_{i=1}^M (\sigma_j^y \sigma_i^z)^2 \|a_j \nabla U_{i,h}^z\|_{L^2(D)}^2. \end{aligned}$$

Proof. The proof can be easily obtained from the relation

$$\int_D a \nabla(u - u_h^1) \cdot \nabla v = \text{I} + \text{II} + \text{III} + \text{IV} \quad \forall v \in H_0^1(D), \text{ a.s. in } \Omega$$

with

$$\begin{aligned} \text{I} &= \int_D f_0 v - \int_D a_0 \nabla u_{0,h} \cdot \nabla v \\ \text{II} &= - \int_D (a - a_0) \nabla u_{0,h} \cdot \nabla v - \varepsilon \int_D a_0 \nabla u_{1,h}^y \cdot \nabla v \\ \text{III} &= \int_D (f - f_0) v - \delta \int_D a_0 \nabla u_{1,h}^z \cdot \nabla v \\ \text{IV} &= -\varepsilon \int_D (a - a_0) \nabla u_{1,h}^y \cdot \nabla v - \delta \int_D (a - a_0) \nabla u_{1,h}^z \cdot \nabla v, \end{aligned}$$

bounding then each term separately. \square

2.3 Numerical results

We consider one-dimensional examples with $D = (0, 1)$. In the results below, the true error is computed with the standard Monte Carlo method with a sample size of $K = 10000$ and a reference solution computed on a uniform partition with mesh size $h_{ref} = 2^{-12}$.

Random forcing term

We consider first the case where only the forcing term is random, that is we set $\varepsilon = 0$ in (2.14). As mentioned above, the efficiency of the stochastic estimator η_δ in (2.28) depends on the input data, due to the use of the Poincaré inequality for the forcing term. To observe this behaviour, we consider the following two cases

$$f(x, \omega) = 1 + \delta \sum_{j=1}^M f_j(x) Z_j(\omega), \quad f_j(x) = \frac{\sin(2\pi j x)}{j} \quad (2.39)$$

and

$$f(x, \omega) = 1 + \delta \sum_{j=1}^M f_j(x) Z_j(\omega), \quad f_j(x) = 0.5 j^{-\frac{1}{2}} e^{-50 j (x-0.5)^2} \quad (2.40)$$

where Z_j , $j = 1, \dots, M$, are uniform random variables in $[-\sqrt{3}, \sqrt{3}]$.

The plot of several realizations of the forcing term for the case (2.39) with $M = 6$ and $M = 50$ is given in Figures 2.1 and 2.2, respectively, where the corresponding solution is also depicted. The forcing term contains much more high oscillating features with $M = 50$ than in the case $M = 6$. The difference between the two cases for the corresponding solutions is not noticeable, but is indeed present.

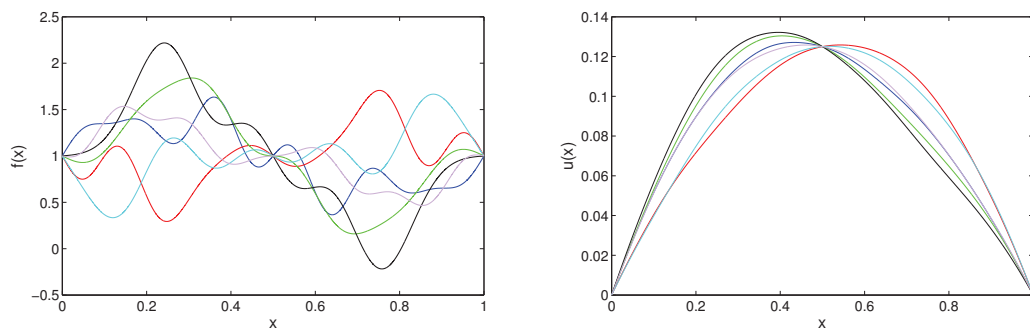


Figure 2.1: Six realizations of the random forcing term f given in (2.39) with $\delta = 0.5$ and $M = 6$ (left) and the corresponding solution (right).

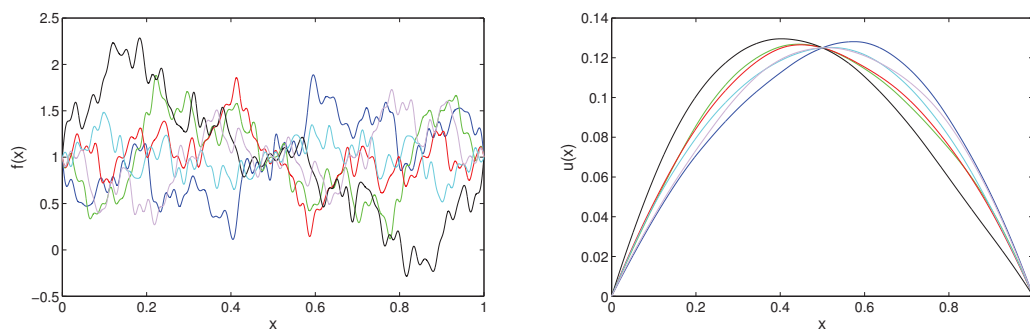


Figure 2.2: Six realizations of the random forcing term f given in (2.39) with $\delta = 0.5$ and $M = 50$ (left) and the corresponding solution (right).

Recall that we have set $\varepsilon = 0$ here, namely only the forcing term is affected by uncertainty, and thus $\eta_\varepsilon = \hat{\eta}_\varepsilon = 0$. We give in Table 2.1 the error $\|u - u_{0,h}\|_{L^2_p(\Omega; H^1_0(D))}$ and the estimators $\eta = (\eta_h^2 + \eta_\delta^2)^{\frac{1}{2}}$ and $\hat{\eta} = (C_{H^1_0}^2 \eta_h^2 + \hat{\eta}_\delta^2)^{\frac{1}{2}}$ with $C_{H^1_0} = 1/3.46$ for the first case (2.39), where η_h , η_δ and $\hat{\eta}_\delta$ are given in (2.26), (2.28) and (2.32), respectively.

	δ	error	η_δ	η	η/error	$\hat{\eta}_\delta$	$\hat{\eta}$	$\hat{\eta}/\text{error}$
$M = 6$	2^0	1.1692e-1	8.6354e-1	8.6357e-1	7.3859	1.1700e-1	1.1702e-1	1.0008
	2^{-2}	2.9361e-2	2.1588e-1	2.1603e-1	7.3575	2.9250e-2	2.9337e-2	0.9993
	2^{-4}	7.6029e-3	5.3971e-2	5.4534e-2	7.1727	7.3124e-3	7.6531e-3	1.0066
	2^{-6}	2.9040e-3	1.3493e-2	1.5591e-2	5.3689	1.8281e-3	2.9052e-3	1.0004
	2^{-8}	2.3004e-3	3.3732e-3	8.5096e-3	3.6992	4.5703e-4	2.3037e-3	1.0015

	δ	error	η_δ	η	η/error	$\hat{\eta}_\delta$	$\hat{\eta}$	$\hat{\eta}/\text{error}$
$M = 50$	2^0	1.1745e-1	9.0142e-1	9.0146e-1	7.6750	1.1706e-1	1.1708e-1	0.9969
	2^{-2}	2.9515e-2	2.2536e-1	2.2549e-1	7.6400	2.9266e-2	2.9353e-2	0.9945
	2^{-4}	7.6573e-3	5.6339e-2	5.6878e-2	7.4280	7.3164e-3	7.6569e-3	0.9999
	2^{-6}	2.8939e-3	1.4085e-2	1.6106e-2	5.5656	1.8291e-3	2.9058e-3	1.0041
	2^{-8}	2.2996e-3	3.5212e-3	8.5694e-3	3.7264	4.5728e-4	2.3038e-3	1.0018

Table 2.1: Efficiency of the two error estimator η and $\hat{\eta}$ for the case (2.39) with $h = 2^{-7}$ ($\eta_h = 7.8125e-3$).

We see that similar results are obtained for the two cases $M = 6$ and $M = 50$. Moreover, the efficiency of the error estimator η varies between 3.7 and 7.7. More precisely, we recover the value of $C_{H^1_0}$ in a physical space error dominant regime while it is about 7.7 when the stochastic error is dominant. The second error estimator $\hat{\eta}$, obtained by taking into account the constant $C_{H^1_0}$ for η_h and by computing M additional Poisson problems (see Proposition 2.2.3), yields an effectivity index close to 1. The results for the second case (2.40), see Figure 2.3 for a plot of some realizations for f and the corresponding solutions, are given in Table 2.2.

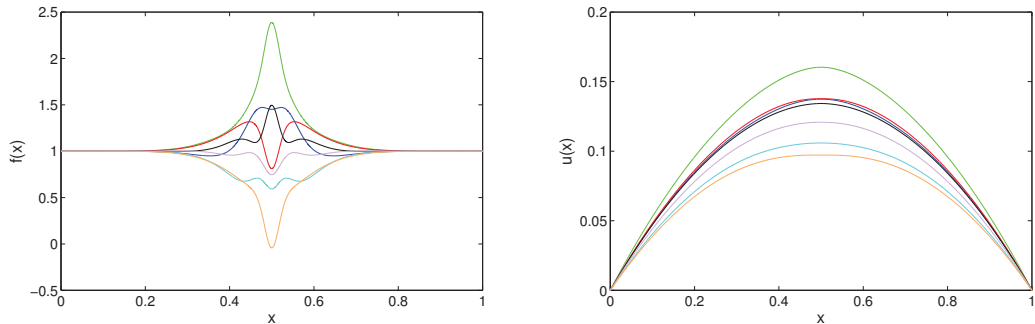


Figure 2.3: Seven realizations of the random forcing term f given in (2.40) with $\delta = 0.5$ and $M = 50$ (left) and the corresponding solution (right).

In this case, the effectivity index of the error estimator η is about 4.5 when the stochastic error

Chapter 2. Elliptic model problems with other sources of uncertainty

	δ	error	η_δ	η	η/error	$\hat{\eta}_\delta$	$\hat{\eta}$	$\hat{\eta}/\text{error}$
$M = 50$	2^0	7.2668e-2	3.2138e-1	3.2148e-1	4.4239	7.2070e-2	7.2106e-2	0.9923
	2^{-2}	1.8098e-2	8.0346e-2	8.0725e-2	4.4605	1.8018e-2	1.8159e-2	1.0034
	2^{-4}	5.0669e-3	2.0086e-2	2.1552e-2	4.2536	4.5044e-3	5.0386e-3	0.9944
	2^{-6}	2.5199e-3	5.0216e-3	9.2872e-3	3.6856	1.1261e-3	2.5232e-3	1.0013
	2^{-8}	2.2718e-3	1.2554e-3	7.9127e-3	3.4830	2.8152e-4	2.2754e-3	1.0016

Table 2.2: Efficiency of the two error estimator η and $\hat{\eta}$ for the case (2.40) with $h = 2^{-7}$ ($\eta_h = 7.8125e-3$).

is dominant, to be compared to about 7.7 for the first example. This highlight the dependence of the efficiency of η with respect to the input data, due to the different loss when using the Poincaré inequality. On the contrary, the second error estimator $\hat{\eta}$ is also very close to 1 for this second example.

Random forcing term and diffusion coefficient

Let us now consider the case of two random inputs with

$$a(x, \omega) = 1 + \varepsilon \sum_{j=1}^{50} a_j(x) Y_j(\omega), \quad a_j(x) = \frac{\sin(2\pi j x)}{(\pi j)^2}, \quad Y_j \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}] \quad (2.41)$$

and

$$f(x, \omega) = 1 + \delta \sum_{j=1}^{50} f_j(x) Z_j(\omega), \quad f_j(x) = 0.5 j^{-\frac{1}{2}} e^{-50 j(x-0.5)^2}, \quad Z_j \sim \mathcal{N}(0, 1).$$

Remark 2.3.1. We mention that the choice of the a_j in (2.41) is the one for which we obtained the largest effectivity index for the stochastic error estimator η_ε , namely the ratio of η_ε over the error is about 1.8 in the pure stochastic error case (with $\delta = 0$). It is still an open question, at least to us, to show if there are cases for which we get a larger constant, i.e. for which the loss due to the use of Cauchy-Schwarz inequality in

$$\int_D (a - a_0) \nabla u_{0,h} \cdot \nabla (u - u_{0,h}) \leq \| (a - a_0) \nabla u_{0,h} \|_{L^2(D)} \| \nabla (u - u_{0,h}) \|_{L^2(D)}$$

is bigger.

We give in Tables 2.3 and 2.4 the results obtained for the cases $h = 2^{-5}$ and $h = 2^{-7}$, respectively. We report the error $\|u - u_{0,h}\|_{L^2_p(\Omega; H^1_0(D))}$, the estimators η_h , η_ε and η_δ defined in (2.26), (2.27) and (2.28), respectively, and the effectivity index of the full estimator $\eta = (\eta_h^2 + \eta_\varepsilon^2 + \eta_\delta^2)^{\frac{1}{2}}$. We also give the efficiency of the *implicit* estimator $\hat{\eta} = (C_{H^1_0}^2 \eta_h^2 + \hat{\eta}_\varepsilon^2 + \hat{\eta}_\delta^2)^{\frac{1}{2}}$ with $\hat{\eta}_\varepsilon$ and $\hat{\eta}_\delta$ defined in (2.35) and $C_{H^1_0} = 1/3.46$.

From the results of Tables 2.3 and 2.4, we see that the efficiency of the full error estimator η is

2.3. Numerical results

ε	δ	error	η_ε	η_δ	η/error	$\hat{\eta}_\varepsilon$	$\hat{\eta}_\delta$	$\hat{\eta}/\text{error}$
2^0	2^0	7.3783e-2	1.9954e-2	3.2249e-1	4.3995	1.1331e-2	7.2013e-2	0.9956
2^{-2}	2^0	7.3191e-2	4.9885e-3	3.2249e-1	4.4272	2.8328e-3	7.2013e-2	0.9924
2^{-4}	2^0	7.2246e-2	1.2471e-3	3.2249e-1	4.4847	7.0821e-4	7.2013e-2	1.0046
2^{-6}	2^0	7.2718e-2	3.1178e-4	3.2249e-1	4.4555	1.7705e-4	7.2013e-2	0.9981
2^0	2^{-2}	2.3233e-2	1.9954e-2	8.0622e-2	3.8195	1.1331e-2	1.8003e-2	0.9947
2^{-2}	2^{-2}	2.0159e-2	4.9885e-3	8.0622e-2	4.2964	2.8328e-3	1.8003e-2	1.0090
2^{-4}	2^{-2}	2.0186e-2	1.2471e-3	8.0622e-2	4.2840	7.0821e-4	1.8003e-2	0.9984
2^{-6}	2^{-2}	2.0131e-2	3.1178e-4	8.0622e-2	4.2952	1.7705e-4	1.8003e-2	1.0006
2^0	2^{-4}	1.5453e-2	1.9954e-2	2.0155e-2	2.7309	1.1331e-2	4.5008e-3	0.9819
2^{-2}	2^{-4}	1.0487e-2	4.9885e-3	2.0155e-2	3.5776	2.8328e-3	4.5008e-3	0.9994
2^{-4}	2^{-4}	1.0114e-2	1.2471e-3	2.0155e-2	3.6789	7.0821e-4	4.5008e-3	1.0002
2^{-6}	2^{-4}	1.0068e-2	3.1178e-4	2.0155e-2	3.6937	1.7705e-4	4.5008e-3	1.0025
2^0	2^{-6}	1.4804e-2	1.9954e-2	5.0388e-3	2.5276	1.1331e-2	1.1252e-3	0.9818
2^{-2}	2^{-6}	9.5369e-3	4.9885e-3	5.0388e-3	3.3600	2.8328e-3	1.1252e-3	0.9995
2^{-4}	2^{-6}	9.1184e-3	1.2471e-3	5.0388e-3	3.4741	7.0821e-4	1.1252e-3	1.0012
2^{-6}	2^{-6}	9.0934e-3	3.1178e-4	5.0388e-3	3.4811	1.7705e-4	1.1252e-3	1.0011

Table 2.3: for $h = 2^{-5}$ ($\eta_h = 3.125e-2$)

ε	δ	error	η_ε	η_δ	η/error	$\hat{\eta}_\varepsilon$	$\hat{\eta}_\delta$	$\hat{\eta}/\text{error}$
2^0	2^0	7.3022e-2	1.9923e-2	3.2138e-1	4.4109	1.1490e-2	7.2070e-2	0.9999
2^{-2}	2^0	7.2376e-2	4.9806e-3	3.2138e-1	4.4423	2.8724e-3	7.2070e-2	0.9971
2^{-4}	2^0	7.2361e-2	1.2452e-3	3.2138e-1	4.4428	7.1811e-4	7.2070e-2	0.9965
2^{-6}	2^0	7.1792e-2	3.1129e-4	3.2138e-1	4.4779	1.7953e-4	7.2070e-2	1.0044
2^0	2^{-2}	2.1710e-2	1.9923e-2	8.0346e-2	3.8299	1.1490e-2	1.8018e-2	0.9898
2^{-2}	2^{-2}	1.8452e-2	4.9806e-3	8.0346e-2	4.3832	2.8724e-3	1.8018e-2	0.9963
2^{-4}	2^{-2}	1.8183e-2	1.2452e-3	8.0346e-2	4.4401	7.1811e-4	1.8018e-2	0.9994
2^{-6}	2^{-2}	1.7873e-2	3.1129e-4	8.0346e-2	4.5165	1.7953e-4	1.8018e-2	1.0160
2^0	2^{-4}	1.2685e-2	1.9923e-2	2.0086e-2	2.3138	1.1490e-2	4.5044e-3	0.9890
2^{-2}	2^{-4}	5.7768e-3	4.9806e-3	2.0086e-2	3.8291	2.8724e-3	4.5044e-3	1.0040
2^{-4}	2^{-4}	5.0541e-3	1.2452e-3	2.0086e-2	4.2715	7.1811e-4	4.5044e-3	1.0070
2^{-6}	2^{-4}	5.0988e-3	3.1129e-4	2.0086e-2	4.2274	1.7953e-4	4.5044e-3	0.9888
2^0	2^{-6}	1.1897e-2	1.9923e-2	5.0216e-3	1.8476	1.1490e-2	1.1261e-3	0.9888
2^{-2}	2^{-6}	3.8361e-3	4.9806e-3	5.0216e-3	2.7471	2.8724e-3	1.1261e-3	0.9966
2^{-4}	2^{-6}	2.6217e-3	1.2452e-3	5.0216e-3	3.5742	7.1811e-4	1.1261e-3	1.0007
2^{-6}	2^{-6}	2.5186e-3	3.1129e-4	5.0216e-3	3.6895	1.7953e-4	1.1261e-3	1.0043

Table 2.4: for $h = 2^{-7}$ ($\eta_h = 7.8125e-3$)

comprised between the efficiency of each of its parts η_h , η_ε and η_δ , depending on which is the predominant source of error. For instance, the effectivity index tends to the value 3.46 when the FE error is dominant, see e.g. Table 2.3 with $\delta = 2^{-6}$ and $\varepsilon = 2^{-4}$ or 2^{-6} , while it is about 4.5 in a δ -error dominant regime as in the similar case (2.40) considered above. Finally, we see that if the error due to the uncertainty in the diffusion coefficient a is largest, the effectivity index tends to the value 1.8 indicated in Remark 2.3.1. In all cases, the *implicit* error estimator $\hat{\eta}$ has an effectivity index close to 1, but more work is required to compute it.

Conclusions

In this chapter, we have extended the results we obtained in Chapter 1 for the linear model problem to include other sources of uncertainty. More precisely, we have considered first the case of Neumann random boundary conditions and then the combination of two random input data, namely the diffusion coefficient and the forcing term. For the latter case, two different sets of random variables have been used to characterize the data affected by uncertainty. We have shown that when the random solution depends linearly on the random input, as it is the case for Neumann boundary conditions or the source term, then the solution is fully described by the first two terms in the expansion, the remaining terms being zero. Moreover, we have seen that when the Poincaré inequality is required in the estimation, the efficiency of the error estimator might change when modifying the input data, even though it has the optimal convergence rate. The same behaviour was observed when considering the error $u - u_{0,h}$ in the $L^2(D)$ norm in Chapter 1. As a remedy to the sensitivity of the error estimator to the input data, we have proposed a second error estimator, see Proposition 2.2.3. It is obtained by solving additional (Poisson) problems, as many as the number of random variables used to characterized the uncertainty in the data. However, we can use the same spatial mesh than the one for $u_{0,h}$ to solve these problems approximately.

3 PDEs in random domains

In this chapter, we consider nonlinear PDEs defined on random domains. The first part consists of the analysis of a 1D problem, namely the viscous Burgers' equation to be solved on an interval of random length. This equation, first introduced by Bateman in [19] and then used by Burgers in [34] for modelling turbulence, can be seen as a simplification of the Navier-Stokes equations to the one-dimensional case. In the second part, whose material is mainly taken from the submitted paper [75], we consider the more involved incompressible Navier-Stokes equations in random domains. We restrict ourselves here to the stationary formulation of these equations.

For both problems, we use the so-called domain mapping method [125]: we introduce a random mapping that transforms the deterministic PDEs defined on a random domain into PDEs on a fixed reference domain with random coefficients. For simplicity, we assume that the uncertainty in the system is only due to the random domain, but the analysis can be straightforwardly extended to include other sources of randomness.

Introduction

Several approaches have been developed to perform analysis and numerical approximation of PDEs in random domains, such as the fictitious domain method [40], the perturbation method based on shape calculus [77] and the domain mapping method initially proposed by [125] and also used for instance in [39, 43, 76]. In the first approach, the PDEs are extended to a fixed reference domain, the so-called fictitious domain, which contains all the random domains. The original boundary condition is then imposed through a Lagrange multiplier yielding a saddle-point problem to be solved in the fictitious domain. In the perturbation method, which is suitable for small perturbations only, the solution is represented using a shape Taylor expansion with respect to the (random) perturbation field of the boundary of the domain. Finally, the domain mapping approach, which is the one considered in this work, transforms the deterministic PDEs defined on a random domain into PDEs on a fixed reference domain with random coefficients via a random mapping. We give in Figure 3.1 an

illustration of the mapping for a given ω between the physical domain and the reference one, supplemented with some notation.

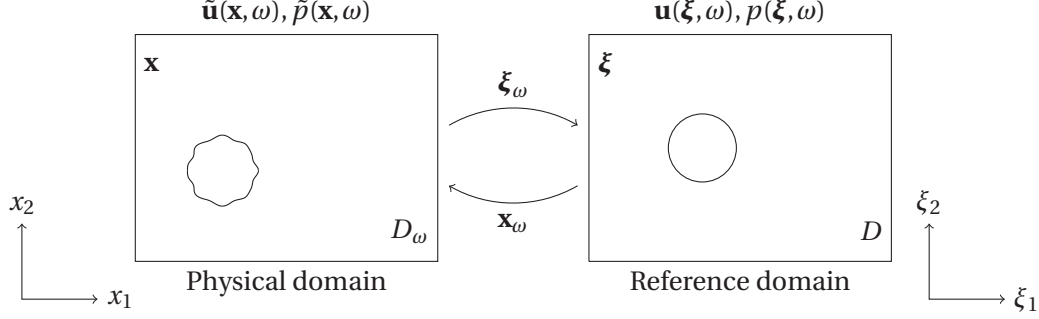


Figure 3.1: Illustration and notation for the domain mapping approach.

Contrary to the method based on shape derivatives, our approach requires the construction of a random mapping defined in the whole domain consistent with the random perturbation of the boundary. If the random mapping is not given analytically, it can be obtained by solving appropriate equations, e.g. Laplace equation as it is done in [125]. The domain mapping method prevents the need of remeshing and can make use of the well-developed theory for PDEs on deterministic domains with random coefficients. Numerical approximation of the solution on the fixed reference domain can indeed be obtained through any of the well-known techniques, such as Monte-Carlo methods [63] and their generalizations as quasi-Monte Carlo [38, 54, 70] and multi-level Monte-Carlo [17, 52, 68, 79], or the stochastic spectral methods comprising the stochastic Galerkin [10, 11, 21, 64, 67] and the stochastic collocation [7, 97, 124] methods.

The (weak) formulation on the reference domain can be obtained using two strategies, as illustrated in Figure 3.2. In general, the two strategies are not equivalent. They yield the same result only in particular cases, for instance if the Jacobian of the mapping does not depend on the physical space variable. In this work, we will use the first strategy s_1 , that is the formulation on the reference domain is obtained performing the change of variables on the weak formulation of the problem on the random domain. We refer for instance to [36] for a version where the second strategy s_2 is used.

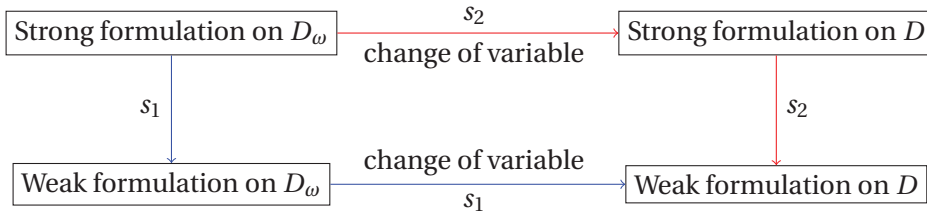


Figure 3.2: Two strategies s_1 and s_2 for the (strong) formulation on the reference domain.

For the stochastic space approximation, we proceed as in the previous chapters and use a perturbation approach [82] to expand the exact random solution with respect to a parameter ε that controls the level of uncertainty in the problem. This approach yields uncoupled deterministic problems for each term in the expansion, which can be solved using for instance the finite element (FE) method. The main goal here is to perform an *a posteriori* error analysis for the error between the exact random solution and the finite element approximation of the first term in the expansion, that is the solution corresponding to the case $\varepsilon = 0$. The error estimators we obtain are made of two parts, namely one part due to the physical space discretization and another one due to the uncertainty. Their computation requires only the FE approximation of the solution of the problem for $\varepsilon = 0$ and the Jacobian matrix of the mapping between the reference domain and the physical random domain. These estimators can be used for instance to adaptively determine a mesh that yields a numerical accuracy comparable with the model uncertainty. Notice that the error estimates we get here using the domain mapping method combined with a perturbation technique are defined for any fixed ε . The only restriction is that ε is sufficiently small for the problem to be well-posed. The more common perturbation method is to use shape calculus [77], thus avoiding to recast the equations in a reference domain. However, the derivation of *a posteriori* error estimates for a fixed value of ε is, in our opinion, not obvious in this context and, to the best of our knowledge, it is still an open question.

We mention that the formulation we obtained in Section 3.2.2 for the Navier-Stokes equations on the reference domain is similar to the one obtained for instance in [71] where a fluid-structure interaction problem is considered or in [91, 108] where the Navier-Stokes equations in parametrized domains are solved approximately using the Reduced Basis Method.

3.1 Steady-state viscous Burgers' equation in random intervals

To start with, we consider a 1D problem on a random domain, namely the (nonlinear) steady-state viscous Burgers' equation. This equation can be viewed as a simplification of the Navier-Stokes equations in the one-dimensional case. We consider a physical domain with uncertain geometry, which reduces here to an interval of random length. We study first the deterministic case, considering the Burgers' equation on a fixed domain, say $[0, 1]$.

3.1.1 Deterministic case

We consider the following nonlinear deterministic problem with mixed Neumann-Dirichlet homogeneous boundary conditions:

find $u : (0, 1) \rightarrow \mathbb{R}$ such that

$$\begin{cases} -au'' + buu' &= f & \text{in } (0, 1) \\ u(0) &= 0 \\ u'(1) &= 0 \end{cases} \quad (3.1)$$

where a and b are positive constants and $f \in L^2(0, 1)$. It can be written in conservation form as

$$-au'' + \frac{b}{2}(u^2)' = f \quad \text{in } (0, 1).$$

Let $V = \{v \in H^1(0, 1) : v(0) = 0\}$ that we endow with the norm $\|\cdot\|_V := |\cdot|_{H^1(0,1)}$. This is possible thanks to the Friedrich-Poincaré inequality, see (2.5), which reads here

$$\|u\|_{L^2(0,1)} \leq C_F \|u'\|_{L^2(0,1)} \quad (3.2)$$

and holds for instance for $C_F = \frac{1}{\sqrt{2}} \leq 1$. The weak form of problem (3.1) is given by

$$\text{find } u \in V : \int_0^1 au'v'dx + \int_0^1 buu'v'dx = \int_0^1 fvdx \quad \forall v \in V. \quad (3.3)$$

We first show, under suitable conditions on the data, that the problem (3.1) is well-posed. Since we do not have an a priori estimate, due to the mixed Neumann-Dirichlet boundary conditions¹, we restrict ourselves to the set of functions whose norm is bounded by a certain constant. More precisely, we consider

$$\mathcal{M} := \{v \in V : \|v'\|_{L^2(0,1)} \leq r\} \quad \text{with} \quad r = \sqrt{\frac{C_F}{b} \|f\|_{L^2(0,1)}}.$$

Since \mathcal{M} is a closed ball in V , it is bounded, convex and closed in V . The well-posedness of the problem under certain assumptions on the data is proved in the following proposition.

Proposition 3.1.1. *If $\frac{a}{b} \geq 2r$, then there exists a solution $u \in \mathcal{M}$ to problem (3.3). Moreover, if $\frac{a}{b} > 2r$, then such solution is unique.*

Strictly speaking, it is enough to assume $\frac{a}{b} > r$ to prove the existence of a solution in \mathcal{M} . Using the definition of r , the condition $\frac{a}{b} > 2r$ can be expressed more explicitly in terms of the given

¹If we have homogeneous Dirichlet conditions in $x = 0$ and $x = 1$, we have an *a priori* estimate. Indeed, it is easy to show that $\|u'\|_{L^2(0,1)} \leq \frac{1}{a} \|f\|_{L^2(0,1)}$ taking $v = u$ in (3.3) and using the fact that $\int_0^1 buu'u = 0$. The existence of a solution can then be proved using for instance Schauder's fixed point theorem while for the uniqueness, it holds under the constraint $C_F \|f\|_{L^2(0,1)} < \frac{a^2}{b}$.

3.1. Steady-state viscous Burgers' equation in random intervals

data by $C_F \|f\|_{L^2(0,1)} < \frac{a^2}{4b}$, which coincides with the one given in [28] replacing $C_F \|f\|_{L^2(0,1)}$ by the dual norm $\|f\|_{V'}$. The proof of Proposition 3.1.1, given below for completeness, uses the Schauder's fixed point theorem for the existence and is inspired by the one given in [120]. The uniqueness is proved using a variational argument.

Proof. Existence: we define the mapping $T : \mathcal{M} \rightarrow V$, $u \mapsto Tu =: w$, where $w \in V$ is the unique solution of

$$\text{find } w \in V : \mathcal{A}_u(w, v) = F(v) \quad \forall v \in V \quad (3.4)$$

with

$$\mathcal{A}_u(w, v) := \int_0^1 a w' v' dx + \int_0^1 b u w' v dx \quad \text{and} \quad F(v) := \int_0^1 f v dx.$$

We show that T is well-defined, maps \mathcal{M} to \mathcal{M} and is compact. Let $u \in \mathcal{M}$, i.e. $\|u'\|_{L^2(0,1)} \leq r$. The fact that $T : \mathcal{M} \rightarrow V$ is well-defined follows directly from Lax-Milgram's lemma. Indeed, for any $v, w \in V$ we have

$$\mathcal{A}_u(w, v) \leq a \|w'\|_{L^2(0,1)} \|v'\|_{L^2(0,1)} + b \|u\|_{L^4(0,1)} \|w'\|_{L^2(0,1)} \|v\|_{L^4(0,1)} \leq (a + br) \|w'\|_{L^2(0,1)} \|v'\|_{L^2(0,1)}$$

using successively Cauchy-Schwarz and Hölder's inequalities and the fact that

$$\|v\|_{L^4(0,1)} \leq C \|v'\|_{L^2(0,1)} \quad \text{holds with } C = 1. \quad (3.5)$$

Moreover, since $u \in \mathcal{M}$ and $\frac{a}{b} \geq 2r$ by assumption, we have

$$-\int_0^1 b u w' w dx \leq b \|u'\|_{L^2(0,1)} \|w'\|_{L^2(0,1)}^2 \leq br \|w'\|_{L^2(0,1)}^2 \leq \frac{a}{2} \|w'\|_{L^2(0,1)}^2$$

and thus

$$\mathcal{A}_u(w, w) = a \|w'\|_{L^2(0,1)}^2 + \int_0^1 b u w' w dx \geq \frac{a}{2} \|w'\|_{L^2(0,1)}^2.$$

Finally, thanks to (3.2) we get

$$F(v) \leq C_F \|f\|_{L^2(0,1)} \|v'\|_{L^2(0,1)}$$

and the assumptions of Lax-Milgram's lemma are satisfied. We now show that T maps \mathcal{M} to itself, i.e. $Tu = w \in \mathcal{M}$. Thanks to the coercivity of \mathcal{A}_u and the continuity of F , taking $v = w$ in (3.4) yields

$$\frac{a}{2} \|w'\|_{L^2(0,1)}^2 \leq \int_0^1 f w dx \leq C_F \|f\|_{L^2(0,1)} \|w'\|_{L^2(0,1)}$$

and thus

$$\|w'\|_{L^2(0,1)} \leq \frac{2}{a} C_F \|f\|_{L^2(0,1)} = \frac{2}{a} br^2 \leq r.$$

We finally show that T is compact. Let $(u_n)_{n \in \mathbb{N}}$ be a bounded sequence in \mathcal{M} . Since $H^1(0,1)$ is compactly embedded in $L^4(0,1)$, there exists a subsequence $(u_{n_j})_{j \in \mathbb{N}}$ which converges in $L^4(0,1)$. Let u_n and u_m be two elements of this subsequence and write w_n and w_m the

corresponding images under T . We have

$$\int_0^1 a(w'_n - w'_m)v' dx + \int_0^1 b[u_n(w'_n - w'_m) + w'_m(u_n - u_m)]v dx = 0 \quad \forall v \in V.$$

If we take $v = w_n - w_m$, using that $u_n \in \mathcal{M}$ and $br \leq \frac{a}{2}$ we can easily show that

$$\frac{a}{2} \|w'_n - w'_m\|_{L^2(0,1)} \leq b \|w'_m\|_{L^2(0,1)} \|u_n - u_m\|_{L^4(0,1)}$$

and thus

$$\|w'_n - w'_m\|_{L^2(0,1)} \leq \|u_n - u_m\|_{L^4(0,1)}$$

since $w_m \in \mathcal{M}$. Therefore, $(w_{n_j})_{j \in \mathbb{N}}$ is a Cauchy sequence in V and thus converges.

Uniqueness: we use a variational argument. Let $u_1, u_2 \in \mathcal{M}$ be two solutions of problem (3.3).

We have

$$\int_0^1 a(u'_1 - u'_2)v dx + \int_0^1 b(u_1 u'_1 - u_2 u'_2)v dx = 0 \quad \forall v \in V.$$

If we take $v = u_1 - u_2$, we obtain

$$\begin{aligned} a \|u'_1 - u'_2\|_{L^2(0,1)}^2 &= - \int_0^1 b(u_1(u'_1 - u'_2) + u'_2(u_1 - u_2))(u_1 - u_2) dx \\ &\leq b \|u_1\|_{L^4(0,1)} \|u_1 - u_2\|_{L^2(0,1)} \|u_1 - u_2\|_{L^4(0,1)} \\ &\quad + b \|u'_2\|_{L^2(0,1)} \|u_1 - u_2\|_{L^4(0,1)}^2 \\ &\leq b (\|u'_1\|_{L^2(0,1)} + \|u'_2\|_{L^2(0,1)}) \|u'_1 - u'_2\|_{L^2(0,1)}^2 \\ &\leq 2br \|u'_1 - u'_2\|_{L^2(0,1)}^2 \end{aligned}$$

and thus

$$(a - 2br) \|u'_1 - u'_2\|_{L^2(0,1)}^2 \leq 0.$$

Since $\frac{a}{b} > 2r$ by assumption, the last inequality implies $u'_1 = u'_2$. The fact that $u_1(0) = u_2(0)$ allows us to conclude that $u_1 = u_2$. \square

Remark 3.1.2. If the solution is assumed to be in $H^2(0, 1)$, we can alternatively use Schaefer's fixed point theorem [62] to prove the existence of a solution to problem (3.1).

We now give an *a posteriori* estimate of the error in the V norm between the exact solution u and its finite element approximation. We thus consider

$$0 = x_0 < x_1 < \dots < x_N < x_{N+1} = 1$$

a partition of $[0, 1]$ and let $h_i = x_{i+1} - x_i$ for $i = 0, \dots, N$. Let $V_h \subset V$ be the finite dimensional space of continuous piecewise polynomials of degree less or equal to one associated to this

3.1. Steady-state viscous Burgers' equation in random intervals

partition (the usual hat functions). The finite element approximation of problem (3.3) reads

$$\text{find } u_h \in V_h : \int_0^1 a u'_h v'_h dx + \int_0^1 b u_h u'_h v_h dx = \int_0^1 f v_h dx \quad \forall v_h \in V_h. \quad (3.6)$$

Similarly to the continuous case, we can show that there exists a unique solution $u_h \in \mathcal{M}_h$ to problem (3.6) if $\frac{a}{b} > 2r$, with $\mathcal{M}_h = \{v_h \in V_h : \|v'_h\|_{L^2(0,1)} \leq r\} \subset \mathcal{M}$. Moreover, if we take $v = v_h$ in (3.3) and subtract (3.6), we get the following so-called Galerkin orthogonality property

$$\int_0^1 a(u' - u'_h) v'_h dx + \int_0^1 b(uu' - u_h u'_h) v_h dx = 0 \quad \forall v_h \in V_h. \quad (3.7)$$

Proposition 3.1.3. *If a, b and f are such that $\frac{a}{b} > 2r$, i.e. $\frac{4b}{a^2} C_F \|f\|_{L^2(0,1)} < 1$, then there exists a constant $C > 0$ independent of h and u such that*

$$\|u' - u'_h\|_{L^2(0,1)} \leq \frac{C}{a} \left(\sum_{i=0}^N \eta_i^2 \right)^{\frac{1}{2}} \quad (3.8)$$

with

$$\eta_i^2 = h_i^2 \int_{x_i}^{x_{i+1}} (f - b u_h u'_h + a u''_h)^2 dx, \quad i = 0, \dots, N. \quad (3.9)$$

Proof. For any $v \in V$, let $\langle \mathcal{R}(u_h), v \rangle = \int_0^1 (f v - b u_h u'_h v - a u'_h v') dx$ denote the residual for u_h . We have

$$\begin{aligned} \int_0^1 a(u' - u'_h) v dx &= \int_0^1 f v dx - \int_0^1 b u u' v dx - \int_0^1 a u'_h v' dx \\ &= \langle \mathcal{R}(u_h), v \rangle - \int_0^1 b(uu' - u_h u'_h) v dx. \end{aligned}$$

If we take $v = u - u_h$, the second term can be bounded by

$$- \int_0^1 b(uu' - u_h u'_h) v dx \leq 2br \|u' - u'_h\|_{L^2(0,1)}^2.$$

Therefore

$$\|u' - u'_h\|_{L^2(0,1)}^2 \leq \frac{1}{a} \langle \mathcal{R}(u_h), u - u_h \rangle + \frac{2br}{a} \|u' - u'_h\|_{L^2(0,1)}^2.$$

Since $\frac{a}{b} > 2r$ by assumption, there exists $\gamma > 0$ such that $\frac{2br}{a} \leq 1 - \gamma$. Therefore, we have

$$\|u' - u'_h\|_{L^2(0,1)}^2 \leq \frac{1}{a\gamma} \mathcal{R}(u - u_h). \quad (3.10)$$

It only remains to give an estimation of the residual. First note that

$$\langle \mathcal{R}(u_h), v_h \rangle = 0 \quad \forall v_h \in V_h.$$

Taking $v_h = r_h v$ the Lagrange interpolant of v and using standard techniques, we get

$$\langle \mathcal{R}(u_h), v \rangle \leq C_I \left(\sum_{i=0}^N h_i^2 \int_{x_i}^{x_{i+1}} (f - bu_h u_h' + au_h'')^2 dx \right)^{\frac{1}{2}} \|v'\|_{L^2(0,1)} \quad (3.11)$$

where C_I is the constant (independent of h and v) in the interpolation error bound

$$\|v - r_h v\|_{L^2(x_i, x_{i+1})} \leq C_I h_i \|v'\|_{L^2(x_i, x_{i+1})}. \quad (3.12)$$

For instance, we can take $C_I = \sqrt{\frac{49}{30}}$. Inserting (3.11) in (3.10) yields (3.8) with $C = \frac{C_I}{\gamma}$. \square

Remark 3.1.4. *The a posteriori error estimate (3.8) holds under the constraint $\frac{2br}{a} < 1$, i.e. $\frac{2br}{a} \leq 1 - \gamma$ for a certain $\gamma > 0$. However, if γ is chosen too small, then the constant C explodes. In practice, it is common to assume that the input data are such that $\frac{2br}{a} \leq \frac{1}{2}$ holds.*

3.1.2 Random case

Let (Ω, \mathcal{F}, P) be a complete probability space and for any $\omega \in \Omega$ let $D_\omega := (0, s(\omega)) \subseteq \hat{D}$ be an interval of random length $s(\omega)$. To simplify the notation, the set

$$\{(x, \omega) : x \in D_\omega, \omega \in \Omega\}$$

will be denoted by $D_\omega \times \Omega$ in the sequel. The goal is to solve the problem:

find $\tilde{u} : D_\omega \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω

$$\begin{cases} -a \frac{\partial^2}{\partial x^2} \tilde{u}(x, \omega) + b \tilde{u}(x, \omega) \frac{\partial}{\partial x} \tilde{u}(x, \omega) &= \tilde{f}(x) & x \in D_\omega \\ \tilde{u}(0, \omega) &= 0 \\ \frac{\partial}{\partial x} \tilde{u}(s(\omega), \omega) &= 0, \end{cases} \quad (3.13)$$

where a and b are positive constants and $\tilde{f} \in L^2(\hat{D})$ is a deterministic forcing term. Let $\tilde{V}_\omega = \{\tilde{v} \in H^1(D_\omega) : \tilde{v}(0, \omega) = 0 \text{ a.s. in } \Omega\}$. The pointwise weak form of problem (3.13) reads:

find $\tilde{u}(\cdot, \omega) \in \tilde{V}_\omega$ such that

$$\int_0^{s(\omega)} a \frac{\partial \tilde{u}(\cdot, \omega)}{\partial x} \frac{\partial \tilde{v}}{\partial x} dx + \int_0^{s(\omega)} b \tilde{u}(\cdot, \omega) \frac{\partial \tilde{u}(\cdot, \omega)}{\partial x} \tilde{v} dx = \int_0^{s(\omega)} \tilde{f} \tilde{v} dx \quad \forall \tilde{v} \in \tilde{V}_\omega. \quad (3.14)$$

For ease of presentation, we will use the short hand notation $\tilde{u}(\omega) = \tilde{u}(\cdot, \omega)$ when no confusion arises. Instead of solving this problem on the stochastic domain D_ω , we will solve it on a fixed reference domain, namely $D = (0, 1)$, by considering the change of variable $x = s(\omega)\xi$. Therefore, assuming $s(\omega) > 0$ a.s. in Ω we define the (random) mapping

$$\begin{aligned} g_\omega : D_\omega &\rightarrow D \\ x &\mapsto \xi = g_\omega(x) = \frac{x}{s(\omega)} \end{aligned} \quad (3.15)$$

3.1. Steady-state viscous Burgers' equation in random intervals

whose inverse is given by

$$\begin{aligned} g_\omega^{-1}: D &\rightarrow D_\omega \\ \xi &\mapsto x = g_\omega^{-1}(\xi) = s(\omega)\xi. \end{aligned}$$

Let $u(\xi, \omega) = \tilde{u}(x, \omega)$ and $f(\xi, \omega) = \tilde{f}(x, \omega)$ denote respectively the velocity and the forcing term on the fixed domain D , i.e. $u(\xi, \omega) = \tilde{u}(g_\omega^{-1}(\xi), \omega)$ and $f(\xi, \omega) = \tilde{f}(g_\omega^{-1}(\xi))$. Finally, let $V = \{v \in H^1(D) : v(0) = 0\}$. Applying the standard chain rule and the change of variable formula, the pointwise weak problem (3.14) can then be rewritten:

find $u(\omega) \in V$ such that

$$\int_0^1 \frac{a}{s(\omega)} \frac{\partial u(\omega)}{\partial \xi} \frac{\partial v}{\partial \xi} d\xi + \int_0^1 b u(\omega) \frac{\partial u(\omega)}{\partial \xi} v d\xi = \int_0^1 s(\omega) f(\omega) v d\xi \quad \forall v \in V. \quad (3.16)$$

The strong form of the problem on the reference domain can be stated as:

find $u : D \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω

$$\begin{cases} -\frac{a}{s(\omega)^2} \frac{\partial^2}{\partial \xi^2} u(\xi, \omega) + \frac{b}{s(\omega)} u(\xi, \omega) \frac{\partial}{\partial \xi} u(\xi, \omega) &= f(\xi, \omega) & \xi \in D \\ u(0, \omega) &= 0 \\ \frac{\partial}{\partial \xi} u(1, \omega) &= 0. \end{cases} \quad (3.17)$$

Notice that here, performing the change of variable on the variational formulation (3.14) of the problem or directly on the strong formulation (3.13) yields the same result, which is not the case in general. This is due to the fact that s does not depend on the physical variable plus the fact that we are considering the pointwise (in ω) weak formulation.

From now on, we assume that the random length of interval $s(\omega)$ has the form

$$s(\omega) = s_0 + \varepsilon Y(\omega),$$

where Y is a random variable with zero mean, unit variance and bounded image Γ . Moreover, we assume that Y is such that $s(\omega)$ is bounded almost surely from below and above by respectively s_{min} and s_{max} . More precisely, we assume that

$$\exists 0 < s_{min} \leq s_{max} < \infty : P(\omega \in \Omega : s_{min} \leq s(\omega) \leq s_{max}) = 1. \quad (3.18)$$

Due to the Doob-Dynkin lemma, the solution u of (3.17) depends on the same random variable as s , i.e. $u(\xi, \omega) = u(\xi, Y(\omega))$. Let $\rho : \Gamma \rightarrow \mathbb{R}^+$ denotes the density function of Y . The solution of problem (3.17) can then be sought either in the probability space (Ω, \mathcal{F}, P) or in its image space $(\Gamma, B(\Gamma), \rho(y) dy)$. The stochastic problem (3.17) can indeed be written in the following

Chapter 3. PDEs in random domains

deterministic parametric form:

find $u : D \times \Gamma \rightarrow \mathbb{R}$ such that ρ -a.e. in Γ we have

$$\begin{cases} -\frac{a}{s(y)^2} \frac{\partial^2}{\partial \xi^2} u(\xi, y) + \frac{b}{s(y)} u(\xi, y) \frac{\partial}{\partial \xi} u(\xi, y) &= f(\xi, y) \quad \xi \in D \\ u(0, y) &= 0 \\ \frac{\partial}{\partial \xi} u(1, y) &= 0. \end{cases} \quad (3.19)$$

From now on, we will drop the dependence of the functions on either ξ , ω or y when no confusion is possible. Furthermore, we will write u' for $\frac{\partial}{\partial \xi}$. Since s is expanded as sum of coefficients, it is more convenient to have all its occurrences in the numerator rather than having division by s . Therefore, we will consider the following weak form of problem (3.19):

find $u(y) \in V$ such that

$$\int_0^1 a u' v' d\xi + \int_0^1 b s(y) u u' v d\xi = \int_0^1 s^2(y) f v d\xi \quad \forall v \in V, \rho\text{-a.e. in } \Gamma. \quad (3.20)$$

Before giving an *a posteriori* error estimation for the problem (3.17), and thus for the problem (3.13), we briefly give a condition on the given data that ensures the well-posedness of the problem. Recall that $f = \tilde{f} \circ g_\omega^{-1}$, i.e. $f(\xi, \omega) = \tilde{f}(s(\omega)\xi)$. Thanks to the uniform bounds on s , we have in particular $s^k f \in L_P^2(\Omega; L^2(D))$ for any k . Notice that it can be shown using only the lower bound s_{\min} or the upper bound s_{\max} depending on the sign of k . For instance, we have for the right-hand side of (3.20)

$$\|s^2(\omega) f(\omega)\|_{L^2(D)} = s^{\frac{3}{2}}(\omega) \|\tilde{f}\|_{L^2(D_\omega)} \leq s_{\max}^{\frac{3}{2}} \|\tilde{f}\|_{L^2(\hat{D})} < \infty \quad \text{a.s. in } \Omega.$$

More generally, we can easily show that the assumption (3.18) ensures that the spaces $L^2(D_\omega)$ and $L^2(D)$, respectively \tilde{V}_ω and V , are isomorphic. This is precisely stated in the following proposition.

Proposition 3.1.5. *Under assumption (3.18), for any $\tilde{f} \in L^2(D_\omega)$ and any $\tilde{v} \in \tilde{V}_\omega$ we have a.s. in Ω*

$$\sqrt{s_{\min}} \|f\|_{L^2(D)} \leq \|\tilde{f}\|_{L^2(D_\omega)} \leq \sqrt{s_{\max}} \|f\|_{L^2(D)}$$

and

$$\frac{1}{\sqrt{s_{\max}}} \left\| \frac{\partial v}{\partial \xi} \right\|_{L^2(D)} \leq \left\| \frac{\partial \tilde{v}}{\partial x} \right\|_{L^2(D_\omega)} \leq \frac{1}{\sqrt{s_{\min}}} \left\| \frac{\partial v}{\partial \xi} \right\|_{L^2(D)}$$

with $f = \tilde{f} \circ g_\omega^{-1}$ and $v = \tilde{v} \circ g_\omega^{-1}$. The same relations hold for any $f \in L^2(D)$ and any $v \in V$ with $\tilde{f} = f \circ g_\omega$ and $\tilde{v} = v \circ g_\omega$.

Similarly to the deterministic problem (3.1), we restrict ourselves to the solutions which lie in \mathcal{M} defined by

$$\mathcal{M} := \{v \in L_P^2(\Omega; V) : \|v(\omega)'\|_{L^2(0,1)} \leq r_\omega \text{ a.s. in } \Omega\} \quad (3.21)$$

3.1. Steady-state viscous Burgers' equation in random intervals

with $r_\omega = \sqrt{\frac{s(\omega)}{b}} C_F \|f(\omega)\|_{L^2(0,1)}$, where C_F is the Friedrich-Poincaré constant on the reference interval D given in (3.2). Since

$$s(\omega) \|f(\omega)\|_{L^2(0,1)} = \frac{s(\omega)}{\sqrt{s(\omega)}} \|\tilde{f}\|_{L^2(D_\omega)} \leq \sqrt{s_{\max}} \|\tilde{f}\|_{L^2(\hat{D})} < \infty \quad \text{a.s. in } \Omega,$$

we have $r_\omega \in L_P^\infty(\Omega)$. Therefore, since $L_P^\infty(\Omega) \subset L_P^2(\Omega)$, \mathcal{M} is a closed ball in $L_P^2(\Omega; V)$ and thus \mathcal{M} is bounded, convex and closed in $L_P^2(\Omega; V)$.

The well-posedness of the stochastic problem can thus be proved following a reasoning similar to the one used in the deterministic case.

Proposition 3.1.6. *If $bs(\omega)r_\omega \leq \frac{a}{2}$ a.s. in Ω , or in other words if $\frac{4bs_{\max}^3}{a^2} C_F \|f(\omega)\|_{L^2(D)} \leq 1$ a.s. in Ω , then there exists a solution $u \in \mathcal{M}$ to problem (3.20). Furthermore, if the inequality is strict, then the solution is unique.*

Remark 3.1.7. *We can show the well-posedness of the problem under the slightly less restrictive assumption*

$$\frac{4C_F b s_{\max}^{5/2}}{a^2} \|\tilde{f}\|_{L^2(D_\omega)} < 1 \quad \text{a.s. in } \Omega, \quad (3.22)$$

setting then $r_\omega = \sqrt{\frac{s(\omega)}{b}} C_F \|f(\omega)\|_{L^2(0,1)}$ in (3.21). The inequality (3.22) holds true if the input data satisfy the assumption of Proposition 3.1.6 since $\|f(\omega)\|_{L^2(D)} \geq s_{\max}^{-\frac{1}{2}} \|\tilde{f}\|_{L^2(D_\omega)}$ by Proposition 3.1.5. We refer to Remark 3.2.9 for the same discussion about the small data assumption for the well-posedness of the Navier-Stokes problem and we mention that the assumption of Proposition 3.1.6 and (3.22) are consistent with (3.43) and (3.41), respectively.

We use a perturbation approach and write

$$u(\xi, Y(\omega)) = u_0(\xi) + \varepsilon u_1(\xi, Y(\omega)) + \mathcal{O}(\varepsilon^2)$$

with ε a small parameter that controls the amplitude of the variation of s . The goal is now to derive an *a posteriori* error estimate for the approximation $u \approx u_{0,h}$ with $u_{0,h}$ the finite element approximation of u_0 . We assume that $\tilde{f} \in H^2(\hat{D})$ which allows us to write $f = f(\xi, Y(\omega))$ as $f = f_0 + \varepsilon f_1 Y + \varepsilon^2 f_2 Y^2$ with

$$f_0(\xi) = \tilde{f}(s_0 \xi), \quad f_1(\xi) = \frac{\partial \tilde{f}}{\partial x}(s_0 \xi) \xi \quad \text{and} \quad f_2(\xi, Y(\omega)) = \xi^2 \int_0^1 (1-t) \frac{\partial^2 \tilde{f}}{\partial x^2}(s_0 \xi + \varepsilon Y(\omega) \xi t) dt,$$

using a Taylor expansion with integral remainder of $\tilde{f}(s\xi)$, $s = s_0 + \varepsilon Y$. The deterministic part u_0 of the solution can be found by solving

$$\begin{cases} -au_0'' + bs_0 u_0 u_0' &= s_0^2 f_0 & \text{in } D \\ u_0(0) &= 0 \\ u_0'(1) &= 0. \end{cases} \quad (3.23)$$

Remark 3.1.8. Notice that we could also choose to take $(0, s_0)$ as the reference domain, i.e. the interval corresponding to the case $\varepsilon = 0$, using then the mapping $g_\omega(x) = \frac{s_0 x}{s(\omega)}$ instead of (3.15). In this case, the problem for u_0 would not contain the coefficient s_0 , contrary to (3.23). We should then be careful when using for instance the Friedrich-Poincaré inequality (3.2) which holds on $(0, s_0)$ up to a factor s_0 .

We use the finite element method to approximate numerically the solution u_0 of problem (3.23). To this aim, we consider $0 = \xi_0 < \xi_1 < \dots < \xi_N < \xi_{N+1} = 1$ a partition of D and let $h_i = \xi_{i+1} - \xi_i$ for $i = 0, \dots, N$. Then, we consider V_h the finite dimensional space of V constituted of the corresponding continuous, piecewise linear finite element functions that vanish in 0. We now give an *a posteriori* estimate of the error between the exact solution u and the finite element approximation $u_{0,h}$ of u_0 in the $L^2_p(\Omega; V)$ norm.

Proposition 3.1.9. If $\frac{2bs(\omega)r_\omega}{a} \leq \frac{1}{2}$ a.s. in Ω , then there exists a constant $C > 0$ depending only on s_0, f_0, f_1 and $\mathbb{E}[Y^k f_2^p]$ for $p = 0, 1, 2$ and some $3 \leq k \leq 8$ such that

$$\mathbb{E} \left[\|u' - u'_{0,h}\|_{L^2(0,1)}^2 \right]^{\frac{1}{2}} \leq \frac{2\sqrt{2}}{a} [\eta_h^2 + \eta_\varepsilon^2]^{\frac{1}{2}} + C\varepsilon^2, \quad (3.24)$$

with

$$\eta_h^2 := C_I^2 \sum_{i=0}^N h_i^2 \int_{\xi_i}^{\xi_{i+1}} (s_0^2 f_0 - bs_0 u_{0,h} u'_{0,h} + a u_{0,h}'')^2 d\xi \quad (3.25)$$

$$\eta_\varepsilon^2 := \varepsilon^2 C_F^2 \|2s_0 f_0 + s_0^2 f_1 - b u_{0,h} u'_{0,h}\|_{L^2(D)}^2, \quad (3.26)$$

where C_I and C_F are the constants in (3.12) and (3.2), respectively.

Remark 3.1.10. The factor 2 in (3.24) comes from the assumption $\frac{2bs(\omega)r_\omega}{a} \leq \frac{1}{2}$ on the input data, which is imposed so that the constant does not explode, see also Remark 3.1.4.

Proof. For any $v \in V$ and a.s. in Ω we can decompose

$$\begin{aligned} \int_0^1 a(u' - u'_{0,h}) v' d\xi &= \underbrace{\int_0^1 (s_0^2 f_0 v - bs_0 u_{0,h} u'_{0,h} v - a u'_{0,h} v') d\xi}_{A_1(v)} + \underbrace{\int_0^1 (s^2 f - s_0^2 f_0) v d\xi}_{A_2(v)} \\ &\quad - \underbrace{\int_0^1 bs(uu' - u_{0,h} u'_{0,h}) v d\xi}_{A_3(v)} - \underbrace{\int_0^1 b(s - s_0) u_{0,h} u'_{0,h} v d\xi}_{A_4(v)} \end{aligned}$$

and thus

$$\|u' - u'_{0,h}\|_{L^2(D)}^2 = \frac{1}{a} [A_1(u - u_{0,h}) + A_2(u - u_{0,h}) + A_3(u - u_{0,h}) + A_4(u - u_{0,h})].$$

Let us consider each term separately. First of all, note that the first term A_1 corresponds to the residual for $u_{0,h}$, the finite element approximation of problem (3.23). Using a standard

procedure, it can be bounded by

$$A_1(v) \leq \left(C_I^2 \sum_{i=0}^N h_i^2 \int_{\xi_i}^{\xi_{i+1}} (s_0^2 f_0 - b s_0 u_{0,h} u'_{0,h} + a u''_{0,h})^2 d\xi \right)^{\frac{1}{2}} \|v'\|_{L^2(D)}$$

with C_I the constant in (3.12). Thanks to the Cauchy-Schwarz and Friedrich-Poincaré inequalities, the second and fourth terms, that we keep together for sharpness², can be bounded by

$$A_2(v) + A_4(v) \leq C_F \|s^2 f - s_0^2 f_0 - b(s - s_0) u_{0,h} u'_{0,h}\|_{L^2(D)} \|v'\|_{L^2(D)}.$$

Finally, we consider the term A_3 which is due to the nonlinear part of the problem. If we take $v = u - u_{0,h} \in V$ a.s. in Ω , it can be bounded by

$$A_3(u - u_{0,h}) \leq 2bs(\omega)r_\omega \|u' - u'_{0,h}\|_{L^2(D)}^2$$

using Hölder's inequality, Sobolev embedded theorem and the fact that $\|u'\|_{L^2(0,1)}$ and $\|u'_{0,h}\|_{L^2(0,1)}$ are bounded by r_ω a.s. in Ω . Thanks to the assumption that $\frac{2bs(\omega)r_\omega}{a} \leq \frac{1}{2}$ a.s. in Ω , we have

$$\frac{1}{a} A_3(u - u_{0,h}) \leq \frac{1}{2} \|u' - u'_{0,h}\|_{L^2(D)}^2.$$

Altogether, we obtain

$$\begin{aligned} \|u' - u'_{0,h}\|_{L^2(D)} &\leq \frac{2}{a} \left[\left(C_I^2 \sum_{i=0}^N h_i^2 \int_{\xi_i}^{\xi_{i+1}} (s_0^2 f_0 - b s_0 u_{0,h} u'_{0,h} + a u''_{0,h})^2 d\xi \right)^{\frac{1}{2}} \right. \\ &\quad \left. + C_F \|s^2 f - s_0^2 f_0 - b(s - s_0) u_{0,h} u'_{0,h}\|_{L^2(D)} \right] \end{aligned}$$

which yields

$$\begin{aligned} \|u' - u'_{0,h}\|_{L^2(D)}^2 &\leq \frac{8}{a^2} \left[C_I^2 \sum_{i=0}^N h_i^2 \int_{\xi_i}^{\xi_{i+1}} (s_0^2 f_0 - b s_0 u_{0,h} u'_{0,h} + a u''_{0,h})^2 d\xi \right. \\ &\quad \left. + C_F^2 \|s^2 f - s_0^2 f_0 - b(s - s_0) u_{0,h} u'_{0,h}\|_{L^2(D)}^2 \right]. \end{aligned}$$

Since Y has zero mean and unit variance, the result follows taking first the expected value and then the square root on both sides of last inequality. Indeed, we have

$$s^2 f - s_0^2 f_0 = \varepsilon(2s_0 f_0 + s_0^2 f_1) Y + \varepsilon^2(f_0 + 2s_0 f_1 + s_0^2 f_2) Y^2 + \varepsilon^3(f_1 + 2s_0 f_2) Y^3 + \varepsilon^4 f_2 Y^4$$

from which we deduce, recalling that $s - s_0 = \varepsilon Y$,

$$\mathbb{E} \left[\|s^2 f - s_0^2 f_0 - b(s - s_0) u_{0,h} u'_{0,h}\|_{L^2(D)}^2 \right] = \varepsilon^2 \|2s_0 f_0 + s_0^2 f_1 - b u_{0,h} u'_{0,h}\|_{L^2(D)}^2 + C_2 \varepsilon^3$$

²Notice that we get comparable results if we bound these two terms separately, in which case the estimator due to the uncertainty reads $\eta_\varepsilon^2 = 2\varepsilon^2 C_F^2 \left(\|2s_0 f_0 + s_0^2 f_1\|_{L^2(D)}^2 + b^2 \|u_{0,h} u'_{0,h}\|_{L^2(D)}^2 \right)$.

where C_2 depends only on $s_0, f_0, f_1, \mathbb{E}[Y^k]$ for $k = 3, \dots, 6, \mathbb{E}[Y^k f_2]$ for $k = 3, \dots, 7$ and $\mathbb{E}[Y^k f_2^2]$ for $k = 4, \dots, 8$. \square

Notice that we have used the Friedrich-Poincaré inequality to bound the terms A_2 and A_4 due to the forcing and nonlinear terms, for which 1 is a uniform bound for $D = (0, 1)$. The loss due to the use of this inequality is different from case to case, therefore affecting the efficiency of the estimator η_ε when changing the input data.

3.1.3 Numerical results

We consider here two numerical examples for the Burgers' equation. We choose $s_0 = 1$ for simplicity. We start with the results for the deterministic case presented in Section 3.1.1.

Deterministic case

Let $a = b = 1$. For the first example, we consider

$$\tilde{u}(x) = -0.3 \tanh(x) + 0.3 \operatorname{sech}(1)^2 x, \quad x \in (0, 1), \quad (3.27)$$

and compute the corresponding right-hand side $\tilde{f} = -a\tilde{u}'' + b\tilde{u}u$ and for the second example we set the source term to

$$\tilde{g}(x) = \sin(\pi x). \quad (3.28)$$

Notice that \tilde{g} does not satisfy the bound $C_F \|\tilde{g}\|_{L^2(0,1)} < \frac{a^2}{4b} = 0.25$ with $C_F = 1/\sqrt{2}$ since $C_F \|\tilde{g}\|_{L^2(0,1)} = 0.5$. We give in Table 3.1 the results for these two cases considering various (uniform) partitions of $[0, 1]$. Here, error stands for the error $\|u' - u'_h\|_{L^2(0,1)}$, while

$$\eta = \frac{1}{a} \left(\sum_{i=0}^N \eta_i^2 \right)^{\frac{1}{2}} \quad \text{with } \eta_i \text{ in (3.9)}$$

and e.i. denotes the ratio between the estimator η and the error. The error is computed with the exact solution for the first case (3.27) and with respect to the reference solution obtained with $h_{ref} = 2^{-12}$ for the second case (3.28).

By looking at the effectivity index for both cases, we see that for h small enough, we recover the value $3.46 \approx 2\sqrt{3}$ obtained in the one-dimensional numerical examples of the previous chapters, see also Appendix 1.C. The slight increase of e.i. for small value of h in the second case (3.28) is due to the fact that the error is computed with respect to a reference solution.

Random case

We consider now the case of random interval $D_\omega = (0, s(\omega))$ with $s(\omega) = s_0 + \varepsilon Y(\omega) = 1 + \varepsilon Y(\omega)$, where Y is a uniform random variable in $[-\sqrt{3}, \sqrt{3}]$. Considering \tilde{f} and \tilde{g} defined above

3.1. Steady-state viscous Burgers' equation in random intervals

h	\tilde{f}			\tilde{g}		
	error	η	e.i.	error	η	e.i.
1/4	1.3440e-2	4.6302e-2	3.4452	5.0731e-2	1.6499e-1	3.2523
1/8	6.7111e-3	2.3217e-2	3.4594	2.4118e-2	8.2193e-2	3.4080
1/16	3.3545e-3	1.1616e-2	3.4629	1.1901e-2	4.1057e-2	3.4499
1/32	1.6771e-3	5.8092e-3	3.4638	5.9306e-3	2.0524e-2	3.4606
1/64	8.3855e-4	2.9047e-3	3.4640	2.9626e-3	1.0261e-2	3.4636
1/128	4.1927e-4	1.4524e-3	3.4641	1.4804e-3	5.1306e-3	3.4656
1/256	2.0964e-4	7.2620e-4	3.4641	7.3909e-4	2.5653e-3	3.4708
1/512	1.0482e-4	3.6340e-4	3.4641	3.6736e-4	1.2826e-3	3.4915
1/1024	5.2409e-5	1.8155e-4	3.4641	1.7925e-4	6.4132e-4	3.5777

Table 3.1: Error, estimator and effectivity index for the deterministic Burgers' equation with mesh size $2^{-2} \leq h \leq 2^{-10}$.

as (deterministic) forcing terms for the problems on the physical random domain D_ω , the corresponding right-hand sides for the problems on the reference interval $(0, 1)$ are then given by $f(\xi, \omega) = \tilde{f}(s(\omega)\xi)$ and $g(\xi, \omega) = \tilde{g}(s(\omega)\xi)$, respectively. We give in Figure 3.3 the graph of the function f and the corresponding solution u of problem (3.20) for different values of s and the results for the second case g can be found in Figure 3.4.

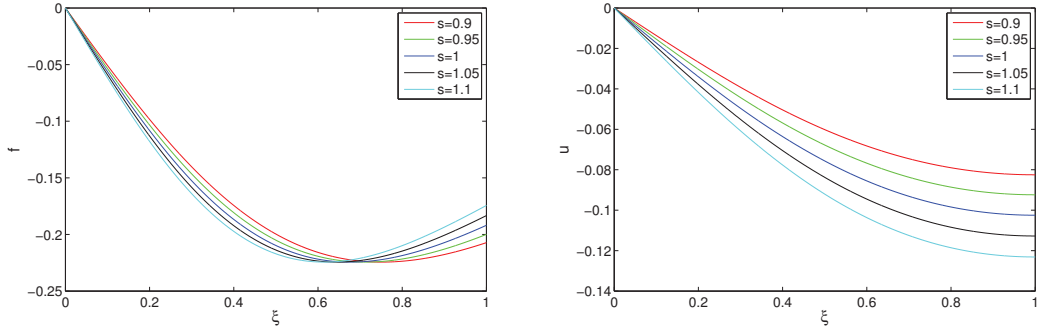


Figure 3.3: Function f and corresponding solution u for various values of s .

We give then in Table 3.2 the error $\|u - u_{0,h}\|_{L^2_p(\Omega;V)}$, the estimators η_h and η_ε defined in (3.25) and (3.26), respectively, and the effectivity index for the first case f . Notice that the error has been computed with the Monte-Carlo method with a sample size $K = 1000$ using a reference solution obtained with $h_{ref} = 2^{-12}$. The results for the second case g are provided in Table 3.3.

As anticipated in the theoretical results, the efficiency of the error estimator η_ε is sensitive to the input data. Indeed, it is about 1.6 and 4.7 for the cases f and g , respectively. One remedy would be to consider an implicit error estimator for η_ε , proceeding similarly to what is done in Proposition 3.2.16 for the Navier-Stokes equations or in Proposition 2.2.3 for the model problem with random forcing term.

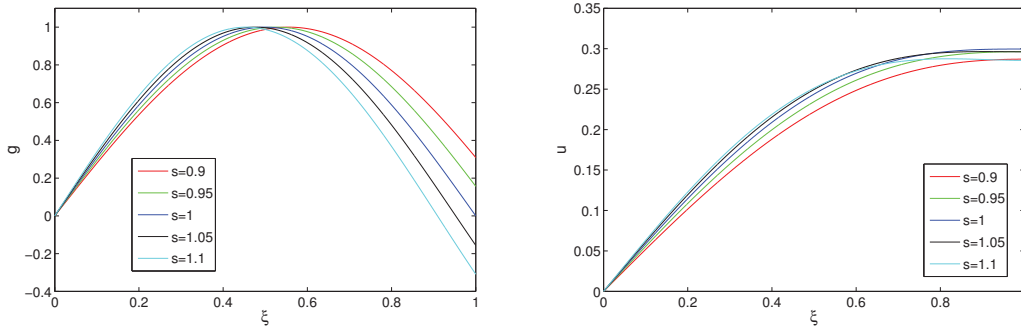


Figure 3.4: Function g and corresponding solution u for various values of s .

h	η_h	$\varepsilon = 0.005$			$\varepsilon = 0.00125$		
		η_ε	error	e.i.	η_ε	error	e.i.
1/4	4.6302e-2	1.9365e-3	1.3522e-2	3.4273	4.8413e-4	1.3469e-2	3.4379
1/8	2.3217e-2	1.9439e-3	6.8240e-3	3.4141	4.8598e-4	6.7214e-3	3.4549
1/16	1.1616e-2	1.9458e-3	3.5705e-3	3.2988	4.8644e-4	3.3693e-3	3.4508
1/32	5.8092e-3	1.9462e-3	2.0823e-3	2.9422	4.8656e-4	1.7037e-3	3.4218
1/64	2.9047e-3	1.9464e-3	1.4531e-3	2.4063	4.8659e-4	8.9101e-4	3.3055
1/128	1.4524e-3	1.9464e-3	1.2835e-3	1.8922	4.8660e-4	5.1987e-4	2.9464
1/256	7.2620e-4	1.9464e-3	1.2266e-3	1.6936	4.8660e-4	3.6041e-4	2.4254
1/512	3.6310e-4	1.9464e-3	1.1935e-3	1.6590	4.8660e-4	3.1774e-4	1.9108
1/1024	1.8155e-4	1.9464e-3	1.2322e-3	1.5865	4.8660e-4	3.0688e-4	1.6924

Table 3.2: Error, estimators and effectivity index for the Burgers' equation in random intervals for the first case f with $\varepsilon = 0.005$ and 0.00125 .

h	η_h	$\varepsilon = 0.01$			$\varepsilon = 0.0025$		
		η_ε	error	e.i.	η_ε	error	e.i.
1/4	1.6499e-1	1.7575e-2	5.0817e-2	3.2652	4.3939e-3	5.0736e-2	3.2532
1/8	8.2193e-2	1.6843e-2	2.4358e-2	3.4444	4.2108e-3	2.4133e-2	3.4103
1/16	4.1057e-2	1.6664e-2	1.2396e-2	3.5745	4.1659e-3	1.1932e-2	3.4587
1/32	2.0524e-2	1.6619e-2	6.9139e-3	3.8196	4.1548e-3	5.9977e-3	3.4914
1/64	1.0261e-2	1.6608e-2	4.6153e-3	4.2299	4.1520e-3	3.0952e-3	3.5763
1/128	5.1306e-3	1.6605e-2	3.8080e-3	4.5640	4.1513e-3	1.7249e-3	3.8261
1/256	2.5653e-3	1.6604e-2	3.6616e-3	4.5885	4.1511e-3	1.1517e-3	4.2369
1/512	1.2826e-3	1.6604e-2	3.5817e-3	4.6496	4.1510e-3	9.6613e-4	4.4970
1/1024	6.4132e-4	1.6604e-2	3.5264e-3	4.7121	4.1510e-3	8.9897e-4	4.6724

Table 3.3: Error, estimators and effectivity index for the Burgers' equation in random intervals for the second case g with $\varepsilon = 0.01$ and 0.0025 .

3.2 Steady-state incompressible Navier-Stokes equations in random domains

We consider now the steady-state incompressible Navier-Stokes equations in random domains. We start with the statement of the problem in Section 3.2.1. We introduce in Section 3.2.2 the corresponding problem on a fixed reference domain using a random mapping and show its well-posedness in Section 3.2.3 under the *small data* assumption and suitable assumptions on the mapping. A specific but rather general form of the random mapping is introduced in Section 3.2.4, namely that it depends linearly on finite (but arbitrary large) number of independent random variables. In Section 3.2.5, which is the core part, an *a posteriori* error analysis is performed with the derivation of two *a posteriori* error estimates for the first order approximation. Finally, numerical experiments are presented in Section 3.2.6 and agree with the theoretical results.

3.2.1 Problem statement

Let $D_\omega \subseteq \hat{D} \subset \mathbb{R}^d$, $d = 2, 3$, be an open bounded domain with Lipschitz continuous boundary that depends on a random parameter $\omega \in \Omega$, where \hat{D} is a fixed bounded domain that contains D_ω for all $\omega \in \Omega$. Here (Ω, \mathcal{F}, P) denotes a complete probability space, where Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ -algebra of events and $P : \mathcal{F} \rightarrow [0, 1]$ is a probability measure. By a slight abuse of notations, we will denote

$$D_\omega \times \Omega := \{(\mathbf{x}, \omega) : \mathbf{x} \in D_\omega, \omega \in \Omega\}.$$

We consider the steady incompressible Navier-Stokes equations in D_ω :

find a velocity $\tilde{\mathbf{u}} : D_\omega \times \Omega \rightarrow \mathbb{R}^d$ and a pressure $\tilde{p} : D_\omega \times \Omega \rightarrow \mathbb{R}$ such that P -almost everywhere (a.e.) in Ω , or in other words almost surely (a.s.), the following equations hold

$$\begin{cases} -\nu \Delta_{\mathbf{x}} \tilde{\mathbf{u}} + (\tilde{\mathbf{u}} \cdot \nabla_{\mathbf{x}}) \tilde{\mathbf{u}} + \nabla_{\mathbf{x}} \tilde{p} &= \tilde{\mathbf{f}} & \mathbf{x} \in D_\omega \\ \nabla_{\mathbf{x}} \cdot \tilde{\mathbf{u}} &= 0 & \mathbf{x} \in D_\omega \\ \tilde{\mathbf{u}} &= \mathbf{0} & \mathbf{x} \in \partial D_\omega, \end{cases} \quad (3.29)$$

where ν is the kinematic viscosity, $\tilde{\mathbf{f}} \in [L^2(\hat{D})]^d$ is the external force field per unit mass that we assume to be deterministic and well-defined for all $\mathbf{x} \in \hat{D}$. Note that \tilde{p} is the pressure divided by the density of the fluid. We consider homogeneous Dirichlet boundary conditions for the sake of simplicity. Should we consider non-homogeneous conditions, a lifting of the boundary conditions could be used which only modifies the right-hand side of the equations. However, the lifting has to satisfy some assumptions for the problem to be well-posed (see [116] for a complete discussion in the deterministic case). In particular, the forcing term would no longer be deterministic. In (3.29), we have used the following notation: if we write $\mathbf{x} = (x_1, \dots, x_d)$ and

$\tilde{\mathbf{u}} = (\tilde{u}_1, \dots, \tilde{u}_d)^T$ then for $i, j = 1, \dots, d$

$$\nabla_{\mathbf{x}} \tilde{p} = \left(\frac{\partial \tilde{p}}{\partial x_1}, \dots, \frac{\partial \tilde{p}}{\partial x_d} \right)^T, \quad (\nabla_{\mathbf{x}} \tilde{\mathbf{u}})_{ij} = \frac{\partial \tilde{u}_i}{\partial x_j}, \quad \nabla_{\mathbf{x}} \cdot \tilde{\mathbf{u}} = \sum_{i=1}^d \frac{\partial \tilde{u}_i}{\partial x_i}$$

and

$$(\Delta_{\mathbf{x}} \tilde{\mathbf{u}})_i = (\nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} \tilde{\mathbf{u}})_i = \sum_{j=1}^d \frac{\partial}{\partial x_j} \frac{\partial \tilde{u}_i}{\partial x_j} = \Delta_{\mathbf{x}} \tilde{u}_i, \quad [(\tilde{\mathbf{u}} \cdot \nabla_{\mathbf{x}}) \tilde{\mathbf{u}}]_i = \sum_{j=1}^d \tilde{u}_j \frac{\partial \tilde{u}_i}{\partial x_j}.$$

Note that we will use the same notation to denote the norm of a scalar, vector or matrix-valued function, with the natural extension $\|\mathbf{v}\|^2 = \sum_{i=1}^d \|v_i\|^2$ (Euclidean norm) and $\|B\|^2 = \sum_{i,j=1}^d \|B_{ij}\|^2$ (Frobenius norm) for any vector $\mathbf{v} = (v_1, \dots, v_d) \in \mathbb{R}^d$ and any matrix $B = (B_{ij})_{i,j=1}^d \in \mathbb{R}^{d \times d}$. In order to write the weak formulation of the problem, we need to introduce some functional spaces. For a given Banach space W with norm $\|\cdot\|_W$, we define the Bochner space

$$L_P^2(\Omega; W) := \{v : \Omega \rightarrow W, v \text{ is strongly measurable and } \|v\|_{L_P^2(\Omega; W)} < +\infty\},$$

where $\|v\|_{L_P^2(\Omega; W)}^2 := \int_{\Omega} \|v(\omega)\|_W^2 dP(\omega) = \mathbb{E}[\|v\|_W^2]$ using the shorthand notation $v(\omega) = v(\cdot, \omega)$ for ease of presentation. Notice that if W is a separable Hilbert space, then $L_P^2(\Omega; W)$ is isomorphic [10] to the tensor product space $L_P^2(\Omega) \otimes W$. Finally, we define $\tilde{V}_{\omega} = [H_0^1(D_{\omega})]^d$ equipped with the gradient norm $\|\cdot\|_{\tilde{V}_{\omega}} := \|\nabla_{\mathbf{x}} \cdot\|_{L^2(D_{\omega})}$ and $\tilde{Q}_{\omega} = L^2(D_{\omega})$. Note that unless otherwise clearly stated, the Lebesgue measure is used in D_{ω} . The (pointwise in ω) weak formulation of problem (3.29) reads:

find $(\tilde{\mathbf{u}}(\omega), \tilde{p}(\omega)) \in \tilde{V}_{\omega} \times \tilde{Q}_{\omega}$ such that

$$\begin{cases} \nu \int_{D_{\omega}} \nabla_{\mathbf{x}} \tilde{\mathbf{u}} : \nabla_{\mathbf{x}} \tilde{\mathbf{v}} d\mathbf{x} + \int_{D_{\omega}} [(\tilde{\mathbf{u}} \cdot \nabla_{\mathbf{x}}) \tilde{\mathbf{u}}] \cdot \tilde{\mathbf{v}} d\mathbf{x} - \int_{D_{\omega}} \tilde{p} \nabla_{\mathbf{x}} \cdot \tilde{\mathbf{v}} d\mathbf{x} = \int_{D_{\omega}} \tilde{\mathbf{f}} \cdot \tilde{\mathbf{v}} d\mathbf{x} \\ - \int_{D_{\omega}} \tilde{q} \nabla_{\mathbf{x}} \cdot \tilde{\mathbf{u}} d\mathbf{x} = 0 \end{cases} \quad (3.30)$$

for all $(\tilde{\mathbf{v}}, \tilde{q}) \in \tilde{V}_{\omega} \times \tilde{Q}_{\omega}$ and a.s. in Ω . Since we impose Dirichlet conditions on the whole boundary, the pressure is only defined up to an additive constant. We come back to this point in the next section (see Remark 3.2.1). Under the assumption of *small data*, the well-posedness of the problem on the family of random domains $(D_{\omega})_{\omega \in \Omega}$ can be proved using two different approaches. The first one would be to consider the Navier-Stokes equations directly on $D_{\omega} \times \Omega$. Another approach, adopted here, consists in mapping the random domain to a reference one, yielding PDEs on a (fixed, deterministic) reference domain with random coefficients.

3.2.2 Formulation on a reference domain

Let $D \subset \mathbb{R}^d$ be an open bounded reference domain with Lipschitz continuous boundary ∂D . We assume that there exists a mapping $\mathbf{x} : D \times \Omega \rightarrow \mathbb{R}^d$ that transforms D into D_{ω} : for each

3.2. Steady-state incompressible Navier-Stokes equations in random domains

$\omega \in \Omega$

$$\begin{aligned} \mathbf{x}_\omega : D &\rightarrow D_\omega \\ \xi &\mapsto \mathbf{x} = \mathbf{x}_\omega(\xi) \end{aligned}$$

where the notation $\mathbf{x}_\omega(\xi)$ stands for $\mathbf{x}(\xi, \omega)$. We assume that for any $\omega \in \Omega$, \mathbf{x}_ω is invertible and sufficiently regular so that everything that follows makes sense, the precise regularity assumptions on the random mapping \mathbf{x} being given in Section 3.2.3. Let ξ_ω be the inverse of \mathbf{x}_ω defined by

$$\begin{aligned} \xi_\omega : D_\omega &\rightarrow D \\ \mathbf{x} &\mapsto \xi = \xi_\omega(\mathbf{x}). \end{aligned}$$

We also introduce the $d \times d$ Jacobian matrices $A^{-1} = A^{-1}(\xi, \omega)$ and $\tilde{A} = \tilde{A}(\mathbf{x}, \omega)$ corresponding respectively to the random transformations \mathbf{x}_ω and ξ_ω and defined by

$$A^{-1} = \left(A_{ij}^{-1} \right)_{1 \leq i, j \leq d} \quad \text{with} \quad A_{ij}^{-1} := \frac{\partial (\mathbf{x}_\omega)_i}{\partial \xi_j}$$

and

$$\tilde{A} = \left(\tilde{A}_{ij} \right)_{1 \leq i, j \leq d} \quad \text{with} \quad \tilde{A}_{ij} := \frac{\partial (\xi_\omega)_i}{\partial x_j}.$$

We mention that the matrix A^{-1} is often denoted F in the continuum mechanics literature. For any function \tilde{g} defined on $D_\omega \times \Omega$, we denote by $g = \tilde{g} \circ \mathbf{x}_\omega$ its corresponding function on $D \times \Omega$, i.e. $g(\xi, \omega) = \tilde{g}(\mathbf{x}, \omega)$ with $\mathbf{x} = \mathbf{x}_\omega(\xi)$. Notice that the matrix $A = \tilde{A} \circ \mathbf{x}_\omega$ is the inverse (in the matrix sense) of A^{-1} . From the chain rule, the following relations hold true

$$\nabla_{\mathbf{x}} = \tilde{A}^T \nabla_{\xi} \quad \text{and} \quad \nabla_{\mathbf{x}} \tilde{\mathbf{u}} = (\nabla_{\xi} \mathbf{u} \circ \xi_\omega) \tilde{A},$$

where $\tilde{A}^T \nabla_{\xi}$ is a matrix-vector product. For the sake of notation, we will write ∇ instead of ∇_{ξ} from now on and use the notation

$$[(B\nabla)p]_i = \sum_{j=1}^d B_{ij} \frac{\partial p}{\partial \xi_j}, \quad (B\nabla) \cdot \mathbf{u} = \sum_{i,j=1}^d B_{ij} \frac{\partial u_i}{\partial \xi_j} = B : \nabla \mathbf{u}$$

and

$$[(B\nabla)\mathbf{u}]_{ij} = \sum_{k=1}^d B_{jk} \frac{\partial u_i}{\partial \xi_k}, \quad [(\mathbf{u} \cdot B\nabla)\mathbf{v}]_i = \sum_{j,k=1}^d u_j B_{jk} \frac{\partial v_i}{\partial \xi_k}$$

for a $d \times d$ matrix $B = (B_{ij})_{1 \leq i, j \leq d}$. Note that $(A\nabla)p = A(\nabla p)$. Moreover, let $J_{\mathbf{x}} = \det(A^{-1})$ denotes the determinant of the Jacobian matrix A^{-1} associated to \mathbf{x}_ω . Finally, we introduce the spaces $V = [H_0^1(D)]^d$ and $Q = L_0^2(D) = \{q \in L^2(D) : \int_D q d\xi = 0\}$.

Remark 3.2.1. We choose to fix the constant part of the pressure by imposing zero average on D and not on D_ω , the goal being not to estimate this constant when performing the error analysis. Notice that if we fix \tilde{p} with zero average on D_ω , then the average of the corresponding pressure $p = \tilde{p} \circ \mathbf{x}_\omega$ on D would be small when \mathbf{x}_ω is a small perturbation of the identity mapping. Indeed, we have $\int_D p d\xi = \int_D p d\xi - \int_{D_\omega} \tilde{p} d\mathbf{x} = \int_D p(1 - |J_{\mathbf{x}}|) d\xi$.

Chapter 3. PDEs in random domains

We are now able to write the weak formulation of problem (3.29) on the reference domain, using the change of variable $\mathbf{x} = \mathbf{x}_\omega(\xi)$:

find $(\mathbf{u}(\omega), p(\omega)) \in V \times Q$ such that

$$\begin{cases} a(\mathbf{u}, \mathbf{v}; \omega) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}; \omega) + b(\mathbf{v}, p; \omega) &= F(\mathbf{v}; \omega) \\ b(\mathbf{u}, q; \omega) &= 0 \end{cases} \quad (3.31)$$

for all $(\mathbf{v}, q) \in V \times Q$ and a.s. in Ω , where

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}; \omega) &:= \nu \int_D (\nabla \mathbf{u} A(\omega)) : (\nabla \mathbf{v} A(\omega)) J_{\mathbf{x}}(\omega) d\xi \\ b(\mathbf{v}, q; \omega) &:= - \int_D q J_{\mathbf{x}}(\omega) (A(\omega)^T \nabla) \cdot \mathbf{v} d\xi \\ c(\mathbf{u}, \mathbf{v}, \mathbf{w}; \omega) &:= \int_D [(\mathbf{u} \cdot A(\omega)^T \nabla) \mathbf{v}] \cdot \mathbf{w} J_{\mathbf{x}}(\omega) d\xi \\ F(\mathbf{v}; \omega) &:= \int_D \mathbf{f}(\omega) \cdot \mathbf{v} J_{\mathbf{x}}(\omega) d\xi. \end{aligned} \quad (3.32)$$

Using the relations (see Appendix 3.C for proofs)

$$(\nabla \mathbf{u} A) : (\nabla \mathbf{u} A) = (\nabla \mathbf{u} A A^T) : (\nabla \mathbf{u}), \quad \nabla \mathbf{u} A = (A^T \nabla) \mathbf{u} \quad (3.33)$$

and

$$- \int_D q J_{\mathbf{x}}(A^T \nabla) \cdot \mathbf{v} d\xi = \int_D J_{\mathbf{x}}(A^T \nabla q) \cdot \mathbf{v} d\xi, \quad (3.34)$$

the strong form of (3.31) can be written:

find $\mathbf{u} : D \times \Omega \rightarrow \mathbb{R}^d$ and $p : D \times \Omega \rightarrow \mathbb{R}$ such that P -almost everywhere in Ω there holds:

$$\begin{cases} -\nu \nabla \cdot [J_{\mathbf{x}} A A^T \nabla] \mathbf{u} + (\mathbf{u} \cdot J_{\mathbf{x}} A^T \nabla) \mathbf{u} + (J_{\mathbf{x}} A^T \nabla) p &= \mathbf{f}_{\mathbf{x}} & \xi \in D \\ (J_{\mathbf{x}} A^T \nabla) \cdot \mathbf{u} &= 0 & \xi \in D \\ \mathbf{u} &= \mathbf{0} & \xi \in \partial D. \end{cases} \quad (3.35)$$

Notice that similarly to the formulation in [71], the continuity equation can be equivalently written $\nabla \cdot (J_{\mathbf{x}} A \mathbf{u})$ thanks to Piola's identity (see Appendix 3.C).

Remark 3.2.2. If homogeneous Neumann boundary conditions $\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - \tilde{p} \tilde{\mathbf{n}} = \mathbf{0}$ are prescribed for problem (3.29) on a part of the boundary ∂D_ω , typically at the outflow part of the boundary, the corresponding boundary conditions for the problem on the reference domain D read $\nu J_{\mathbf{x}} \nabla \mathbf{u} A A^T \mathbf{n} - p J_{\mathbf{x}} A^T \mathbf{n} = \mathbf{0}$. However, the problem might no longer be well-posed due to the loss of (uniform) coercivity of $a(\cdot, \cdot; \omega) + c(\cdot, \cdot; \omega)$ or its counter part on D_ω . Indeed, we are not able to control the negative part of the boundary integral. Braack and al. proved in [29] the existence and uniqueness of a solution to the Navier-Stokes equations with small data and homogeneous Neumann conditions on a part of the boundary after introducing what they called a directed-do-nothing condition, adding a (boundary integral) term in the weak

formulation of the problem. From a physical point of view, a force per unit area is prescribed by imposing $v(\nabla_{\mathbf{x}}\tilde{\mathbf{u}} + (\nabla_{\mathbf{x}}\tilde{\mathbf{u}})^T)\tilde{\mathbf{n}} - \tilde{p}\tilde{\mathbf{n}} = \tilde{\mathbf{g}}$, corresponding to $vJ_{\mathbf{x}}(\nabla\mathbf{u}A + (\nabla\mathbf{u}A)^T)A^T\mathbf{n} - pJ_{\mathbf{x}}A^T\mathbf{n} = \mathbf{g}$ on the reference domain. In such a case, $\Delta_{\mathbf{x}}\tilde{\mathbf{u}}$ in (3.29) should be replaced by $\nabla_{\mathbf{x}} \cdot (\nabla_{\mathbf{x}}\tilde{\mathbf{u}} + (\nabla_{\mathbf{x}}\tilde{\mathbf{u}})^T)$.

3.2.3 Well-posedness of the problem

The goal is now to show the well-posedness of problem (3.29), under suitable conditions on the family of random mapping $(\mathbf{x}_{\omega})_{\omega \in \Omega}$ and restriction on the input data. We will show that there exists a unique solution (\mathbf{u}, p) to problem (3.31), the weak solution of problem (3.29) being then given by $(\tilde{\mathbf{u}}, \tilde{p}) = (\mathbf{u} \circ \xi_{\omega}, p \circ \xi_{\omega})$.

For any $\omega \in \Omega$, we assume that $\mathbf{x}_{\omega} : D \rightarrow D_{\omega}$, with $D_{\omega} = \mathbf{x}_{\omega}(D)$, is a one-to-one mapping such that $\mathbf{x}_{\omega} \in [W^{1,\infty}(D)]^d$, $\xi_{\omega} \in [W^{1,\infty}(D_{\omega})]^d$ and D_{ω} is bounded with Lipschitz continuous boundary ∂D_{ω} . Since \mathbf{x}_{ω} is invertible, the determinant $J_{\mathbf{x}}$ of its Jacobian matrix A^{-1} does not vanish. Without loss of generality, we can assume that $J_{\mathbf{x}} > 0$, namely that the mapping is orientation-preserving. Moreover, we make the following assumption [43, 76] on the singular values σ_i of A^{-1} : there exist two constants $\sigma_{\min}, \sigma_{\max}$ such that for $i = 1, \dots, d$

$$0 < \sigma_{\min} \leq \sigma_i(A^{-1}(\xi, \omega)) \leq \sigma_{\max} < \infty \quad \text{a.e. in } D \text{ and a.s. in } \Omega. \quad (3.36)$$

Notice that the singular values of A are then bounded uniformly from below and above by σ_{\max}^{-1} and σ_{\min}^{-1} , respectively. Therefore, the random mapping \mathbf{x} have finite moment of any order and with the above regularity assumption we have $\mathbf{x} \in L_p^{\infty}(\Omega; [W^{1,\infty}(D)]^d)$. Moreover, the following properties are immediate consequences of assumption (3.36).

Proposition 3.2.3. *Under assumption (3.36), we have a.e. in D and a.s. in Ω*

- $\sigma_{\min}^d \leq \det(A^{-1}) \leq \sigma_{\max}^d$,
- $\sigma_{\max}^{-2} \leq \lambda_i(AA^T) \leq \sigma_{\min}^{-2}$ for $i = 1, \dots, d$,

where $\lambda_i(AA^T)$, $i = 1, \dots, d$, denote the eigenvalues of AA^T .

Proof. Since the eigenvalues of $A^{-1}A^{-T}$ (and thus of the so-called (right) Cauchy-Green strain tensor $A^{-T}A^{-1}$) are the square of the singular values of A^{-1} , the first relation follows directly from (3.36) and the fact that

$$\det(A^{-1}) = \sqrt{\det(A^{-1}A^{-T})} = \sqrt{\prod_{i=1}^d \lambda_i(A^{-1}A^{-T})} = \prod_{i=1}^d \sigma_i(A^{-1}).$$

The second relation is just a consequence of $\lambda_i(AA^T) = \sigma_i(A)^2$. □

The following proposition ensures that the spaces $L^2(D_{\omega})$ and $L^2(D)$, respectively $[H_0^1(D_{\omega})]^d$ and $[H_0^1(D)]^d$, are isomorphic.

Proposition 3.2.4. *Under assumption (3.36), for any $\tilde{g} \in L^2(D_\omega)$ and $\tilde{\mathbf{v}} \in [H^1(D_\omega)]^d$ we have a.s. in Ω*

$$\sigma_{\min}^{\frac{d}{2}} \|g\|_{L^2(D)} \leq \|\tilde{g}\|_{L^2(D_\omega)} \leq \sigma_{\max}^{\frac{d}{2}} \|g\|_{L^2(D)} \quad (3.37)$$

and

$$\frac{\sigma_{\min}^{\frac{d}{2}}}{\sigma_{\max}} \|\nabla \mathbf{v}\|_{L^2(D)} \leq \|\nabla_{\mathbf{x}} \tilde{\mathbf{v}}\|_{L^2(D_\omega)} \leq \frac{\sigma_{\max}^{\frac{d}{2}}}{\sigma_{\min}} \|\nabla \mathbf{v}\|_{L^2(D)} \quad (3.38)$$

with $g = \tilde{g} \circ \mathbf{x}_\omega$ and $\mathbf{v} = \tilde{\mathbf{v}} \circ \mathbf{x}_\omega$. The same relations hold true for any $g \in L^2(D)$ and $\mathbf{v} \in [H^1(D)]^d$ with $\tilde{g} = g \circ \xi_\omega$ and $\tilde{\mathbf{v}} = \mathbf{v} \circ \xi_\omega$.

Proof. Let $\tilde{g} \in L^2(D_\omega)$ and $\tilde{\mathbf{v}} \in [H^1(D_\omega)]^d$. The proof of (3.37) is immediate using the uniform bounds on $\det(A^{-1})$ given by Proposition 3.2.3. For (3.38), we use the fact that $\sigma_{\min}^d \sigma_{\max}^{-2}$ and $\sigma_{\max}^d \sigma_{\min}^{-2}$ are uniform bounds for the eigenvalues (or equivalently singular values) of the symmetric positive definite matrix $\det(A^{-1}) A A^T$ and the relation

$$\|\nabla_{\mathbf{x}} \tilde{\mathbf{u}}\|_{L^2(D_\omega)}^2 = \int_D (\nabla \mathbf{u} A) : (\nabla \mathbf{u} A) \det(A^{-1}) d\xi = \int_D \sum_{i=1}^d (\det(A^{-1}) A A^T \nabla u_i) \cdot \nabla u_i d\xi.$$

The proof of (3.37) and (3.38) for the case $g \in L^2(D)$ and $\mathbf{v} \in [H^1(D)]^d$ is similar using the relations $\sigma_{\max}^{-d} \leq \det(A) \leq \sigma_{\min}^{-d}$ and $\sigma_{\max}^{-2} \sigma_{\min}^2 \leq \lambda_i(\det(A) A^{-1} A^{-T}) \leq \sigma_{\min}^{-2} \sigma_{\max}^2$ a.e. in D and a.s. in Ω . \square

To show the well-posedness of problem (3.31), the forms a , b and c defined in (3.32) have to satisfy (uniformly) some properties, which we verify in the following proposition.

Proposition 3.2.5. *For any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and any $q \in L^2(D)$ we have a.s. in Ω*

- *a is continuous:* $|a(\mathbf{u}, \mathbf{v}; \omega)| \leq \nu M \|\nabla \mathbf{u}\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)}$ with $M = \sigma_{\min}^{-2} \sigma_{\max}^d$,
- *a is coercive:* $a(\mathbf{v}, \mathbf{v}; \omega) \geq \nu \alpha \|\nabla \mathbf{v}\|_{L^2(D)}^2$ with $\alpha = \sigma_{\max}^{-2} \sigma_{\min}^d$,
- *b is continuous:* $|b(\mathbf{v}, q; \omega)| \leq \sigma_{\max}^d \sigma_{\min}^{-1} \|q\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)}$,
- *c is continuous:* $|c(\mathbf{u}, \mathbf{v}, \mathbf{w}; \omega)| \leq \hat{C} \|\nabla \mathbf{u}\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)} \|\nabla \mathbf{w}\|_{L^2(D)}$ with $\hat{C} = C_I^2 \sigma_{\max}^d \sigma_{\min}^{-1}$,

where $C_I = C_I(D)$ is the constant in $\|\mathbf{v}\|_{L^4(D)} \leq C_I \|\nabla \mathbf{v}\|_{L^2(D)}$, resulting from Sobolev embedding's theorem and Poincaré's inequality on D .

Proof. The proof is immediate from Proposition 3.2.3, Hölder's inequality and the Sobolev embedding theorem. The relation (see e.g. [106])

$$\int_D (\nabla \cdot \mathbf{v})(\nabla \cdot \mathbf{v}) d\xi + \int_D (\nabla \times \mathbf{v}) \cdot (\nabla \times \mathbf{v}) d\xi = \int_D \nabla \mathbf{v} : \nabla \mathbf{v} d\xi \quad \forall \mathbf{v} \in V,$$

3.2. Steady-state incompressible Navier-Stokes equations in random domains

where $\nabla \times \mathbf{v}$ denotes the *curl* of \mathbf{v} , is used to prove the continuity of b .

□

Notice that we do not include the parameter ν in the constants α and M linked to the coercivity and continuity of a , respectively, because we will track its occurrence in the derivation of our *a posteriori* error estimates, the goal being to minimize the sensitivity of the effectivity index with respect to ν . We mention that b is also continuous on $[H^1(D)]^d$ with the same constant as in Proposition 3.2.5 up to a multiplication by a factor \sqrt{d} and satisfies the so-called (Brezzi [32]) inf-sup condition $\inf_{q \in Q} \sup_{\mathbf{v} \in V} \frac{b(\mathbf{v}, q; \omega)}{\|q\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)}} \geq \frac{\sigma_{\min}}{\sigma_{\max}^d} \beta_\omega > 0$ for any $\omega \in \Omega$ since D_ω is a Lipschitz domain. Moreover, we assume that there exists a constant $\beta > 0$ such that the inf-sup condition holds uniformly with respect to ω , i.e.

$$\inf_{q \in Q} \sup_{\mathbf{v} \in V} \frac{b(\mathbf{v}, q; \omega)}{\|q\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)}} \geq \beta \quad \text{a.s. in } \Omega. \quad (3.39)$$

Remark 3.2.6. The inf-sup condition (3.39) can be easily shown under the assumption that the mapping satisfies $\mathbf{x} \in L_p^\infty(\Omega; [W^{2,\infty}(D)]^d)$, proceeding similarly to [71]. Indeed, for any $q \in Q$ there exists $\mathbf{z} \in V$ such that $\nabla \cdot \mathbf{z} = q$ and $\|\nabla \mathbf{z}\|_{L^2(D)} \leq C_1 \|q\|_{L^2(D)}$ with a constant C_1 depending only on the reference domain D , see for instance [69]. Setting $\mathbf{v} = -(J_{\mathbf{x}} A)^{-1} \mathbf{z}$ we get a.s. in Ω

$$b(\mathbf{v}, q; \omega) = \|q\|_{L^2(D)}^2 \geq \frac{1}{C_1} \|q\|_{L^2(D)} \|\nabla \mathbf{z}\|_{L^2(D)} \quad \text{and} \quad \|\nabla \mathbf{v}\|_{L^2(D)} \leq C_2 \|(J_{\mathbf{x}} A)^{-1}\|_{W^{1,\infty}(D)} \|\nabla \mathbf{z}\|_{L^2(D)},$$

where C_2 depends only on the Poincaré constant on D . From these two inequalities, we deduce that $\frac{b(\mathbf{v}, q; \omega)}{\|\nabla \mathbf{v}\|_{L^2(D)}} \geq \beta \|q\|_{L^2(D)}$ a.s. in Ω with $\beta^{-1} = C_1 C_2 \|(J_{\mathbf{x}} A)^{-1}\|_{L_p^\infty(\Omega; [W^{1,\infty}(D)]^{d \times d})}$.

Let us introduce the subspace $\tilde{V}_{\text{div}, \omega} \subset \tilde{V}_\omega$ constituted of all (weakly) divergence-free functions of \tilde{V}_ω , and its counterpart on D given by

$$V_{\text{div}, \omega} := \{\mathbf{v} \in V : b(\mathbf{v}, q; \omega) = 0 \quad \forall q \in Q, \text{ a.s. in } \Omega\}.$$

We can then formulate the (reduced, pointwise in ω) weak formulation of problem (3.31):

find $\mathbf{u}(\omega) \in V_{\text{div}, \omega}$ such that

$$a(\mathbf{u}, \mathbf{v}; \omega) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}; \omega) = F(\mathbf{v}; \omega) \quad \forall \mathbf{v} \in V_{\text{div}, \omega}, \text{ a.s. in } \Omega. \quad (3.40)$$

Proposition 3.2.7. For $\mathbf{u}(\omega) \in V_{\text{div}, \omega}$ solution of (3.40), there exists a unique pressure $p(\omega) \in Q$ so that (\mathbf{u}, p) is a solution of (3.31), a.s in Ω .

Proof. Follows from the inf-sup condition (see [69, p.283]).

□

Therefore, to show the well-posedness of problem (3.31), and thus of the original problem

(3.30), it only remains to prove that the nonlinear problem (3.40) admits a unique solution. Recalling that F is defined in (3.32) with $\mathbf{f} = \tilde{\mathbf{f}} \circ \mathbf{x}_\omega$, the following proposition gives a sufficient condition on the input data so that problem (3.40) is well-posed.

Proposition 3.2.8. *If there exists $\theta \in [0, 1)$ such that*

$$\frac{C_P C_I^2 \sigma_{\max}^{\frac{3d}{2}+4}}{\nu^2 \sigma_{\min}^{2d+1}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \leq \theta < 1 \quad \text{a.s. in } \Omega, \quad (3.41)$$

where $C_P = C_P(D)$ denotes the Poincaré constant on D , then problem (3.40) has a unique solution. Moreover, its solution satisfies

$$\|\nabla \mathbf{u}(\omega)\|_{L^2(D)} \leq \theta \frac{\nu \sigma_{\min}^{d+1}}{C_I^2 \sigma_{\max}^{d+2}} = \theta \frac{\nu \alpha}{\hat{C}} \quad \text{a.s. in } \Omega, \quad (3.42)$$

with α and \hat{C} defined in Proposition 3.2.5.

Remark 3.2.9. Notice that if condition (3.41) holds, then $\frac{\hat{C}}{(\nu \alpha)^2} \|F(\cdot; \omega)\|_{V'_{\text{div}, \omega}} < 1$ a.s. in Ω , where the norm on the dual space is defined in the usual way, which is nothing else but the standard small data assumption for uniqueness (see e.g. [60, 69, 116]). Indeed, we have

$$\frac{\hat{C}}{(\nu \alpha)^2} \|F(\cdot; \omega)\|_{V'_{\text{div}, \omega}} = \frac{\hat{C}}{(\nu \alpha)^2} \sup_{\mathbf{v} \in V_{\text{div}, \omega}} \frac{|F(\mathbf{v}; \omega)|}{\|\nabla \mathbf{v}\|_{L^2(D)}} \leq \frac{C_P C_I^2 \sigma_{\max}^{\frac{3d}{2}+4}}{\nu^2 \sigma_{\min}^{2d+1}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \quad \text{a.s. in } \Omega,$$

where for the last inequality we used the relation

$$|F(\mathbf{v}; \omega)| \leq \sigma_{\max}^{\frac{d}{2}} \|\mathbf{f} J_{\mathbf{x}}^{\frac{1}{2}}\|_{L^2(D)} \|\mathbf{v}\|_{L^2(D)} \leq C_P \sigma_{\max}^{\frac{d}{2}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \|\nabla \mathbf{v}\|_{L^2(D)} \quad \text{a.s. in } \Omega.$$

Moreover, instead of (3.41), we could impose that

$$\frac{C_P C_I^2 \sigma_{\max}^{2(d+2)}}{\nu^2 \sigma_{\min}^{2d+1}} \|\mathbf{f}(\omega)\|_{L^2(D)} \leq \theta < 1 \quad \text{a.s. in } \Omega \quad (3.43)$$

since $\|\mathbf{f}(\omega)\|_{L^2(D)} \geq \sigma_{\max}^{-\frac{d}{2}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)}$ by Proposition 3.2.4, and thus (3.43) implies (3.41).

The proof of Proposition 3.2.8 follows the same procedure as the one proposed in [109] for deterministic steady Navier-Stokes equations in a given domain and is based on a fixed point argument.

Proof. In this proof, the explicit dependence of the functions with respect to $\omega \in \Omega$ will not necessarily be indicated, unless ambiguity holds. Moreover, with little abuse of notation we define the space

$$L_P^2(\Omega; V_{\text{div}, \omega}) := \{\mathbf{v} \in L_P^2(\Omega; V) : \mathbf{v}(\omega) \in V_{\text{div}, \omega} \text{ a.s. in } \Omega\}.$$

3.2. Steady-state incompressible Navier-Stokes equations in random domains

First of all, we can show that

$$c(\mathbf{u}, \mathbf{v}, \mathbf{v}; \omega) = 0 \quad \forall \mathbf{u} \in V_{\text{div}, \omega}, \forall \mathbf{v} \in V, \text{ a.s. in } \Omega. \quad (3.44)$$

Indeed, if we write $\tilde{\mathbf{u}} = \mathbf{u} \circ \xi_\omega$ and $\tilde{\mathbf{v}} = \mathbf{v} \circ \xi_\omega$ then $\tilde{\mathbf{u}} \in \tilde{V}_{\text{div}, \omega}$, $\tilde{\mathbf{v}} \in \tilde{V}_\omega$ and

$$\begin{aligned} c(\mathbf{u}, \mathbf{v}, \mathbf{v}; \omega) &= \int_D [(\mathbf{u} \cdot A^T \nabla) \mathbf{v}] \cdot \mathbf{v} J_{\mathbf{x}} d\xi = \int_{D_\omega} [(\tilde{\mathbf{u}} \cdot \nabla_{\mathbf{x}}) \tilde{\mathbf{v}}] \cdot \tilde{\mathbf{v}} d\mathbf{x} \\ &= -\frac{1}{2} \int_{D_\omega} (\nabla_{\mathbf{x}} \cdot \tilde{\mathbf{u}}) |\tilde{\mathbf{v}}|^2 d\mathbf{x} + \frac{1}{2} \int_{\partial D_\omega} (\tilde{\mathbf{u}} \cdot \mathbf{n}) |\tilde{\mathbf{v}}|^2 ds = 0 \end{aligned}$$

using the fact that we have imposed homogeneous Dirichlet boundary conditions. Now, for any $\mathbf{u} \in L_P^2(\Omega; V_{\text{div}})$ we define the (pointwise in ω) bilinear form $\mathcal{A}_{\mathbf{u}(\omega)}(\cdot, \cdot; \omega) : V_{\text{div}, \omega} \times V_{\text{div}, \omega} \rightarrow \mathbb{R}$ by

$$\mathcal{A}_{\mathbf{u}(\omega)}(\mathbf{w}, \mathbf{v}; \omega) := a(\mathbf{w}, \mathbf{v}; \omega) + c(\mathbf{u}(\omega), \mathbf{w}, \mathbf{v}; \omega),$$

which is uniformly continuous and coercive (on V and thus on $V_{\text{div}, \omega}$) thanks to Proposition 3.2.5 and relation (3.44). Since $\|\mathbf{f} J_{\mathbf{x}}\|_{L^2(D)} \leq \sigma_{\max}^{d/2} \|\tilde{\mathbf{f}}\|_{L^2(\hat{D})} < +\infty$ a.s. in Ω , in particular $\mathbf{f} J_{\mathbf{x}} \in L_P^2(\Omega; L^2(D))$ and Lax-Milgram's lemma ensures the existence of a unique solution to the problem:

for every $\omega \in \Omega$, find $\mathbf{w}(\omega) \in V_{\text{div}, \omega}$ such that

$$\mathcal{A}_{\mathbf{u}(\omega)}(\mathbf{w}, \mathbf{v}; \omega) = F(\mathbf{v}; \omega) \quad \forall \mathbf{v} \in V_{\text{div}, \omega}, \text{ a.s. in } \Omega. \quad (3.45)$$

Moreover, taking $\mathbf{v} = \mathbf{w}(\omega)$ in (3.45) and using the coercivity of $\mathcal{A}_{\mathbf{u}}(\cdot, \cdot; \omega)$ we have a.s. in Ω

$$\nu \sigma_{\min}^d \sigma_{\max}^{-2} \|\nabla \mathbf{w}\|_{L^2(D)}^2 \leq \mathcal{A}_{\mathbf{u}}(\mathbf{w}, \mathbf{w}; \omega) = F(\mathbf{w}; \omega) \leq C_P \sigma_{\max}^{d/2} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \|\nabla \mathbf{w}\|_{L^2(D)}$$

and thus

$$\|\nabla \mathbf{w}\|_{L^2(D)} \leq \frac{C_P \sigma_{\max}^{d/2+2}}{\nu \sigma_{\min}^d} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \leq \frac{C_P \sigma_{\max}^{d/2+2}}{\nu \sigma_{\min}^d} \|\tilde{\mathbf{f}}\|_{L^2(\hat{D})} < \infty \quad (3.46)$$

from which we deduce that $\mathbf{w} \in L_P^2(\Omega; V_{\text{div}, \omega})$. Notice that a fixed point of the application $\Phi : L_P^2(\Omega; V_{\text{div}, \omega}) \rightarrow L_P^2(\Omega; V_{\text{div}, \omega})$, which maps \mathbf{u} to the unique solution \mathbf{w} of (3.45), is a solution of problem (3.40). Therefore, it only remains to prove that Φ is a strict contraction. Let $\mathbf{w} = \Phi(\mathbf{u})$ with $\mathbf{u} \in L_P^2(\Omega; V_{\text{div}, \omega})$. First, using relation (3.46) we directly get that $\Phi(L_P^2(\Omega; V_{\text{div}})) \subset \mathcal{M}$, where the ball $\mathcal{M} \subset L_P^2(\Omega; V_{\text{div}, \omega})$ is defined by

$$\mathcal{M} := \{\mathbf{v} \in L_P^2(\Omega; V_{\text{div}, \omega}) : \|\nabla \mathbf{v}\|_{L^2(D)} \leq \frac{C_P \sigma_{\max}^{d/2+2}}{\nu \sigma_{\min}^d} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \text{ a.s. in } \Omega\}.$$

Finally, we show that Φ is a contraction, i.e. that there exists a constant $0 < k < 1$ such that

$$\|\Phi(\mathbf{u}) - \Phi(\tilde{\mathbf{u}})\|_{L_P^2(\Omega; V)} \leq k \|\mathbf{u} - \tilde{\mathbf{u}}\|_{L_P^2(\Omega; V)} \quad \forall \mathbf{u}, \tilde{\mathbf{u}} \in L_P^2(\Omega; V_{\text{div}, \omega}).$$

Chapter 3. PDEs in random domains

Let $\mathbf{w} = \Phi(\mathbf{u})$ and $\bar{\mathbf{w}} = \Phi(\bar{\mathbf{u}})$. Since \mathbf{w} and $\bar{\mathbf{w}}$ satisfy problem (3.45) with $\mathcal{A}_{\mathbf{u}}(\cdot, \cdot; \omega)$ and $\mathcal{A}_{\bar{\mathbf{u}}}(\cdot, \cdot; \omega)$, respectively, we have

$$a(\mathbf{w} - \bar{\mathbf{w}}, \mathbf{v}; \omega) + c(\mathbf{u}, \mathbf{w}, \mathbf{v}; \omega) - c(\bar{\mathbf{u}}, \bar{\mathbf{w}}, \mathbf{v}; \omega) = 0 \quad \forall \mathbf{v} \in V_{\text{div}, \omega}, \text{ a.s. in } \Omega,$$

from which we deduce

$$a(\mathbf{w} - \bar{\mathbf{w}}, \mathbf{v}; \omega) + c(\mathbf{u} - \bar{\mathbf{u}}, \bar{\mathbf{w}}, \mathbf{v}; \omega) + c(\mathbf{u}, \mathbf{w} - \bar{\mathbf{w}}, \mathbf{v}; \omega) = 0,$$

or in other words

$$\mathcal{A}_{\mathbf{u}}(\mathbf{w} - \bar{\mathbf{w}}, \mathbf{v}; \omega) = -c(\mathbf{u} - \bar{\mathbf{u}}, \bar{\mathbf{w}}, \mathbf{v}; \omega).$$

Since $\bar{\mathbf{w}} \in \mathcal{M}$, taking $\mathbf{v} = \mathbf{w} - \bar{\mathbf{w}}$ in the last equation yields a.s. in Ω

$$\begin{aligned} \nu \sigma_{\min}^d \sigma_{\max}^{-2} \|\nabla(\mathbf{w} - \bar{\mathbf{w}})\|_{L^2(D)}^2 &\leq \mathcal{A}_{\mathbf{u}}(\mathbf{w} - \bar{\mathbf{w}}, \mathbf{w} - \bar{\mathbf{w}}; \omega) = -c(\mathbf{u} - \bar{\mathbf{u}}, \bar{\mathbf{w}}, \mathbf{w} - \bar{\mathbf{w}}; \omega) \\ &\leq C_I^2 \sigma_{\max}^d \sigma_{\min}^{-1} \|\nabla(\mathbf{u} - \bar{\mathbf{u}})\|_{L^2(D)} \|\nabla \bar{\mathbf{w}}\|_{L^2(D)} \|\nabla(\mathbf{w} - \bar{\mathbf{w}})\|_{L^2(D)} \\ &\leq \frac{C_P C_I^2 \sigma_{\max}^{\frac{3d}{2}+2}}{\nu \sigma_{\min}^{d+1}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \|\nabla(\mathbf{u} - \bar{\mathbf{u}})\|_{L^2(D)} \|\nabla(\mathbf{w} - \bar{\mathbf{w}})\|_{L^2(D)}. \end{aligned}$$

Therefore

$$\|\nabla(\mathbf{w} - \bar{\mathbf{w}})\|_{L^2(D)} \leq \frac{C_P C_I^2 \sigma_{\max}^{\frac{3d}{2}+4}}{\nu^2 \sigma_{\min}^{2d+1}} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \|\nabla(\mathbf{u} - \bar{\mathbf{u}})\|_{L^2(D)} \quad \text{a.s. in } \Omega$$

which proves that Φ is a contraction under the assumption that (3.41) holds. By the Banach contraction theorem, we know that there exists a unique fixed point $\mathbf{u} = \Phi(\mathbf{u})$, which is solution of problem (3.40). The fact that any solution of (3.40) is in \mathcal{M} and is a fixed point of Φ achieves the proof of well-posedness of the problem. Finally, recalling that α and \hat{C} are defined in Proposition 3.2.5, the bound (3.42) is immediate since

$$\|\nabla \mathbf{u}\|_{L^2(D)} \leq \frac{C_P \sigma_{\max}^{\frac{d}{2}+2}}{\nu \sigma_{\min}^d} \|\tilde{\mathbf{f}}\|_{L^2(D_\omega)} \leq \theta \frac{\nu \sigma_{\min}^{d+1}}{C_I^2 \sigma_{\max}^{d+2}} = \theta \frac{\nu \sigma_{\max}^{-2} \sigma_{\min}^d}{C_I^2 \sigma_{\max}^d \sigma_{\min}^{-1}} = \theta \frac{\nu \alpha}{\hat{C}}$$

where we have used that $\mathbf{u} \in \mathcal{M}$ for the first inequality and relation (3.41) for the second one. \square

3.2.4 Specific form of the random mapping

We assume from now on that the random mapping $\mathbf{x}(\xi, \omega)$ is parametrized by L mutually independent random variables and write $\mathbf{x}(\xi, \omega) = \mathbf{x}(\xi, Y_1(\omega), \dots, Y_L(\omega))$ with a slight abuse of notation. This assumption with L finite, usually referred to as *finite dimensional noise assumption*, is necessary to make the problem feasible for numerical simulation. Such approximation of a random field can be achieved by several techniques, for instance using truncated Karhunen-Loève or Fourier expansions. More precisely, we assume that the mapping \mathbf{x}_ω from

3.2. Steady-state incompressible Navier-Stokes equations in random domains

D to D_ω writes

$$\mathbf{x}_\omega(\xi) = \boldsymbol{\varphi}_0(\xi) + \varepsilon \sum_{j=1}^L \boldsymbol{\varphi}_j(\xi) Y_j(\omega), \quad (3.47)$$

where the Y_j , $j = 1, \dots, L$, are independent random variables with zero mean and unit variance, the deterministic functions $\boldsymbol{\varphi}_j : D \rightarrow \mathbb{R}^d$ are assumed to be smooth so that $\nabla \boldsymbol{\varphi}_0 \in [W^{1,\infty}(D)]^{d \times d}$ and $\nabla \boldsymbol{\varphi}_j \in [L^\infty(D)]^{d \times d}$ for $j = 1, \dots, L$, and $\varepsilon \in [0, \varepsilon_{\max}]$ is a parameter that controls the amount of randomness. We assume that the random variables Y_j , $j = 1, \dots, L$, and the functions $\boldsymbol{\varphi}_j$, $j = 0, 1, \dots, L$, are independent of ε . Without loss of generality, we can assume that $\boldsymbol{\varphi}_0$ is the identity mapping (see [76]), i.e.

$$\mathbf{x}_\omega(\xi) = \xi + \varepsilon \sum_{j=1}^L \boldsymbol{\varphi}_j(\xi) Y_j(\omega). \quad (3.48)$$

The Jacobian matrix A^{-1} associated to \mathbf{x}_ω therefore reads

$$A^{-1}(\xi, \omega) = I + \varepsilon A_1(\xi, \omega) \quad \text{with} \quad A_1(\xi, \omega) = \sum_{j=1}^L \nabla \boldsymbol{\varphi}_j(\xi) Y_j(\omega),$$

where I denotes the $d \times d$ identity matrix and $\nabla \boldsymbol{\varphi}_j(\xi)$ is the Jacobian matrix of $\boldsymbol{\varphi}_j$ for $j = 1, \dots, L$. Finally, we make the following additional assumptions to ensure that (3.36) is satisfied:

$$Y_j(\Omega) = [-\gamma_j, \gamma_j] =: \Gamma_j \quad \text{with} \quad \gamma_j > 0, \quad j = 1, \dots, L, \quad (3.49)$$

and

$$\varepsilon_{\max} < \frac{1}{\delta} \quad \text{with} \quad \delta \text{ such that } \sum_{j=1}^L \gamma_j \|\nabla \boldsymbol{\varphi}_j(\xi)\|_2 \leq \delta \quad \text{a.e. in } D, \quad (3.50)$$

where $\|\cdot\|_2$ is the spectral norm. It is straightforward to show that under assumptions (3.49) and (3.50), then (3.36) is fullfield for any $\varepsilon \in [0, \varepsilon_{\max}]$ with $\sigma_{\min} = 1 - \varepsilon_{\max} \delta$ and $\sigma_{\max} = 1 + \varepsilon_{\max} \delta$.

Remark 3.2.10. A (truncated) Karhunen-Loève expansion of the random vector field \mathbf{x}_ω (see [76, 87, 88]) yields a characterization of \mathbf{x}_ω that can be recast into the form (3.47). In this case, the functions $\boldsymbol{\varphi}_j$, $j = 1, \dots, L$, write $\boldsymbol{\varphi}_j = \sqrt{\lambda_j} \boldsymbol{\psi}_j$ with $\{\lambda_j, \boldsymbol{\psi}_j\}$ the eigenpairs of the (compact, self-adjoint) integral operator associated with the covariance kernel $V : D \times D \rightarrow \mathbb{R}^{d \times d}$ given by

$$V(\xi, \xi') := \frac{1}{\varepsilon^2} \mathbb{E} [(\mathbf{x}_\omega(\xi) - \boldsymbol{\varphi}_0(\xi))(\mathbf{x}_\omega(\xi') - \boldsymbol{\varphi}_0(\xi'))^T].$$

We underline that in this work, we do not take into account the error made when the random mapping is approximated via a finite number of random variables. Therefore, we assume here that (3.47) is an exact representation of the random mapping introduced in Section 3.2.2.

Due to the Doob-Dynkin Lemma, the solutions \mathbf{u} and p of (3.35) depend on the same random variables as \mathbf{x}_ω . Defining the random vector $\mathbf{Y} = (Y_1, \dots, Y_L)$, we can thus write $\mathbf{u}(\xi, \omega) = \mathbf{u}(\xi, \mathbf{Y}(\omega))$ and $p(\xi, \omega) = p(\xi, \mathbf{Y}(\omega))$. The complete probability space (Ω, \mathcal{F}, P) can thus be

replaced by $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $\Gamma = \Gamma_1 \times \dots \times \Gamma_L$, $B(\Gamma)$ is the Borel σ -algebra on Γ and $\rho(\mathbf{y})d\mathbf{y}$ is the probability measure of the random vector \mathbf{Y} . Notice that since the random variables Y_j , $j = 1, \dots, L$, are assumed independent, the joint density function ρ factorizes as $\rho(\mathbf{y}) = \prod_{j=1}^L \rho_j(y_j)$ for all $\mathbf{y} = (y_1, \dots, y_L) \in \Gamma$. Therefore, for any integrable function $\hat{g} : \Gamma \rightarrow \mathbb{R}$ on $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, the expectation of the random variable $g = g(\omega) = \hat{g}(\mathbf{Y}(\omega))$ is by definition given by

$$\mathbb{E}[g] = \int_{\Omega} g(\omega) dP(\omega) = \int_{\Omega} \hat{g}(\mathbf{Y}(\omega)) dP(\omega) = \int_{\Gamma} \hat{g}(\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y}.$$

With a little abuse of notation, we will not distinguish g and \hat{g} in what follows. The problem (3.31) can then be rewritten into the following parametric form:

find $(\mathbf{u}, p) \in L^2_{\rho}(\Gamma; V) \times L^2_{\rho}(\Gamma; Q)$ such that

$$\begin{cases} a(\mathbf{u}(\mathbf{y}), \mathbf{v}; \mathbf{y}) + c(\mathbf{u}(\mathbf{y}), \mathbf{u}(\mathbf{y}), \mathbf{v}; \mathbf{y}) + b(\mathbf{v}, p(\mathbf{y}); \mathbf{y}) &= F(\mathbf{v}; \mathbf{y}) \\ b(\mathbf{u}(\mathbf{y}), q; \mathbf{y}) &= 0 \end{cases} \quad (3.51)$$

for all $(\mathbf{v}, q) \in V \times Q$ and ρ -a.e. in Γ , where the various forms are defined as in (3.32) with $A(\xi, \omega)$, $A^{-1}(\xi, \omega)$, $J_{\mathbf{x}}(\xi, \omega)$ and $\mathbf{f}(\xi, \omega)$ replaced by $A(\xi, \mathbf{y})$, $A^{-1}(\xi, \mathbf{y})$, $J_{\mathbf{x}}(\xi, \mathbf{y})$ and $\mathbf{f}(\xi, \mathbf{y})$, respectively. This problem is well-posed under the so-called *small data* assumption (3.41) with $\mathbf{f}(\omega)$ replaced by $\mathbf{f}(\mathbf{y})$ and a.s. in Ω replaced by ρ -a.e. in Γ , the proof being essentially the same as the proof of Proposition 3.2.8. The random weak solution of problem (3.35), i.e. the solution of (3.51), is then given by $(u(\mathbf{Y}(\omega)), p(\mathbf{Y}(\omega)))$ with (\mathbf{u}, p) the parametric solution of (3.51).

Remark 3.2.11. Notice that for any $\mathbf{y} \in \Gamma$, the partial derivative with respect to y_j of the solutions $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\mathbf{x}, \mathbf{y})$ and $\tilde{p} = \tilde{p}(\mathbf{x}, \mathbf{y})$ of the problem defined on $D_{\mathbf{y}}$ is given for $j = 1, \dots, L$ by

$$\frac{\partial \tilde{\mathbf{u}}}{\partial y_j} = \frac{\partial \mathbf{u}}{\partial y_j} \circ \xi_{\mathbf{y}} + \left(\frac{\partial \xi_{\mathbf{y}}}{\partial y_j} \cdot \nabla_{\xi} \right) \mathbf{u} \circ \xi_{\mathbf{y}} \quad \text{and} \quad \frac{\partial \tilde{p}}{\partial y_j} = \frac{\partial p}{\partial y_j} \circ \xi_{\mathbf{y}} + \left(\frac{\partial \xi_{\mathbf{y}}}{\partial y_j} \cdot \nabla_{\xi} \right) p \circ \xi_{\mathbf{y}}. \quad (3.52)$$

In other words, the (Eulerian) partial derivative with respect to y_j of $\tilde{\mathbf{u}}$ (resp. \tilde{p}) is equal to the material derivative with respect to y_j of $\mathbf{u} = \tilde{\mathbf{u}} \circ \mathbf{x}_{\mathbf{y}}$ (resp. $p = \tilde{p} \circ \mathbf{x}_{\mathbf{y}}$), transported back to $D_{\mathbf{y}}$. Moreover, we have the relation

$$\left(\frac{\partial \xi_{\mathbf{y}}}{\partial y_j} \cdot \nabla_{\xi} \right) \mathbf{u} \circ \xi_{\mathbf{y}} = - \left(\frac{\partial \mathbf{x}_{\mathbf{y}}}{\partial y_j} \circ \xi_{\mathbf{y}} \cdot \nabla_{\mathbf{x}} \right) \tilde{\mathbf{u}} \quad (3.53)$$

and using it in (3.52) we recognize an analogy with the Arbitrary Lagrangian Eulerian (ALE) formulation of PDEs on moving domains [27, 56], where the (Eulerian) partial time-derivative is replaced by the partial time-derivative on the ALE frame written in the Eulerian coordinate plus the convective-type term of the right-hand side of (3.53) in which the so-called domain velocity is involved.

3.2.5 *A posteriori* error analysis

To simplify the presentation, we assume from now on that $d = 2$ and that $\tilde{\mathbf{f}} \in [H^2(\hat{D})]^2$. Since the forcing term on D is given by $\mathbf{f} = \tilde{\mathbf{f}} \circ \mathbf{x}_Y$ and we assumed $\boldsymbol{\varphi}_0$ to be the identity mapping, the regularity assumption on $\tilde{\mathbf{f}}$ allows us to write $\mathbf{f} = \mathbf{f}(\boldsymbol{\xi}, \omega) = \mathbf{f}(\boldsymbol{\xi}, \mathbf{Y}(\omega))$ as

$$\mathbf{f}(\mathbf{Y}) = \mathbf{f}_0 + \varepsilon \mathbf{f}_1(\mathbf{Y}) + \mathcal{O}(\varepsilon^2) \quad \text{with} \quad \mathbf{f}_0 := \tilde{\mathbf{f}}, \quad \mathbf{f}_1(\mathbf{Y}) := \sum_{j=1}^L \mathbf{F}_j Y_j, \quad \mathbf{F}_j := (\nabla_{\mathbf{x}} \tilde{\mathbf{f}}) \boldsymbol{\varphi}_j. \quad (3.54)$$

The constant in the term of order ε^2 in (3.54) depends on the second derivatives of $\tilde{\mathbf{f}}$ and products $\boldsymbol{\varphi}_i \boldsymbol{\varphi}_j$, $i, j = 1, \dots, L$. Moreover, since $d = 2$ we have

$$J_{\mathbf{x}} = \det(A^{-1}) = \det(I + \varepsilon A_1) = 1 + \varepsilon \operatorname{tr}(A_1) + \varepsilon^2 \det(A_1) \quad \text{with} \quad \det(A_1) \leq \delta^2 \quad (3.55)$$

using assumption (3.50) to bound $\det(A_1)$ and

$$A = I - \varepsilon A_1 + \sum_{k=2}^{\infty} (-1)^k \varepsilon^k A_1^k \quad \text{with} \quad \left\| \sum_{k=2}^{\infty} (-1)^k \varepsilon^k A_1^k \right\|_2 \leq \frac{\varepsilon^2 \delta^2}{1 - \varepsilon \delta} \leq \frac{\varepsilon^2 \delta^2}{\sigma_{\min}}, \quad (3.56)$$

where we have used a von Neumann series to expand $A = (I + \varepsilon A_1)^{-1}$. We use a perturbation approach expanding the solution (\mathbf{u}, p) on the reference domain D with respect to ε up to a certain order as

$$(\mathbf{u}(\boldsymbol{\xi}, \mathbf{Y}(\omega)), p(\boldsymbol{\xi}, \mathbf{Y}(\omega))) = (\mathbf{u}_0(\boldsymbol{\xi}), p_0(\boldsymbol{\xi})) + \varepsilon (\mathbf{u}_1(\boldsymbol{\xi}, \mathbf{Y}(\omega)), p_1(\boldsymbol{\xi}, \mathbf{Y}(\omega))) + \dots \quad (3.57)$$

where (\mathbf{u}_0, p_0) is the solution of the *standard* Navier-Stokes equations on D , i.e. it solves:

find $\mathbf{u}_0 : D \rightarrow \mathbb{R}^d$ and $p_0 : D \rightarrow \mathbb{R}$ such that:

$$\begin{cases} -\nu \Delta \mathbf{u}_0 + (\mathbf{u}_0 \cdot \nabla) \mathbf{u}_0 + \nabla p_0 &= \mathbf{f}_0, & \boldsymbol{\xi} \in D \\ \nabla \cdot \mathbf{u}_0 &= 0, & \boldsymbol{\xi} \in D \\ \mathbf{u}_0 &= \mathbf{0}, & \boldsymbol{\xi} \in \partial D. \end{cases} \quad (3.58)$$

Writing $\mathbf{u}_1 = \sum_{j=1}^L \mathbf{U}_j Y_j$ and $p_1 = \sum_{j=1}^L P_j Y_j$, it can be shown that the couple (\mathbf{u}_1, p_1) is obtained by solving the L (linear) problems:

for $j = 1, \dots, L$, find $\mathbf{U}_j : D \rightarrow \mathbb{R}^d$ and $P_j : D \rightarrow \mathbb{R}$ such that:

$$\begin{cases} -\nu \Delta \mathbf{U}_j + (\mathbf{u}_0 \cdot \nabla) \mathbf{U}_j + (\mathbf{U}_j \cdot \nabla) \mathbf{u}_0 + \nabla P_j &= g_j(\mathbf{u}_0, p_0), & \boldsymbol{\xi} \in D \\ \nabla \cdot \mathbf{U}_j &= h_j(\mathbf{u}_0), & \boldsymbol{\xi} \in D \\ \mathbf{U}_j &= \mathbf{0}, & \boldsymbol{\xi} \in \partial D, \end{cases} \quad (3.59)$$

where

$$\begin{aligned} g_j(\mathbf{u}_0, p_0) &= (tr(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j) + \nu \nabla \cdot [(\hat{B}_j \nabla) \mathbf{u}_0] - (\mathbf{u}_0 \cdot B_j \nabla) \mathbf{u}_0 - (B_j \nabla) p_0, \\ h_j(\mathbf{u}_0) &= -(B_j \nabla) \cdot \mathbf{u}_0 \end{aligned}$$

with

$$B_j := tr(\nabla \boldsymbol{\varphi}_j) I - \nabla \boldsymbol{\varphi}_j^T \quad \text{and} \quad \hat{B}_j := tr(\nabla \boldsymbol{\varphi}_j) I - (\nabla \boldsymbol{\varphi}_j + \nabla \boldsymbol{\varphi}_j^T). \quad (3.60)$$

Some details about the derivation of problems (3.58) and (3.59) are given in Appendix 3.A. Here, we approximate the solution of the deterministic problem (3.58) using the finite element method to obtain an approximation $(\mathbf{u}_{0,h}, p_{0,h})$ and we provide an *a posteriori* error estimate of $(\mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h})$. For any $h > 0$, let \mathcal{T}_h be a family of shape regular partitions (see [49]) of D into d -simplices K of diameter $h_K \leq h$. Moreover, let (V_h, Q_h) with $V_h \subset V$ and $Q_h \subset Q$ be a pair of inf-sup stable finite element spaces, such as mini-elements $\mathbb{P}_{1b} - \mathbb{P}_1$ (see [5] or [69, p.175] for a proof of stability of these spaces) or Taylor-Hood $\mathbb{P}_2 - \mathbb{P}_1$. We denote by $(\mathbf{u}_{0,h}, p_{0,h})$ the FE approximation of the (weak) solution (\mathbf{u}_0, p_0) of problem (3.58). Writing $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$, it is obtained by solving:

find $(\mathbf{u}_{0,h}, p_{0,h}) \in V_h \times Q_h$ such that

$$\begin{cases} a(\mathbf{u}_{0,h}, \mathbf{v}_h; \mathbf{y}_0) + c(\mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}_h; \mathbf{y}_0) + b(\mathbf{v}_h, p_{0,h}; \mathbf{y}_0) &= F(\mathbf{v}_h; \mathbf{y}_0) \\ b(\mathbf{u}_{0,h}, q_h; \mathbf{y}_0) &= 0 \end{cases} \quad (3.61)$$

for all $(\mathbf{v}_h, q_h) \in V_h \times Q_h$. The rest of this section is devoted to an *a posteriori* error analysis for the error $\|(\mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h})\|$, where the norm $\|\cdot\|$ is defined for any $(\mathbf{v}, q) \in L_P^2(\Omega; V) \times L_P^2(\Omega; Q)$ by

$$\|(\mathbf{v}, q)\| := \left(\mathbb{E} \left[\nu \|\nabla \mathbf{v}\|_{L^2(D)}^2 + \frac{1}{\nu} \|q\|_{L^2(D)}^2 \right] \right)^{\frac{1}{2}}.$$

Remark 3.2.12. Notice that we obtain the same results if we use the norm $\nu^2 \|\nabla \mathbf{v}\|^2 + \|q\|^2$ or $\|\nabla \mathbf{v}\|^2 + \frac{1}{\nu^2} \|q\|^2$ on $V \times Q$. This choice of scaling is guided by the dimension unit of ν , p and $\nabla \mathbf{u}$. This is moreover the natural scaling that arises when analysing the *a priori* estimates on the solution or when performing the *a posteriori* error analysis (see Appendix 3.B for more details).

As we will see in the following, the error estimate consists of two parts, namely a part due to the finite element approximation (in h) and another one due to the uncertainty (in ε). Let us define for any $\mathbf{y} \in \Gamma$ the residual $R(\cdot; \mathbf{y}) : V \times Q \rightarrow \mathbb{R}$, which depends on $(\mathbf{u}_{0,h}, p_{0,h})$, by $R((\mathbf{v}, q); \mathbf{y}) = R_1(\mathbf{v}; \mathbf{y}) + R_2(q; \mathbf{y})$ with

$$\begin{aligned} R_1(\mathbf{v}; \mathbf{y}) &:= F(\mathbf{v}; \mathbf{y}) - a(\mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) - b(\mathbf{v}, p_{0,h}; \mathbf{y}) - c(\mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) \\ R_2(q; \mathbf{y}) &:= -b(\mathbf{u}_{0,h}, q; \mathbf{y}). \end{aligned}$$

The first step in the residual-based error estimation consists in linking the error to the residual. The norm of the residual is then bounded by a computable quantity (possibly up to a

multiplicative constant).

Proposition 3.2.13. *Let σ_{min} , σ_{max} , β and θ be defined in (3.36), (3.39) and (3.41), respectively. If h is small enough, then there exists a constant $C > 0$ depending only on θ , σ_{min} , σ_{max} and β such that a.s. in Ω*

$$\nu \|\nabla(\mathbf{u}(\mathbf{Y}) - \mathbf{u}_{0,h})\|_{L^2(D)}^2 + \frac{1}{\nu} \|p(\mathbf{Y}) - p_{0,h}\|_{L^2(D)}^2 \leq C \left(\frac{1}{\nu} \|R_1(\cdot, \mathbf{Y})\|_{V'}^2 + \nu \|R_2(\cdot, \mathbf{Y})\|_{Q'}^2 \right). \quad (3.62)$$

We mention that the closer θ to 1, the larger C in Proposition 3.2.13, see relation (3.71). Similarly, the closer σ_{min} to 0, the larger C will be. The proof of this proposition is inspired by what is done in [2] for the deterministic steady Navier-Stokes equations. In order to simplify the notation, we will write $\|\cdot\|$ instead of $\|\cdot\|_{L^2(D)}$ in the sequel.

Proof. In what follows, all equations depending on \mathbf{y} hold ρ -a.e. in Γ , without specifically mentioning it. Moreover, the dependence of the functions with respect to $\mathbf{y} \in \Gamma$ will not necessarily be indicated. Let $\mathbf{e}(\mathbf{y}) := \mathbf{u}(\mathbf{y}) - \mathbf{u}_{0,h}$ and $E(\mathbf{y}) := p(\mathbf{y}) - p_{0,h}$. Then (3.51) yields

$$a(\mathbf{e}, \mathbf{v}; \mathbf{y}) + b(\mathbf{v}, E; \mathbf{y}) + b(\mathbf{e}, q; \mathbf{y}) + D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) = R((\mathbf{v}, q); \mathbf{y}) \quad (3.63)$$

for all $(\mathbf{v}, q) \in V \times Q$, where

$$D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) := c(\mathbf{u}, \mathbf{u}, \mathbf{v}; \mathbf{y}) - c(\mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}).$$

We can show that

$$D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) \leq (2\theta\nu\alpha + \hat{C}\|\nabla\mathbf{e}_0\|)\|\nabla\mathbf{e}\|\|\nabla\mathbf{v}\| \quad (3.64)$$

and

$$D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{u} - \mathbf{u}_{0,h}; \mathbf{y}) \leq (\theta\nu\alpha + \hat{C}\|\nabla\mathbf{e}_0\|)\|\nabla\mathbf{e}\|^2 \quad (3.65)$$

where $\mathbf{e}_0 := \mathbf{u}_0 - \mathbf{u}_{0,h}$ and M , α and \hat{C} are defined in Proposition 3.2.5. Indeed, for any $\mathbf{v} \in V$ we have

$$\begin{aligned} D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) &= c(\mathbf{u}, \mathbf{u} - \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) + c(\mathbf{u} - \mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) \\ &\leq \hat{C}(\|\nabla\mathbf{u}\| + \|\nabla\mathbf{u}_0\| + \|\nabla\mathbf{e}_0\|)\|\nabla\mathbf{e}\|\|\nabla\mathbf{v}\| \\ &\leq \hat{C}\left(2\theta\frac{\alpha\nu}{\hat{C}} + \|\nabla\mathbf{e}_0\|\right)\|\nabla\mathbf{e}\|\|\nabla\mathbf{v}\| \end{aligned}$$

thanks to (3.42), which proves relation (3.64). Relation (3.65) is proved analogously using the fact that $c(\mathbf{u}, \mathbf{v}, \mathbf{v}; \mathbf{y}) = 0$ for any $\mathbf{v} \in V$. The rest of the proof consists of two steps, first the derivation of a bound on $\|E\|$ and then a bound on $\|\nabla\mathbf{e}\|$.

Using the inf-sup condition (3.39) for b , the bound (3.64) on D , the continuity of a and the

relation (3.63) with $q = 0$, we have

$$\begin{aligned} \|E\| &\leq \frac{1}{\beta} \sup_{\mathbf{v} \in V} \frac{|b(\mathbf{v}, p - p_{0,h}; \mathbf{y})|}{\|\nabla \mathbf{v}\|} = \frac{1}{\beta} \sup_{\mathbf{v} \in V} \frac{|R_1(\mathbf{v}; \mathbf{y}) - a(\mathbf{u} - \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) - D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y})|}{\|\nabla \mathbf{v}\|} \\ &\leq \frac{1}{\beta} [\|R_1(\cdot; \mathbf{y})\|_{V'} + (\nu M + 2\nu\alpha + \hat{C}\|\nabla \mathbf{e}_0\|)\|\nabla \mathbf{e}\|]. \end{aligned} \quad (3.66)$$

Therefore, using the relation $(a + b)^2 \leq 2(a^2 + b^2)$ we obtain

$$\frac{1}{\nu} \|E\|^2 \leq \frac{2}{\beta^2 \nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + \frac{2(M + 2\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\|)^2}{\beta^2} \nu \|\nabla \mathbf{e}\|^2. \quad (3.67)$$

We now give a bound on the error $\|\nabla \mathbf{e}\|$ for the velocity. Using the inequalities (3.65) and (3.66), the coercivity of the bilinear form a , Young's inequality several times and taking $\mathbf{v} = \mathbf{e}$ and $q = -E$ in (3.63), we get

$$\begin{aligned} \nu\alpha \|\nabla \mathbf{e}\|^2 &\leq a(\mathbf{e}, \mathbf{e}; \mathbf{y}) = R_1(\mathbf{e}; \mathbf{y}) - R_2(E; \mathbf{y}) - D(\mathbf{u}, \mathbf{u}_{0,h}, \mathbf{e}) \\ &\leq \|R_1(\cdot; \mathbf{y})\|_{V'} \|\nabla \mathbf{e}\| + \|R_2(\cdot; \mathbf{y})\|_{Q'} \|E\| + (\theta\nu\alpha + \hat{C}\|\nabla \mathbf{e}_0\|)\|\nabla \mathbf{e}\|^2 \\ &\leq \frac{1}{2\gamma_1\nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + \frac{\nu}{2\beta^2\gamma_2} \|R_2(\cdot; \mathbf{y})\|_{Q'}^2 + \frac{1}{\beta} \|R_1(\cdot; \mathbf{y})\|_{V'} \|R_2(\cdot; \mathbf{y})\|_{Q'} \\ &\quad + \left(\frac{\gamma_1}{2} + \frac{\gamma_2(M + 2\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\|)^2}{2} + \theta\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\| \right) \nu \|\nabla \mathbf{e}\|^2 \\ &\leq \frac{c_1}{\nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + c_2 \nu \|R_2(\cdot; \mathbf{y})\|_{Q'}^2 \\ &\quad + \left(\frac{\gamma_1}{2} + \frac{\gamma_2(M + 2\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\|)^2}{2} + \theta\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\| \right) \nu \|\nabla \mathbf{e}\|^2, \end{aligned} \quad (3.68)$$

with

$$c_1 = \frac{1}{2\gamma_1} + \frac{1}{2} \quad \text{and} \quad c_2 = \frac{1}{2\gamma_2\beta^2} + \frac{1}{2\beta^2}.$$

Recalling that $\theta \in [0, 1[$ and using the convergence of $\mathbf{u}_{0,h}$ to \mathbf{u}_0 as h tends to 0, we can choose h , γ_1 and γ_2 small enough so that

$$\frac{\gamma_1}{2} + \frac{\gamma_2(M + 2\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\|)^2}{2} + \theta\alpha + \frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\| \leq \frac{1+\theta}{2} \alpha. \quad (3.69)$$

For instance, we can choose h small enough so that

$$\frac{\hat{C}}{\nu} \|\nabla \mathbf{e}_0\| \leq \frac{1-\theta}{6} \alpha \quad (3.70)$$

and take

$$\gamma_1 = \frac{1-\theta}{3} \alpha \quad \text{and} \quad \gamma_2 = \frac{1-\theta}{3(M + 2\alpha + \frac{1-\theta}{6} \alpha)^2} \alpha$$

3.2. Steady-state incompressible Navier-Stokes equations in random domains

which depends only on θ , σ_{min} and σ_{max} . Therefore, the last term of the right-hand side of inequality (3.68) can be moved to the left and we get

$$\nu \|\nabla \mathbf{e}\|^2 \leq \frac{2}{(1-\theta)\alpha} \left[\frac{c_1}{\nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + c_2 \nu \|R_2(\cdot; \mathbf{y})\|_{Q'}^2 \right]. \quad (3.71)$$

Using this bound in (3.67) together with (3.70) we get

$$\frac{1}{\nu} \|E\|^2 \leq \left(\frac{2}{\beta^2} + \frac{4c_1}{3\gamma_2\beta^2} \right) \frac{1}{\nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + \frac{4c_2}{3\gamma_2\beta^2} \nu \|R_2(\cdot; \mathbf{y})\|_{Q'}^2.$$

Replacing finally \mathbf{y} by $\mathbf{Y}(\omega)$, the combination of last two inequalities permits to conclude the proof since c_1 and c_2 depend only on β as well as γ_1 and γ_2 , which in turn depend only on θ , σ_{min} et σ_{max} .

□

From Proposition 3.2.13, we deduce the following bound on the error in the $\|\cdot\|$ norm

$$\|\mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h}\| \leq \sqrt{C} \left(\frac{1}{\nu} \mathbb{E} [\|R_1\|_{V'}^2] + \nu \mathbb{E} [\|R_2\|_{Q'}^2] \right)^{\frac{1}{2}} \quad (3.72)$$

by simply taking first the expected value and then the square root on both sides of inequality (3.62). The goal is now to derive a computable (deterministic) error estimator by estimating the residuals that appear in the right-hand side of (3.72). We use a standard procedure to estimate the part due to the space discretization and proceed in two different ways for the part due to the uncertainty, more precisely the truncation in (3.57). The first one is straightforward and does not require the resolution of additional problems. However, it uses the triangle inequality as well as the Poincaré inequality (on the fixed domain D) to bound the terms due to the external forces and the convection. Even though the Poincaré constant is a uniform bound, the loss when using Poincaré's inequality can be different depending of the problem, affecting the sharpness of the error estimate from case to case. The second procedure consists in computing the dual norm of some functional, and therefore requires the resolution of additional (linear) problems. However, it has the advantage of requiring the use of Cauchy-Schwarz's inequality only and thus does not suffer from the drawback mentioned above.

First error estimate

Let $[\cdot]_{\mathbf{n}_e}$ denotes the jump across an edge $e \in \mathcal{T}_h$ in the direction \mathbf{n}_e defined by

$$[\mathbf{g}]_{\mathbf{n}_e}(\xi) := \lim_{t \rightarrow 0^+} [\mathbf{g}(\xi + t\mathbf{n}_e) - \mathbf{g}(\xi - t\mathbf{n}_e)],$$

where \mathbf{n}_e is a unit normal vector to e of arbitrary (but fixed) direction for internal edges and the outward unit vector for boundary edges. Since we impose homogeneous Dirichlet conditions at the boundary, we set the jump to zero for boundary edges. We now have all the ingredients

necessary to derive our first error estimate.

Proposition 3.2.14. *Let (\mathbf{u}, p) be the (weak) solution of problem (3.35) and let $(\mathbf{u}_{0,h}, p_{0,h})$ be the solution of problem (3.61). If the assumptions of Proposition 3.2.13 are satisfied, then there exist positive constants C_1, C_2 and C_3 independent of h and ε such that*

$$\| \mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h} \| \leq \sqrt{2C} (C_1 \eta_h^2 + C_2 \eta_\varepsilon^2)^{\frac{1}{2}} + \sqrt{C} C_3 \varepsilon^2 \quad \text{with} \quad \eta_h^2 = \sum_{K \in \mathcal{T}_h} \eta_K^2 \quad \text{and} \quad \eta_\varepsilon^2 = \sum_{j=1}^L \eta_j^2, \quad (3.73)$$

where C is the constant in Proposition 3.2.13 and

$$\eta_K^2 := \frac{1}{v} \eta_{K,1}^2 + v \eta_{K,2}^2 \quad \text{and} \quad \eta_j^2 := \frac{1}{v} \eta_{j,1}^2 + v \eta_{j,2}^2 \quad (3.74)$$

with

$$\begin{aligned} \eta_{K,1}^2 &:= h_K^2 \| \mathbf{f}_0 + v \Delta \mathbf{u}_{0,h} - (\mathbf{u}_{0,h} \cdot \nabla) \mathbf{u}_{0,h} - \nabla p_{0,h} \|_{L^2(K)}^2 + \sum_{e \in K} h_e \left\| \frac{1}{2} [v(\nabla \mathbf{u}_{0,h}) \mathbf{n}_e - p_{0,h} \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \\ \eta_{K,2}^2 &:= \| \nabla \cdot \mathbf{u}_{0,h} \|_{L^2(K)}^2 \\ \eta_{j,1}^2 &:= \varepsilon^2 \left(\| \text{tr}(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j \|^2 + v^2 \| (\hat{B}_j \nabla) \mathbf{u}_{0,h} \|^2 + \| p_{0,h} B_j \|^2 + \| (\mathbf{u}_{0,h} \cdot B_j \nabla) \mathbf{u}_{0,h} \|^2 \right) \\ \eta_{j,2}^2 &:= \varepsilon^2 \| (B_j \nabla) \cdot \mathbf{u}_{0,h} \|^2, \end{aligned} \quad (3.75)$$

B_j and \hat{B}_j being defined in (3.60), \mathbf{f}_0 and \mathbf{F}_j in (3.54). Moreover, C_1 depends only on the mesh aspect ratio while C_2 depends only on the Poincaré constant on D .

Remark 3.2.15. Notice that if $\varepsilon_{\max} \delta$ is close to 1, or in other words σ_{\min} is close to 0, then the constant C_3 in Proposition 3.2.14 might be large, see (3.56). Therefore, in order for the last term of (3.73) to be negligible, we need to assume small perturbations of the domain, for instance by imposing $\varepsilon_{\max} \leq \frac{1}{2\delta}$.

Proof. Similarly to the proof of Proposition 3.2.13, it is understood that all equations depending on \mathbf{y} hold ρ -a.e. in Γ unless explicitly stated. Thanks to (3.72), we only need to bound the expectation of $\frac{1}{v} \| R_1(\cdot; \mathbf{Y}) \|_{V'}^2$ and $v \| R_2(\cdot; \mathbf{Y}) \|_{Q'}^2$, that is

$$\int_{\Gamma} \frac{1}{v} \| R_1(\cdot; \mathbf{y}) \|_{V'}^2 \rho(\mathbf{y}) d\mathbf{y} \quad \text{and} \quad \int_{\Gamma} v \| R_2(\cdot; \mathbf{y}) \|_{Q'}^2 \rho(\mathbf{y}) d\mathbf{y},$$

by computable quantities. We decompose each term R_1 and R_2 into two parts which control the FE error and the error due to truncation in the expansion (3.57), respectively. For $\mathbf{y}_0 = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$ and for all $\mathbf{y} \in \Gamma$, $\mathbf{v} \in V$ and $q \in Q$ we write

$$R_1(\mathbf{v}; \mathbf{y}) = R_1(\mathbf{v}; \mathbf{y}_0) + [R_1(\mathbf{v}; \mathbf{y}) - R_1(\mathbf{v}; \mathbf{y}_0)]$$

and

$$R_2(q; \mathbf{y}) = R_2(q; \mathbf{y}_0) + [R_2(q; \mathbf{y}) - R_2(q; \mathbf{y}_0)].$$

3.2. Steady-state incompressible Navier-Stokes equations in random domains

Using standard procedure (Galerkin orthogonality, Clément interpolation [50]), see for instance [118], and taking the contribution of the constant ν into account, the deterministic quantities can be bounded by

$$\frac{1}{\nu} \|R_1(\cdot; \mathbf{y}_0)\|_{V'}^2 + \nu \|R_2(\cdot; \mathbf{y}_0)\|_{Q'}^2 \leq C_1 \sum_{K \in \mathcal{T}_h} \eta_K^2$$

where C_1 depends only on the *Clément interpolation constant* and the regularity of the mesh and the local error estimator η_K is defined in (3.74). We now bound the terms due to the uncertainty. We have

$$R_1(\mathbf{v}; \mathbf{y}) - R_1(\mathbf{v}; \mathbf{y}_0) = \Pi_1 + \Pi_2 + \Pi_3 + \Pi_4 \quad \text{and} \quad R_2(q; \mathbf{y}) - R_2(q; \mathbf{y}_0) = \Pi_5$$

with

$$\begin{aligned} \Pi_1 &:= F(\mathbf{v}; \mathbf{y}) - F(\mathbf{v}; \mathbf{y}_0) \leq C_P \|J_{\mathbf{x}} \mathbf{f} - \mathbf{f}_0\| \|\nabla \mathbf{v}\| \\ \Pi_2 &:= a(\mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) - a(\mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}_0) \leq \nu \|(J_{\mathbf{x}} A A^T - I) \nabla \mathbf{u}_{0,h}\| \|\nabla \mathbf{v}\| \\ \Pi_3 &:= b(\mathbf{v}, p_{0,h}; \mathbf{y}) - b(\mathbf{v}, p_{0,h}; \mathbf{y}_0) \leq \|(J_{\mathbf{x}} A^T - I) p_{0,h}\| \|\nabla \mathbf{v}\| \\ \Pi_4 &:= c(\mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}) - c(\mathbf{u}_{0,h}, \mathbf{u}_{0,h}, \mathbf{v}; \mathbf{y}_0) \leq C_P \|\mathbf{u}_{0,h} \cdot (J_{\mathbf{x}} A^T - I) \nabla \mathbf{u}_{0,h}\| \|\nabla \mathbf{v}\| \\ \Pi_5 &:= b(\mathbf{u}_{0,h}, q; \mathbf{y}) - b(\mathbf{u}_{0,h}, q; \mathbf{y}_0) \leq \|(J_{\mathbf{x}} A^T - I) \nabla \mathbf{u}_{0,h}\| \|q\|. \end{aligned}$$

The bound for each term is straightforward, except the one for the term Π_3 which can be obtained by writing it in component form, see Appendix 3.C for details. Therefore, we obtain

$$\frac{1}{\nu} \|R_1(\cdot; \mathbf{y})\|_{V'}^2 + \nu \|R_2(\cdot; \mathbf{y})\|_{Q'}^2 \leq C_1 \eta_h^2 + C_2 \kappa_\varepsilon(\mathbf{y})^2,$$

where C_2 is a (deterministic) constant that depends only on C_P and

$$\begin{aligned} \kappa_\varepsilon^2 &:= \frac{1}{\nu} \|J_{\mathbf{x}} \mathbf{f} - \mathbf{f}_0\|^2 + \nu \|(J_{\mathbf{x}} A A^T - I) \nabla \mathbf{u}_{0,h}\|^2 + \frac{1}{\nu} \|(J_{\mathbf{x}} A^T - I) p_{0,h}\|^2 \\ &\quad + \frac{1}{\nu} \|\mathbf{u}_{0,h} \cdot (J_{\mathbf{x}} A^T - I) \nabla \mathbf{u}_{0,h}\|^2 + \nu \|(J_{\mathbf{x}} A^T - I) \nabla \mathbf{u}_{0,h}\|^2. \end{aligned}$$

Since the independent random variables $\{Y_j\}$ are assumed to be of zero mean and unit variance, we have $\mathbb{E}[Y_j] = 0$ and $\mathbb{E}[Y_i Y_j] = \delta_{ij}$ for $i, j = 1, \dots, L$ and thus, using Young's inequality and the

relations (3.54), (3.55) and (3.56), among others, we get

$$\begin{aligned}
 \mathbb{E} [\|J_{\mathbf{x}} \mathbf{f} - \mathbf{f}_0\|^2] &= \varepsilon^2 \sum_{j=1}^L \|tr(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j\|^2 + \mathcal{O}(\varepsilon^3) \\
 \mathbb{E} [\|[(J_{\mathbf{x}} A A^T - I) \nabla] \mathbf{u}_{0,h}\|^2] &= \varepsilon^2 \sum_{j=1}^L \|(\hat{B}_j \nabla) \mathbf{u}_{0,h}\|^2 + \mathcal{O}(\varepsilon^3) \\
 \mathbb{E} [\|(J_{\mathbf{x}} A^T - I) p_{0,h}\|^2] &= \varepsilon^2 \sum_{j=1}^L \|p_{0,h} B_j\|^2 + \mathcal{O}(\varepsilon^3) \\
 \mathbb{E} [\|[\mathbf{u}_{0,h} \cdot (J_{\mathbf{x}} A^T - I) \nabla] \mathbf{u}_{0,h}\|^2] &= \varepsilon^2 \sum_{j=1}^L \|(\mathbf{u}_{0,h} \cdot B_j \nabla) \mathbf{u}_{0,h}\|^2 + \mathcal{O}(\varepsilon^3) \\
 \mathbb{E} [\|(J_{\mathbf{x}} A^T - I) \nabla \cdot \mathbf{u}_{0,h}\|^2] &= \varepsilon^2 \sum_{j=1}^L \|(B_j \nabla) \cdot \mathbf{u}_{0,h}\|^2 + \mathcal{O}(\varepsilon^3)
 \end{aligned}$$

with B_j and \hat{B}_j defined in (3.60). Therefore, for some constant $c_3 > 0$ independent of ε and h we get

$$\frac{1}{\nu} \mathbb{E} [\|R_1\|_{V'}^2] + \nu \mathbb{E} [\|R_2\|_{Q'}^2] \leq C_1 \sum_{K \in \mathcal{T}_h} \eta_K^2 + C_2 \sum_{j=1}^L \eta_j^2 + c_3 \varepsilon^3, \quad (3.76)$$

where η_j is defined in (3.74). To conclude the proof, it only remains to take the square root on both sides of inequality (3.76). Indeed, using the notation η_h and η_ε introduced in (3.73), we have

$$\left(\frac{1}{\nu} \mathbb{E} [\|R_1\|_{V'}^2] + \nu \mathbb{E} [\|R_2\|_{Q'}^2] \right)^{\frac{1}{2}} \leq (C_1 \eta_h^2 + C_2 \eta_\varepsilon^2 + c_3 \varepsilon^3)^{\frac{1}{2}} \leq \sqrt{C_1} \eta_h + (C_2 \eta_\varepsilon^2 + c_3 \varepsilon^3)^{\frac{1}{2}}$$

thanks to the inequality $\sqrt{a^2 + b^2} \leq a + b$ for any $a, b \geq 0$. Moreover, since $\eta_\varepsilon = \mathcal{O}(\varepsilon)$ we get for some constant $C_3 > 0$ independent of ε and h

$$(C_2 \eta_\varepsilon^2 + c_3 \varepsilon^3)^{\frac{1}{2}} = \sqrt{C_2} \eta_\varepsilon \left(1 + \frac{c_3 \varepsilon^3}{C_2 \eta_\varepsilon^2} \right)^{\frac{1}{2}} = \sqrt{C_2} \eta_\varepsilon \left(1 + \frac{1}{2} \frac{c_3 \varepsilon^3}{C_2 \eta_\varepsilon^2} - \frac{1}{8} \left(\frac{c_3 \varepsilon^3}{C_2 \eta_\varepsilon^2} \right)^2 + \dots \right) \leq \sqrt{C_2} \eta_\varepsilon + C_3 \varepsilon^2.$$

Finally, using the inequality $a + b \leq \sqrt{2} (a^2 + b^2)^{\frac{1}{2}}$ we obtain

$$\left(\frac{1}{\nu} \mathbb{E} [\|R_1\|_{V'}^2] + \nu \mathbb{E} [\|R_2\|_{Q'}^2] \right)^{\frac{1}{2}} \leq \sqrt{C_1} \eta_h + \sqrt{C_2} \eta_\varepsilon + C_3 \varepsilon^2 \leq \sqrt{2} (C_1 \eta_h^2 + C_2 \eta_\varepsilon^2)^{\frac{1}{2}} + C_3 \varepsilon^2,$$

which yields (3.73) thanks to (3.72).

□

Second error estimate

As mentioned above, the use of the triangle inequality to bound each term linked to R_1 separately, plus the Poincaré inequality for some of them (namely Π_1 and Π_4), in the derivation of the error estimate controlling the randomness of the problem can affect the sharpness of the error estimator η_ε . However, it has the advantage to require the resolution of only one (nonlinear) problem, namely the problem for $(\mathbf{u}_{0,h}, p_{0,h})$. We propose in this section a second error estimate for which the use of these inequalities is not required. It is obtained by computing, approximately, the dual norm of the residual $R_1(\mathbf{v}; \mathbf{y}) - R_1(\mathbf{v}; \mathbf{y}_0)$. Similarly to the error estimate of Proposition 3.2.14, the terms of higher order are neglected.

Proposition 3.2.16. *Under the assumptions of Proposition 3.2.14, there exist constants C_1 , C_3 and C_4 independent of h and ε and $s \in (0, 1]$ such that*

$$\|\mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h}\| \leq \sqrt{2C} (C_1 \eta_h^2 + \hat{\eta}_\varepsilon^2)^{\frac{1}{2}} + \sqrt{C} (C_3 \varepsilon^2 + C_4 h^s \varepsilon) \quad \text{with} \quad \hat{\eta}_\varepsilon^2 = \sum_{j=1}^L \hat{\eta}_j^2, \quad (3.77)$$

where η_h is as in (3.73) and

$$\hat{\eta}_j^2 := \frac{1}{\nu} \hat{\eta}_{j,1}^2 + \nu \eta_{j,2}^2$$

with $\eta_{j,2}$ given in (3.75) and $\hat{\eta}_{j,1}^2 := \varepsilon^2 \|\nabla \mathbf{w}_{j,h}\|_{L^2(D)}^2$ for $j = 1, \dots, L$, and $\mathbf{w}_{j,h} \in V_h$ is the solution of

$$\begin{aligned} \int_D \nabla \mathbf{w}_{j,h} : \nabla \mathbf{v}_h d\xi &= \int_D (tr(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j) \cdot \mathbf{v}_h d\xi - \nu \int_D (\hat{B}_j \nabla) \mathbf{u}_{0,h} : \nabla \mathbf{v}_h d\xi + \int_D p_{0,h} (B_j \nabla) \cdot \mathbf{v}_h d\xi \\ &\quad - \int_D [(\mathbf{u}_{0,h} \cdot B_j \nabla) \mathbf{u}_{0,h}] \cdot \mathbf{v}_h d\xi \end{aligned} \quad (3.78)$$

for all $\mathbf{v}_h \in V_h$. Moreover, the constant C_1 depends only on the mesh aspect ratio.

Notice that contrary to the error estimate of Proposition 3.2.14, there is no *internal* constant multiplying $\hat{\eta}_\varepsilon$ in (3.77), the constant $C_2 = C_2(C_P)$ appearing in (3.73) being indeed no longer present.

Proof. The proof is similar to the one of Proposition 3.2.14. The only difference is the estimation of the term $r(\mathbf{v}; \mathbf{y}) := R_1(\mathbf{v}; \mathbf{y}) - R_1(\mathbf{v}; \mathbf{y}_0)$ in the V' norm. We have $\|r(\cdot; \mathbf{y})\|_{V'} = \|\nabla \mathbf{w}(\mathbf{y})\|_{L^2(D)}$, where \mathbf{w} denotes the Riesz representant of r , i.e. $\mathbf{w}(\mathbf{y}) \in V$ is such that $\int_D \nabla \mathbf{w}(\mathbf{y}) : \nabla \mathbf{v} = r(\mathbf{v}; \mathbf{y})$ for all $\mathbf{v} \in V$ and ρ -a.e. in Γ . If we keep only the terms of order ε and use the properties of the random variables Y_j , $j = 1, \dots, L$, taking the expected value of $\|r(\cdot; \mathbf{Y})\|_{V'}^2$, we get

$$\mathbb{E} [\|r\|_{V'}^2] \leq \varepsilon^2 \sum_{j=1}^L \|\nabla \mathbf{w}_j\|_{L^2(D)}^2 + \mathcal{O}(\varepsilon^3)$$

where \mathbf{w}_j is the solution of

$$\int_D \nabla \mathbf{w}_j : \nabla \mathbf{v} d\xi = r_j(\mathbf{v}; \mathbf{u}_{0,h}, p_{0,h}) \quad \forall \mathbf{v} \in V$$

with

$$\begin{aligned} r_j(\mathbf{v}; \mathbf{u}_{0,h}, p_{0,h}) &:= \int_D (tr(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j) \cdot \mathbf{v} d\xi - \nu \int_D (\hat{B}_j \nabla) \mathbf{u}_{0,h} : \nabla \mathbf{v} d\xi + \int_D p_{0,h} (B_j \nabla) \cdot \mathbf{v} d\xi \\ &\quad - \int_D [(\mathbf{u}_{0,h} \cdot B_j \nabla) \mathbf{u}_{0,h}] \cdot \mathbf{v} d\xi. \end{aligned}$$

Obviously, the solution \mathbf{w}_j cannot be computed exactly. However, replacing \mathbf{w}_j by its finite element approximation $\mathbf{w}_{j,h} \in V_h$ introduces an error of higher order, namely an error of order εh^s with $s \in (0, 1]$. Indeed, introducing for $j = 1, \dots, L$ the solution $\boldsymbol{\psi}_j \in V$ of

$$\int_D \nabla \boldsymbol{\psi}_j : \nabla \mathbf{v} = r_j(\mathbf{v}; \mathbf{u}_0, p_0) \quad \mathbf{v} \in V$$

and its finite element approximation $\boldsymbol{\psi}_{j,h} \in V_h$, we have thanks to triangle's inequality

$$\begin{aligned} \|\nabla \mathbf{w}_j\|_{L^2(D)} &\leq \|\nabla(\boldsymbol{\psi}_j - \mathbf{w}_j)\|_{L^2(D)} + \|\nabla(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j,h})\|_{L^2(D)} + \|\nabla(\boldsymbol{\psi}_{j,h} - \mathbf{w}_{j,h})\|_{L^2(D)} + \|\nabla \mathbf{w}_{j,h}\|_{L^2(D)} \\ &\leq \|r_j(\cdot; \mathbf{u}_0, p_0) - r_j(\cdot; \mathbf{u}_{0,h}, p_{0,h})\|_{V'} + \|\nabla(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j,h})\|_{L^2(D)} + \|\nabla \mathbf{w}_{j,h}\|_{L^2(D)} \\ &\leq C_4 h^s + \|\nabla \mathbf{w}_{j,h}\|_{L^2(D)} \end{aligned}$$

where $s \in (0, 1]$ depends only on the regularity of \mathbf{u}_0 , p_0 , $\boldsymbol{\psi}_j$, $j = 1, \dots, L$, and the domain D [53, 72] and C_4 is independent of h and ε but depends on the mesh aspect ratio, $|\mathbf{u}_0|_{H^{1+s}(D)}$, $|p_0|_{H^s(D)}$ and $|\boldsymbol{\psi}_j|_{H^{1+s}(D)}$, $j = 1, \dots, L$. \square

Based on Propositions 3.2.14 and 3.2.16, we can define two computable error estimators $\eta = (\eta_h^2 + \eta_\varepsilon^2)^{\frac{1}{2}}$ and $\hat{\eta} = (\hat{\eta}_h^2 + \hat{\eta}_\varepsilon^2)^{\frac{1}{2}}$, where η_h and η_ε are defined in (3.73) and $\hat{\eta}_\varepsilon$ is defined in (3.77). From a computational point of view, the computation of $\hat{\eta}$ requires the solution of L additional (linear) problems compared to the cost of getting the error estimator η . However, the gain of the second error estimator is twofold: it does not use the triangle inequality to bound each term of $r(\mathbf{v}; \mathbf{y})$ separately and it does not require the use of the Poincaré inequality. The numerical tests of the next section provide an illustration of the theoretical results obtained so far.

3.2.6 Numerical results

We present now two numerical examples to test the error estimators derived in the previous section. We consider the problem of a flow past a cylinder and consider two different types of perturbation of the domain, namely a perturbation along the vertical axis of the position of the cylinder and a perturbation of its shape. The true error $\|\mathbf{u} - \mathbf{u}_{0,h}, p - p_{0,h}\|$ is approximated

with the standard Monte Carlo method using

$$\|\mathbf{v}, q\| \approx \left(\frac{1}{K} \sum_{k=1}^K \left\{ \nu \|\nabla \mathbf{v}(\mathbf{y}_k)\|_{L^2(D)}^2 + \frac{1}{\nu} \|q(\mathbf{y}_k)\|_{L^2(D)}^2 \right\} \right)^{\frac{1}{2}}$$

where $\{\mathbf{y}_k\} \in \Gamma$ are i.i.d. realizations of the random vector \mathbf{Y} . We choose a sample size of $K = 1000$ in which case the variance of the estimation of the error is at least a factor $2 \cdot 10^{-4}$ smaller than the estimated error in all considered test cases. In what follows, whenever we refer to *error* it should be understood that the true error has been computed by the Monte Carlo procedure. Finally, the approximate solution $(\mathbf{u}_{0,h}, p_{0,h})$ is computed using $\mathbb{P}_{1b} - \mathbb{P}_1$ finite elements and, since the exact solution (\mathbf{u}, p) of the problem is not known, we compute a reference solution using $\mathbb{P}_2 - \mathbb{P}_1$ finite elements on the finest mesh considered.

First example

For this first problem, based on a well-known benchmark problem described in [110], we consider the geometry presented in Figure 3.5 and assume that it corresponds to the reference domain D . More precisely, D consists of the rectangle $[a_1, b_1] \times [a_2, b_2]$ with a hole of radius R located at $\mathbf{c} = (c_1, c_2)$. We assume that the rectangle is fixed and that the center \mathbf{c} of the cylinder is randomly moved along the vertical axis, namely that it is given by $(c_1, c_2 + \varepsilon Y)$ in D_ω with Y a uniform random variable in $[-1, 1]$. We take $\tilde{\mathbf{f}} = \mathbf{0}$ and we prescribe the following inflow and

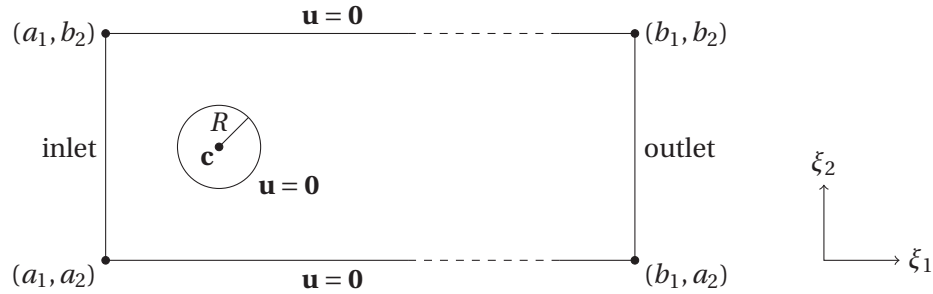


Figure 3.5: Geometry with prescribed boundary conditions for the first example.

outflow (parabolic) velocity profile on the inlet and outlet part of ∂D_ω

$$\tilde{\mathbf{u}}(a_1, x_2) = \tilde{\mathbf{u}}(b_1, x_2) = (4U_{\max}(x_2 - a_2)(b_2 - x_2)/(b_2 - a_2)^2, 0)^T \quad \text{for } a_2 \leq x_2 \leq b_2,$$

with a maximum velocity $U_{\max} = 0.3$ achieved at $x_2 = \frac{a_2 + b_2}{2}$. We impose homogeneous Dirichlet boundary conditions on the remaining parts of the boundary. The Reynolds number is then given by $\frac{2}{3}U_{\max}(2R)\nu^{-1}$, where $\frac{2}{3}U_{\max}$ corresponds to the mean velocity.

We choose a mapping \mathbf{x}_ω , consistent with the perturbation mentioned above, such that all the boundary nodes are fixed. In such a case, the boundary conditions for the equivalent problem on D are the same than the ones on D_ω . More precisely, we consider the mapping $\mathbf{x}_\omega : D \rightarrow D_\omega$

given componentwise by:

$$\begin{cases} x_1 &= \xi_1 \\ x_2 &= \xi_2 + \varepsilon \varphi_1(\xi_1) \varphi_2(\xi_2) Y(\omega), \end{cases}$$

where for $i = 1, 2$

$$\varphi_i(\xi_i) = \begin{cases} \frac{\xi_i - a_i}{c_i - R - a_i} - \tau \frac{(\xi_i - a_i)(\xi_i - c_i + R)}{(c_i - R - a_i)^2} & \text{if } \xi_i \in [a_i, c_i - R[\\ 1 & \text{if } \xi_i \in [c_i - R, c_i + R] \\ \frac{\xi_i - b_i}{c_i + R - b_i} - \tau \frac{(\xi_i - b_i)(\xi_i - c_i - R)}{(c_i + R - b_i)^2} & \text{if } \xi_i \in]c_i + R, b_i], \end{cases} \quad (3.79)$$

which can be written under the form (3.48) as $\mathbf{x}(\boldsymbol{\xi}, \omega) = \boldsymbol{\xi} + \varepsilon \boldsymbol{\varphi}(\boldsymbol{\xi}) Y(\omega) / \sqrt{3}$ with Y a uniform random variable in $[-\sqrt{3}, \sqrt{3}]$ and $\boldsymbol{\varphi}(\boldsymbol{\xi}) = (0, \varphi_1(\xi_1) \varphi_2(\xi_2))^T$. The function φ_2 alone fits the required perturbation of the domain but we use the function φ_1 to fix the nodes on the inlet and outlet boundaries. Moreover, the parameter $\tau \in \{0, 1\}$ is used to control the regularity of the mapping. Indeed, choosing $\tau = 1$ implies that all the functions appearing in the Jacobian matrix A^{-1} of the mapping \mathbf{x}_ω are continuous. From now on, according to [110], we fix the value of the various geometry parameters to $a_1 = a_2 = 0$, $b_1 = 2.2$, $b_2 = 0.41$, $c_1 = c_2 = 0.2$ and $R = 0.05$, and we choose $\tau = 1$. The functions φ_1 and φ_2 for these values of the various geometrical parameters are given in Figure 3.6.

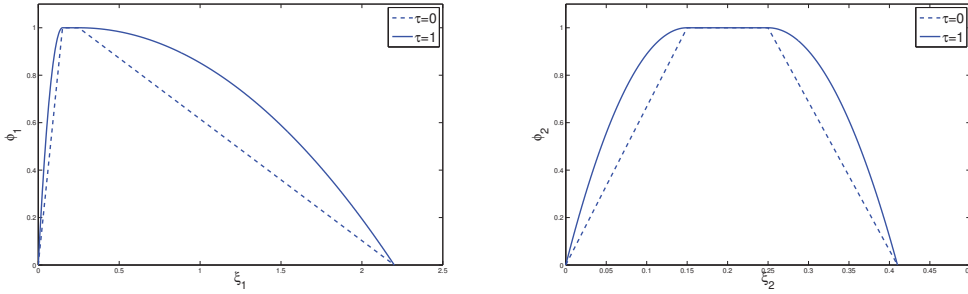


Figure 3.6: Functions $\varphi_1(\xi_1)$, $\xi_1 \in [0, 2.2]$ (left) and $\varphi_2(\xi_2)$, $\xi_2 \in [0, 0.41]$ (right) defined in (3.79).

The numerical tests are performed using FreeFem++ 3.19.1-1 [78]. The mesh is constructed with a Delaunay triangulation using n equispaced points on the left and right boundaries, $5n$ on the upper and lower boundaries and $2n$ along the hole. The mesh size is then given by $h \approx (\sqrt{2}n)^{-1}$ while the number of elements and vertices are about $12n^2$ and $7n^2$, respectively. Notice that we are using piecewise linear triangular elements to mesh the physical domain D whose boundary has a curved part, namely the hole modelling the cylinder. We are not accounting this error here and we refer to [31, Chapter 10] or [48, Chapter VI] for an analysis of such *variational crime*, introducing for instance *isoparametric finite elements*. Finally, we recall that the error estimates derived in Sections 3.2.5 and 3.2.5 are valid for homogeneous Dirichlet boundary conditions. In the case of inhomogeneous conditions, as considered here, an additional term due to the approximation of the Dirichlet data should be included.

However, thanks to the fact that the later is not affected by the mapping, it is a higher order term in h (see for instance [16]) and thus we do not take it into account in the numerical results.

Deterministic case

We first consider the deterministic case, namely when ε is set to zero. The reference values in [110] include the drag (c_D) and lift (c_L) coefficients and the pressure difference $\Delta p = p(0.15, 0.2) - p(0.25, 0.2)$ between the value at the front and the end point of the cylinder. Using $\mathbb{P}_2 - \mathbb{P}_1$ FE on a mesh with $n = 80$, we obtain the values $c_D = 5.57469$, $c_L = 0.0104584$ and $\Delta p = 0.117525$ which are consistent with the bounds given in [110].

We give in Figure 3.7 the velocity magnitude, the two components u_1 and u_2 and the pressure obtained using $\mathbb{P}_2 - \mathbb{P}_1$ finite elements on the finest mesh, i.e. $n = 64$.

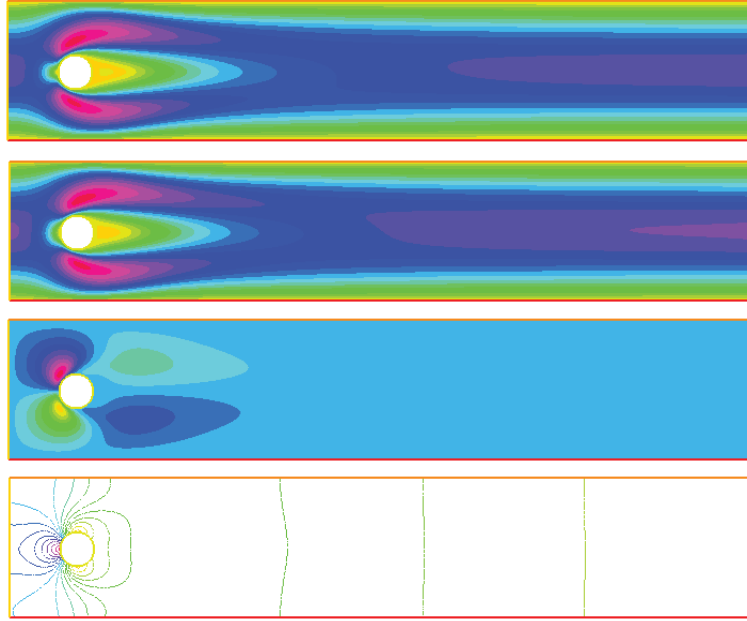


Figure 3.7: Velocity magnitude, components u_1 and u_2 and pressure for the first problem in the case $\varepsilon = 0$ and $\nu = 0.001$.

In Table 3.4, we give the results obtained for various values of n and ν , where err , η and e.i. denote respectively the error, the error estimator $(\eta_h^2 + \eta_\varepsilon^2)^{\frac{1}{2}}$ with η_h and η_ε defined in (3.73) and the effectivity index, namely the ratio between the error estimator and the error. Notice that $\eta_\varepsilon = 0$ here since $\varepsilon = 0$. We can see that in all cases, for h small enough, the effectivity index is about 2.8. This value is consistent with the one obtained in Appendix 1.C, see Table 1.15.

	$\nu = 0.001$			$\nu = 0.01$			$\nu = 0.1$			$\nu = 1$		
n	err	η	e.i.	err	η	e.i.	err	η	e.i.	err	η	e.i.
4	0.136	0.566	4.17	0.158	0.310	1.96	0.514	0.963	1.87	1.628	3.052	1.87
8	0.039	0.150	3.87	0.060	0.135	2.27	0.188	0.415	2.20	0.596	1.312	2.20
16	0.015	0.044	2.87	0.028	0.070	2.55	0.086	0.216	2.52	0.271	0.684	2.52
32	0.007	0.019	2.73	0.013	0.034	2.70	0.039	0.105	2.69	0.124	0.333	2.69
64	0.003	0.009	2.75	0.006	0.017	2.78	0.019	0.052	2.78	0.060	0.166	2.78

Table 3.4: Error, error estimator and effectivity index for the deterministic case ($\varepsilon = 0$) and various viscosities for the first example.

Random case

We treat now the random case by considering values of ε between 0 and 0.05. With $\varepsilon = 0.05$, the random position of the cylinder on the vertical axis lies between 0.15 and 0.25 with nominal value in 0.2, which is quite a large perturbation considering that the height of the rectangle is equal to 0.41.

The velocity magnitude for the case $\nu = 0.001$ when the cylinder is moved from 0.2 to 0.25 is given in Figure 3.8. We plot the solution obtained when performing the computation on the physical domain and on the reference domain, with the appropriate modification of the coefficients in the equations for the latter case. The solution for the case $\varepsilon = 0$ is again given for comparison.

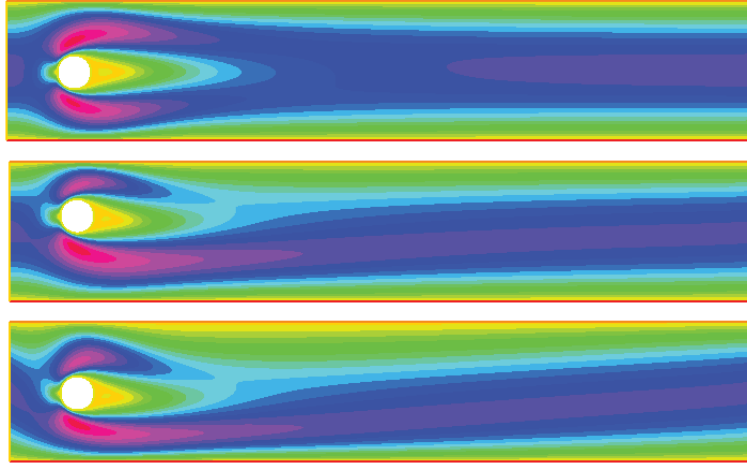


Figure 3.8: Velocity magnitude for $\nu = 0.001$ in the case $\varepsilon = 0$ (top) and $\varepsilon = 0.05$ with $Y = 1$ computed on D_ω (middle) and on D (bottom) for the first example.

We give in Table 3.5 the numerical results obtained for $\nu = 0.001$ and $\nu = 1$ and various values of n and ε .

3.2. Steady-state incompressible Navier-Stokes equations in random domains

n	ε	$\nu = 0.001$				$\nu = 1$			
		err	η_h	η_ε	e.i.	err	η_h	η_ε	e.i.
4	0.05	0.1389	0.5656	1.0649	8.68	1.8881	3.0521	2.4890	2.09
8	0.05	0.0591	0.1503	0.6797	11.78	1.0157	1.3124	2.3458	2.65
16	0.05	0.0452	0.0440	0.5487	12.19	0.8110	0.6839	2.3018	2.96
32	0.05	0.0429	0.0190	0.5288	12.32	0.7713	0.3333	2.2887	3.00
64	0.05	0.0428	0.0091	0.5246	12.25	0.7526	0.1655	2.2856	3.05
4	0.025	0.1361	0.5656	0.5325	5.71	1.6989	3.0521	1.2445	1.94
8	0.025	0.0436	0.1503	0.3399	8.52	0.7159	1.3124	1.1729	2.46
16	0.025	0.0249	0.0440	0.2743	11.15	0.4701	0.6839	1.1509	2.85
32	0.025	0.0205	0.0190	0.2644	12.96	0.3916	0.3333	1.1444	3.04
64	0.025	0.0194	0.0091	0.2623	13.51	0.3831	0.1655	1.1428	3.01
4	0.0125	0.1356	0.5656	0.2662	4.61	1.6458	3.0521	0.6223	1.89
8	0.0125	0.0401	0.1503	0.1699	5.66	0.6291	1.3124	0.5865	2.29
16	0.0125	0.0181	0.0440	0.1372	7.98	0.3310	0.6839	0.5755	2.70
32	0.0125	0.0119	0.0190	0.1322	11.25	0.2264	0.3333	0.5722	2.92
64	0.0125	0.0100	0.0091	0.1311	13.13	0.2056	0.1655	0.5714	2.89
4	0.00625	0.1356	0.5656	0.1331	4.29	1.6324	3.0521	0.3111	1.88
8	0.00625	0.0392	0.1503	0.0850	4.41	0.6043	1.3124	0.2932	2.23
16	0.00625	0.0160	0.0440	0.0686	5.08	0.2872	0.6839	0.2877	2.58
32	0.00625	0.0084	0.0190	0.0661	8.17	0.1559	0.3333	0.2861	2.82
64	0.00625	0.0058	0.0091	0.0656	11.45	0.1117	0.1655	0.2857	2.96
4	0.003125	0.1355	0.5656	0.0666	4.20	1.6324	3.0521	0.1556	1.88
8	0.003125	0.0389	0.1503	0.0425	4.01	0.6043	1.3124	0.1466	2.23
16	0.003125	0.0155	0.0440	0.0343	3.60	0.2872	0.6839	0.1439	2.58
32	0.003125	0.0074	0.0190	0.0330	5.18	0.1328	0.3333	0.1430	2.73
64	0.003125	0.0041	0.0091	0.0328	8.32	0.0760	0.1655	0.1429	2.88

Table 3.5: The error, the two contributions η_h and η_ε of the error estimator in (3.73) and the effectivity index for $\nu = 0.001$ and $\nu = 1$ for the first example.

We recall that we use different FE spaces for the reference and the approximate solution and thus, even in the case where the same mesh is used for both solutions, there is still an error due to space discretization. We can see in Table 3.5 that the effectivity index tends to the one obtained in Table 3.4 when the spatial error is dominating while when the statistical error dominates, it is about 13 and 3 for $\nu = 0.001$ and $\nu = 1$, respectively. This highlights the dependence of the error estimate given in Section 3.2.5 with respect to the input data. However, we can see that when both h and ε are divided by 2 then the effectivity index remains constant, this observation being tempered by the fact that the effectivity index for $\varepsilon = 0$ is not constant for the various meshes considered (see Table 3.4). For instance, in the case $\nu = 0.001$ and $\varepsilon = (5n)^{-1}$, which corresponds to $h \approx 3.5\varepsilon$, the effectivity index is about 8. We study now the efficiency of the second error estimate with respect to the viscosity. In Figure 3.9, we give the effectivity index with respect to ν for both error estimators η and $\hat{\eta} = (\eta_h^2 + \hat{\eta}_\varepsilon^2)^{\frac{1}{2}}$, where $\hat{\eta}_\varepsilon$ is given in (3.77), in the case $\varepsilon = 0.025$, $n = 64$ and $n_{ref} = 64$, which corresponds to a statistical error dominant regime.

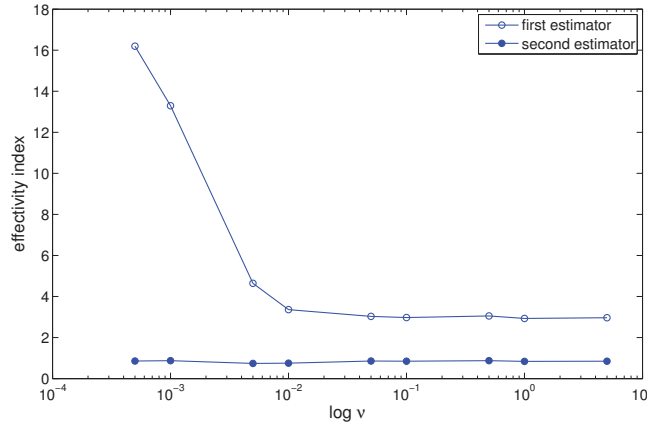


Figure 3.9: Effectivity index with respect to the viscosity ν for the two error estimators η and $\hat{\eta}$ defined in (3.73) and (3.77) for the first example.

We can see that the effectivity index of the first error estimator η remains constant for viscosities greater than 0.01 while below this value, it starts increasing as ν decreases. The situation is different for the second estimator $\hat{\eta}$ of Section 3.2.5, whose efficiency is not sensitive to the value of ν .

Remark 3.2.17. *In order to have the correct balance of the two terms appearing in the error estimator η or $\hat{\eta}$, we could estimate numerically the constants in front of each term η_h and η_ε or $\hat{\eta}_\varepsilon$. The estimation of these constants can also be used to construct a sharp error estimator, namely an error estimator with effectivity index close to 1. According to the results in Table 3.4, the term η_h should be multiplied by a factor 1/2.8. For the term due to uncertainty, we obtain that $\hat{\eta}_\varepsilon$ should be multiplied by about 1.5, considering for instance same FE spaces and fine mesh for both the reference and approximate solutions, whereas the constant in front of η_ε*

depends on the viscosity as seen in Table 3.5 or Figure 3.9 (for instance $1/13$ for $\nu = 0.001$ or $1/3$ for $\nu \geq 0.01$).

To conclude the analysis of this first example, we mention that similar results are obtained if we use homogeneous Neumann boundary conditions on the outlet part of the boundary. Notice that in this case, the jump term should be modified appropriately since it is no longer zero on the boundary edges belonging to the outlet.

Second example

For this second example, the reference geometry D consists in a square $[-H, H]^2$ with $H = 0.5$ and a circular hole of radius $R = 0.15$ centred at the origin, as depicted in Figure 3.10 where the prescribed boundary conditions are also indicated. The shape of the hole is given on D

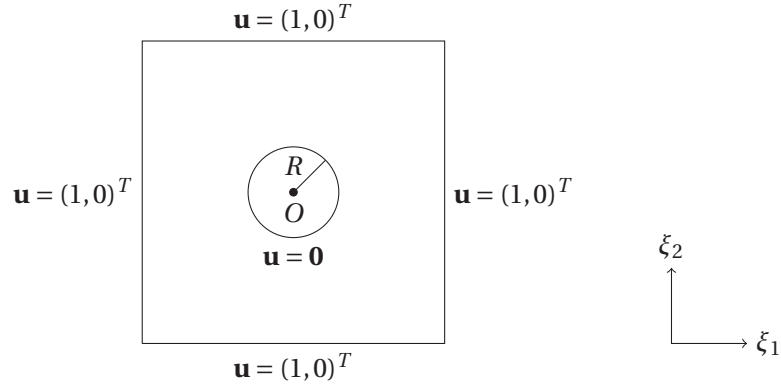


Figure 3.10: Geometry with prescribed boundary conditions for the second example.

by $(\xi_1, \xi_2) = (R \cos(\theta), R \sin(\theta))$ with $\theta \in [0, 2\pi]$. We perturb this hole by modifying its radius with respect to the angle by the formula $R + \varepsilon d_\theta$, where $d_\theta = \sum_{j=1}^L \alpha_j \cos(k_j \theta) Y_j$ and Y_j are i.i.d uniform random variables in $[-1, 1]$. The coefficients k_j and α_j control the frequency and the amplitude of each term, respectively. We mention that a similar perturbation is considered in [125], where the mapping is not constructed explicitly but computed through solutions of Laplace equations. We consider here the following mapping \mathbf{x}_ω from D to D_ω which fits the above perturbation: denoting $r = \sqrt{\xi_1^2 + \xi_2^2}$ and $\theta = \arctan(\frac{\xi_2}{\xi_1})$ the polar coordinates of any point $\xi = (\xi_1, \xi_2)$ of D , we take

$$\mathbf{x} = \xi + \varepsilon \sum_{j=1}^L \boldsymbol{\varphi}_j(\xi) Y_j(\omega), \quad \boldsymbol{\varphi}_j(\xi) = \alpha_j \cos(k_j \theta) g(\xi) \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \quad (3.80)$$

where the *cutoff function* g is such that it vanishes at the boundary of the domain and is equal to 1 in the hole, namely we use

$$g(\xi) = \begin{cases} 1 & \text{if } r \in [0, R] \\ \frac{(\xi_1^2 - H^2)(\xi_2^2 - H^2)}{(R^2 \xi_1^2 r^{-2} - H^2)(R^2 \xi_2^2 r^{-2} - H^2)} & \text{otherwise.} \end{cases} \quad (3.81)$$

The graph of this function is depicted in Figure 3.11

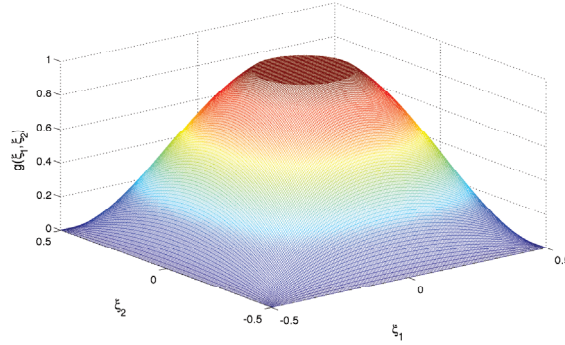


Figure 3.11: Function $g = g(\xi_1, \xi_2)$ defined in (3.81).

The mesh is again built with a Delaunay triangulation using n equispaced points on the boundaries of the square and $2n$ on the hole for various values of n with corresponding mesh size $h \approx 1.5n^{-1}$ and number of elements and vertices of about $3.5n^2$ and $2n^2$, respectively.

Remark 3.2.18. *Contrary to the previous example, the choice of the boundary conditions on the outlet has an impact on the solution of this problem, due to the fact that the outlet is close to the cylinder. This is especially true for small viscosities, in which case some flow is re-entering the domain when homogeneous Neumann conditions are used while the solution presents a boundary layer when Dirichlet conditions are enforced.*

Deterministic case

We consider first the deterministic case taking $\varepsilon = 0$. The plot of the velocity magnitude, the two components u_1 and u_2 and the pressure obtained using $\mathbb{P}_2 - \mathbb{P}_1$ FE and the finest mesh ($n = 160$) is given in Figure 3.12.

Moreover, we give in Table 3.6 the results we get for various values of n and ν . Similarly to the previous example, the effectivity index is about 2.8 in all cases, when h is small enough.

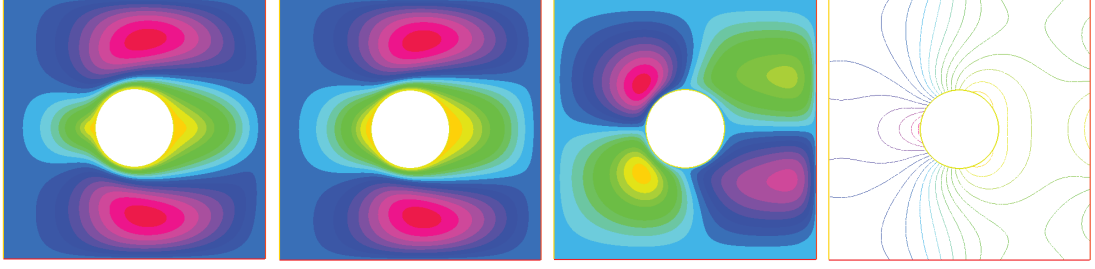


Figure 3.12: From left to right: velocity magnitude, components u_1 and u_2 and pressure for the second problem in the case $\varepsilon = 0$ and $\nu = 0.05$.

	$\nu = 0.05$			$\nu = 0.1$			$\nu = 0.5$			$\nu = 1$		
n	err	η	e.i.	err	η	e.i.	err	η	e.i.	err	η	e.i.
10	0.477	1.149	2.41	0.621	1.405	2.26	1.364	2.988	2.19	1.930	4.221	2.19
20	0.230	0.579	2.51	0.278	0.697	2.51	0.590	1.470	2.49	0.833	2.074	2.49
40	0.112	0.294	2.63	0.132	0.353	2.67	0.279	0.745	2.68	0.393	1.052	2.68
80	0.055	0.148	2.71	0.064	0.176	2.75	0.134	0.371	2.76	0.190	0.523	2.76
160	0.026	0.073	2.77	0.031	0.087	2.80	0.066	0.184	2.80	0.096	0.259	2.80

Table 3.6: Error, error estimator and effectivity index for the deterministic case ($\varepsilon = 0$) and various viscosities for the second example.

Random case

We consider first $L = 1$ random variable, we fix $\alpha_1 = 1$ and $k_1 = 6$ in the definition of d_θ and we let $0 \leq \varepsilon \leq 0.01$. The vorticity of the velocity \mathbf{u} and the pressure p in the case $\varepsilon = 0.01$, $\nu = 0.05$ and $Y = 1$ is given in Figure 3.13, where the solution obtained by solving the problem defined on D_ω as well as the solution for the case $\varepsilon = 0$ are also given for comparison.

We give in Table 3.7 the numerical results obtained for $\nu = 0.05$ and $\nu = 1$ and various values of n and ε .

Similarly to the previous example, we observe that the effectivity index tends to the one obtained for the deterministic case ($\varepsilon = 0$) when the error in h is dominating, while it is about 6 and 1.5 for $\nu = 0.05$ and $\nu = 1$, respectively, when the statistical error dominates. This shows again the sensitivity of the efficiency of the first error estimator with respect to the input data but, as before, the effectivity index remains about constant when both h and ε are divided by 2. Indeed, for instance for $\nu = 0.05$ and $\varepsilon = (10n)^{-1}$, corresponding to $h \approx 15\varepsilon$, it stays between 3.81 and 4.05. Finally, the same behaviour than in the previous example is observed for the efficiency of the second error estimator $\hat{\eta}$ with respect to the viscosity, as can be seen in Figure 3.14 where the results are given for the case $\varepsilon = 0.005$, $n = 160$ and $n_{ref} = 160$.

The results are similar when we consider other kinds of perturbation. For instance, let consider

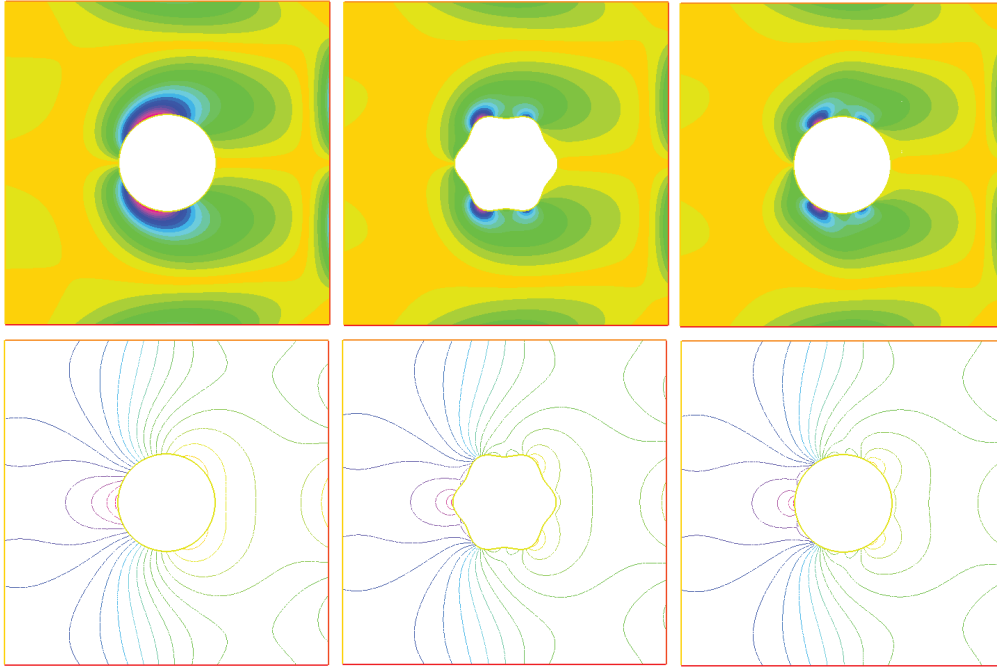


Figure 3.13: Vorticity of the velocity and pressure for $\nu = 0.05$ in the case $\varepsilon = 0$ (left) and $\varepsilon = 0.01$ with $Y = 1$ computed D_ω (middle) and on D (right) for the second example.

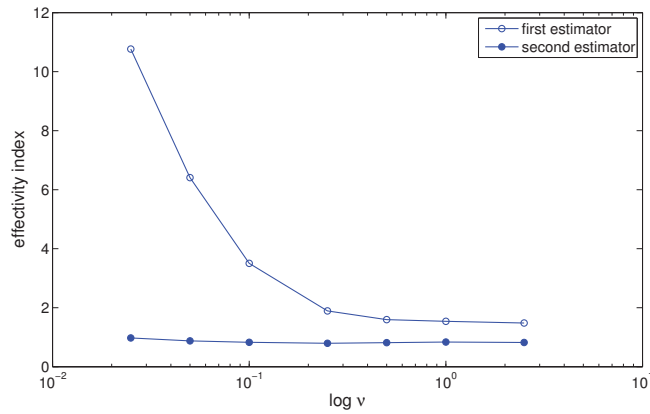


Figure 3.14: Effectivity index with respect to the viscosity ν for the two error estimators η and $\hat{\eta}$ defined in (3.73) and (3.77) for the second example.

3.2. Steady-state incompressible Navier-Stokes equations in random domains

n	ε	$\nu = 0.05$				$\nu = 1$			
		err	η_h	η_ε	e.i.	err	η_h	η_ε	e.i.
10	0.01	0.5125	1.1492	1.6181	3.87	2.0403	4.2209	1.4479	2.19
20	0.01	0.3251	0.5785	1.5682	5.14	1.2200	2.0741	1.3862	2.04
40	0.01	0.2625	0.2937	1.5552	6.03	1.0216	1.0524	1.3730	1.69
80	0.01	0.2486	0.1478	1.5519	6.27	1.0040	0.5233	1.3696	1.46
160	0.01	0.2431	0.07279	1.5511	6.39	0.9630	0.2594	1.3687	1.45
10	0.005	0.4859	1.1492	0.8090	2.89	1.9575	4.2209	0.7240	2.19
20	0.005	0.2556	0.5785	0.7841	3.81	0.9477	2.0741	0.6931	2.31
40	0.005	0.1628	0.2937	0.7776	5.11	0.6163	1.0524	0.6865	2.04
80	0.005	0.1340	0.1478	0.7759	5.91	0.5149	0.5233	0.6848	1.67
160	0.005	0.1238	0.0728	0.7755	6.29	0.4891	0.2594	0.6843	1.50
10	0.0025	0.4792	1.1492	0.4045	2.54	1.9363	4.2209	0.3620	2.19
20	0.0025	0.2370	0.5785	0.3921	2.95	0.8602	2.0741	0.3465	2.44
40	0.0025	0.1263	0.2937	0.3888	3.86	0.4538	1.0524	0.3433	2.44
80	0.0025	0.0808	0.1478	0.3880	5.14	0.3085	0.5233	0.3424	2.03
160	0.0025	0.0662	0.0728	0.3878	5.96	0.2584	0.2594	0.3422	1.66
10	0.00125	0.4776	1.1492	0.2023	2.44	1.9317	4.2209	0.1810	2.19
20	0.00125	0.2319	0.5785	0.1960	2.63	0.8399	2.0741	0.1733	2.48
40	0.00125	0.1154	0.2937	0.1944	3.05	0.4098	1.0524	0.1716	2.60
80	0.00125	0.0624	0.1478	0.1940	3.91	0.2237	0.5233	0.1712	2.46
160	0.00125	0.0405	0.0728	0.1939	5.12	0.1517	0.2594	0.1711	2.05
10	0.000625	0.4772	1.1492	0.1011	2.42	1.9304	4.2209	0.0905	2.19
20	0.000625	0.2306	0.5785	0.0980	2.54	0.8347	2.0741	0.0866	2.49
40	0.000625	0.1125	0.2937	0.0972	2.75	0.3977	1.0524	0.0858	2.66
80	0.000625	0.0565	0.1479	0.0970	3.13	0.1987	0.5233	0.0856	2.67
160	0.000625	0.0304	0.0728	0.0970	3.99	0.1101	0.2594	0.0855	2.48

Table 3.7: The error, the two contributions η_h and η_ε of the estimator in (3.73) and the effectivity index for $\nu = 0.05$ and $\nu = 1$.

(3.80) with $L = 2$ with $k_1 = 6$, $k_2 = 11$, $\alpha_1 = 1$ and $\alpha_2 = 0.8$. The results we obtained, given in Figure 3.15 and in Table 3.8, are very similar to those presented in Table 3.7. The results for the second error estimator $\hat{\eta}$ with the estimated constant, see Remark 3.2.17, are also provided. We can see that for h small enough, namely when the effectivity index for the spatial error estimator is about 2.8 (see Table 3.6), the error estimator is sharp.

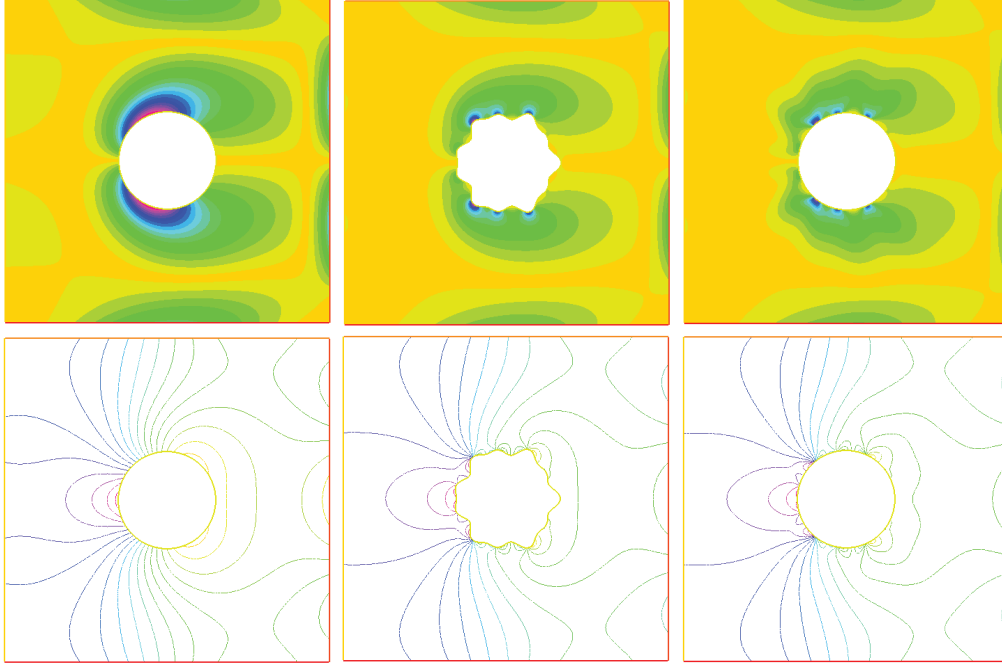


Figure 3.15: Vorticity of the velocity and pressure for $\nu = 0.05$ in the case $\varepsilon = 0$ (left) and $\varepsilon = 0.01$ with $Y = 1$ computed D_ω (middle) and on D (right) for the second example with $L = 2$.

Conclusions

In this chapter, we have considered steady-state nonlinear PDEs on random domains, namely the one-dimensional viscous Burgers' equation and the incompressible Navier-Stokes equations. We have used the *domain mapping method* to transform them into PDEs on a fixed reference domain with random coefficients.

We have first studied the deterministic Burgers' equation with mixed Dirichlet-Neumann boundary conditions. We have shown the well-posedness of the problem under suitable assumptions on the input data and we have derived an *a posteriori* error estimate. Then, the case of random intervals has been considered, performing all the analysis on the fixed reference domain. Finally, we have presented two numerical examples both in the deterministic and random cases.

For the Navier-Stokes equations, we started the analysis by showing the well-posedness of the

3.2. Steady-state incompressible Navier-Stokes equations in random domains

n	ε	err	η_h	η_ε	η/err	$\eta_h/2.8$	$1.5\hat{\eta}_\varepsilon$	$\hat{\eta}/\text{err}$
10	0.005	0.4994	1.1492	1.4301	3.67	0.4104	0.1849	0.90
20	0.005	0.2924	0.5785	1.3884	5.14	0.2066	0.1992	0.98
40	0.005	0.22061	0.2937	1.3768	6.38	0.1049	0.2054	1.05
80	0.005	0.1983	0.1478	1.3739	6.97	0.0528	0.2072	1.08
160	0.005	0.1928	0.0728	1.3732	7.13	0.0260	0.2077	1.09
10	0.0025	0.4826	1.1492	0.7151	2.80	0.4104	0.0924	0.87
20	0.0025	0.2477	0.5785	0.6942	3.65	0.2066	0.0996	0.93
40	0.0025	0.1464	0.2937	0.6884	5.11	0.1049	0.1027	1.00
80	0.0025	0.1080	0.1478	0.6869	6.51	0.0528	0.1036	1.08
160	0.0025	0.0988	0.0728	0.6866	6.99	0.0260	0.1038	1.08
10	0.00125	0.4784	1.1492	0.3575	2.52	0.4104	0.0462	0.86
20	0.00125	0.2345	0.5785	0.3471	2.88	0.2066	0.0498	0.91
40	0.00125	0.1212	0.2937	0.3442	3.73	0.1049	0.0513	0.96
80	0.00125	0.0731	0.1478	0.3435	5.12	0.0528	0.0518	1.01
160	0.00125	0.0545	0.0728	0.3433	6.44	0.0260	0.0519	1.06

Table 3.8: Effectivity index of the two error estimators in the case $\nu = 0.05$ for the second example with $L = 2$.

problem under suitable assumptions on the input data and the mapping, before performing an *a posteriori* error analysis. Using a perturbation method, we obtained two error estimates for the *first order* approximation $(\mathbf{u}, p) \approx (\mathbf{u}_{0,h}, p_{0,h})$. Both estimates are constituted of two parts, namely one part due to space discretization in h and one due to the uncertainty in ε . They already give useful information, especially when the problem contains small uncertainties. They can indeed be used to adaptively find a spatial mesh that balances the two sources of error. Further mesh refinement should then be avoided since it would not decrease the total error, the statistical error being dominant. The latter can only be decreased by adding more terms in the expansion of the solution. Notice that if we want to analyse higher order approximations in ε , then we should impose additional regularity assumptions on f and on the random mapping, namely that the Jacobian matrix $\nabla \boldsymbol{\varphi}_j$ belongs to $[W^{1,\infty}(D)]^{d \times d}$ for $j = 0, 1, \dots, L$ and not only for $j = 0$. Indeed, we have that the residual for the FE approximation $(U_{j,h}, P_{j,h})$ of (U_j, P_j) belongs to $L^2(D)$ for $j = 1, \dots, L$, where (U_j, P_j) is the solution of (3.59) and appears in the second term of the expansion of the solution. The same holds for the residual of the higher order terms.

Each of the two error estimators η and $\hat{\eta}$ that we obtained presents its advantages and drawbacks. The first one can be computed by solving only one nonlinear problem, namely the *standard* Navier-Stokes equations on the reference domain. We have seen however that the sharpness of this estimator might be affected when changing the input data, as predicted by the theory. In the two numerical examples considered here, the effectivity index remains constant for moderate Reynolds numbers but then starts to increase as the viscosity diminishes. The second error estimator shows promising results, its efficiency being indeed independent

of the input data for all the cases we have considered. The extra cost to pay is the resolution of L additional linear problems. Finally, as mentioned in Remark 3.2.17, the constant in front of the two terms in h and ε can be estimated numerically (once for all for the second estimator) to get a sharp error estimator, that is an estimator with effectivity index close to 1.

3.A Derivation of problems (3.58) and (3.59)

We give here some details about the derivation of the problems (3.58) and (3.59) that we need to solve to obtain the first two terms in the expansion of the solution (\mathbf{u}, p) , namely (\mathbf{u}_0, p_0) and (\mathbf{u}_1, p_1) . These problems are obtained by replacing each term in (3.35), the problem in strong form for (\mathbf{u}, p) , by its expansion with respect to ε and keeping only the appropriate terms. Using relations (3.55) and (3.56), we can write

$$\begin{aligned} J_{\mathbf{x}} A A^T &= (1 + \varepsilon \operatorname{tr}(A_1) + \mathcal{O}(\varepsilon^2))(I - \varepsilon A_1 + \mathcal{O}(\varepsilon^2))(I - \varepsilon A_1^T + \mathcal{O}(\varepsilon^2)) \\ &= I + \varepsilon(\operatorname{tr}(A_1)I - A_1 - A_1^T) + \mathcal{O}(\varepsilon^2) \end{aligned}$$

and similarly

$$J_{\mathbf{x}} A^T = I + \varepsilon(\operatorname{tr}(A_1)I - A_1^T) + \mathcal{O}(\varepsilon^2).$$

Therefore, considering for instance the convection term, we get

$$\begin{aligned} (\mathbf{u} \cdot J_{\mathbf{x}} A^T \nabla) \mathbf{u} &= ((\mathbf{u}_0 + \varepsilon \mathbf{u}_1 + \mathcal{O}(\varepsilon^2)) \cdot (I + \varepsilon(\operatorname{tr}(A_1)I - A_1^T) + \mathcal{O}(\varepsilon^2)) \nabla) (\mathbf{u}_0 + \varepsilon \mathbf{u}_1 + \mathcal{O}(\varepsilon^2)) \\ &= (\mathbf{u}_0 \cdot \nabla) \mathbf{u}_0 + \varepsilon [(\mathbf{u}_1 \cdot \nabla) \mathbf{u}_0 + (\mathbf{u}_0 \cdot \nabla) \mathbf{u}_1 + (\mathbf{u}_0 \cdot (\operatorname{tr}(A_1)I - A_1^T) \nabla) \mathbf{u}_0] + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Proceeding similarly for all the terms involved in the first equation of (3.35) and keeping the $\mathcal{O}(1)$ terms with respect to ε we obtain

$$-\nu \Delta \mathbf{u}_0 + (\mathbf{u}_0 \cdot \nabla) \mathbf{u}_0 + \nabla p_0 = \mathbf{f}_0$$

which is the first equation of (3.58). If we collect now the terms of order $\mathcal{O}(\varepsilon)$ we get

$$\begin{aligned} -\nu \Delta \mathbf{u}_1 + (\mathbf{u}_0 \cdot \nabla) \mathbf{u}_1 + (\mathbf{u}_1 \cdot \nabla) \mathbf{u}_0 + \nabla p_1 &= \operatorname{tr}(A_1) \mathbf{f}_0 + \mathbf{f}_1 + \nu \nabla \cdot [((\operatorname{tr}(A_1)I - A_1 - A_1^T) \nabla) \mathbf{u}_0] \\ &\quad - (\mathbf{u}_0 \cdot (\operatorname{tr}(A_1)I - A_1^T) \nabla) \mathbf{u}_0 - ((\operatorname{tr}(A_1)I - A_1^T) \nabla) p_0. \end{aligned} \tag{3.82}$$

Finally, since

$$A_1 = \sum_{j=1}^L \nabla \boldsymbol{\varphi}_j y_j, \quad \mathbf{f}_1 = \sum_{j=1}^L \mathbf{F}_j y_j, \quad \mathbf{u}_1 = \sum_{j=1}^L \mathbf{U}_j y_j \quad \text{and} \quad p_1 = \sum_{j=1}^L P_j y_j,$$

equation (3.82) is satisfied if

$$\begin{aligned}
 -\nu \Delta U_j + (\mathbf{u}_0 \cdot \nabla) \mathbf{U}_j + (\mathbf{U}_j \cdot \nabla) \mathbf{u}_0 + \nabla P_j &= \operatorname{tr}(\nabla \boldsymbol{\varphi}_j) \mathbf{f}_0 + \mathbf{F}_j \\
 &+ \nu \nabla \cdot \left[((\operatorname{tr}(\nabla \boldsymbol{\varphi}_j) I - \nabla \boldsymbol{\varphi}_j - \nabla \boldsymbol{\varphi}_j^T) \nabla) \mathbf{u}_0 \right] \\
 &- (\mathbf{u}_0 \cdot (\operatorname{tr}(\nabla \boldsymbol{\varphi}_j) I - \nabla \boldsymbol{\varphi}_j^T) \nabla) \mathbf{u}_0 \\
 &- ((\operatorname{tr}(\nabla \boldsymbol{\varphi}_j) I - \nabla \boldsymbol{\varphi}_j^T) \nabla) p_0
 \end{aligned} \tag{3.83}$$

for $j = 1, \dots, L$, which is the second equation of problem (3.59). In fact, relations (3.82) and (3.83) are equivalent since the random variables $\{Y_j\}$ are independent, with zero mean and unit variance and thus form an orthonormal set. The second equation of (3.35), corresponding to the incompressibility constraint, is treated analogously.

3.B Choice of the norm

We give here three *justifications* about the choice of the norm on the space $V \times Q$ for the couple (\mathbf{u}, p) , more precisely about the scaling with respect to the kinematic viscosity ν . We claim that the appropriate scaling is given by

$$\|\mathbf{v}, q\|_k^2 := \nu^k \|\nabla \mathbf{v}\|^2 + \nu^{k-2} \|q\|^2 \quad \text{for any choice } k = 0, 1, 2. \tag{3.84}$$

First of all, we can perform a dimensional analysis. The dimension unit of the kinematic viscosity is $[\nu] = \frac{m^2}{s}$ while we have, recall that p corresponds to the pressure divided by the density of the fluid,

$$\|\nabla \mathbf{u}\|^2 = \left(\frac{1}{m} \cdot \frac{m}{s} \right)^2 = \frac{1}{s^2} \quad \text{and} \quad [p^2] = \left(\frac{N}{m^2} \cdot \frac{m^3}{kg} \right)^2 = \frac{m^4}{s^4},$$

from which we deduce that $[\nu^k \|\nabla \mathbf{u}\|^2] = [\nu^{k-2} p^2]$ for all k . This is also the natural choice of scaling that arises when looking at the *a priori* estimates on the solution (\mathbf{u}, p) or when performing *a posteriori* error estimation. For simplicity, let us consider the (deterministic) Stokes problem given under the weak form by:

find $(\mathbf{u}, p) \in V \times Q$ such that

$$\begin{aligned}
 a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= F(\mathbf{v}) & \forall \mathbf{v} \in V \\
 b(\mathbf{u}, q) &= 0 & \forall q \in Q,
 \end{aligned}$$

with $V = [H_0^1(D)]^d$, $Q = L_0^2(D)$, $a(\mathbf{u}, \mathbf{v}) = \nu \int_D \nabla \mathbf{u} : \nabla \mathbf{v}$, $b(\mathbf{v}, q) = - \int_D q \nabla \cdot \mathbf{v}$ and $F(\mathbf{v}) = \int_D \mathbf{f} \cdot \mathbf{v}$. The bilinear form a is continuous and coercive on V with constant ν and b is continuous on V with constant 1 and satisfies the inf-sup condition with constant $\beta = \beta(D)$. The problem is

thus well-posed (see [32]) and the following *a priori* estimates are satisfied

$$\|\nabla \mathbf{u}\| \leq \frac{1}{\nu} \|F\|_{V'} \quad \text{and} \quad \|p\| \leq \frac{1}{\beta} (\|F\|_{V'} + \nu \|\nabla \mathbf{u}\|) \leq \frac{2}{\beta} \|F\|_{V'}.$$

Therefore, we have

$$\nu^{k/2} \|\nabla \mathbf{u}\| + \nu^{k/2-1} \|p\| \leq C \nu^{k/2-1} \|\mathbf{f}\|_{V'} \quad \forall k,$$

where $C = (1 + 2/\beta)$ is independent of ν , which is consistent with the scaling (3.84). Finally, for the *a posteriori* error analysis, denoting $e = \mathbf{u} - \mathbf{u}_h$ and $E = p - p_h$ with \mathbf{u}_h and p_h the finite element approximation of \mathbf{u} and p , respectively, we have for any $(\mathbf{v}, q) \in V \times Q$

$$a(\mathbf{e}, \mathbf{v}) + b(\mathbf{v}, E) + b(\mathbf{e}, q) = R_1(\mathbf{v}) + R_2(q), \quad (3.85)$$

with

$$R_1(\mathbf{v}) := F(\mathbf{v}) - a(\mathbf{u}_h, \mathbf{v}) - b(\mathbf{v}, p_h) \quad \text{and} \quad R_2(q) := -b(\mathbf{u}_h, q).$$

Using relation (3.85), Young's inequality and the properties of a and b , we can easily show that

$$\|E\| \leq \frac{1}{\beta} \|R_1\|_{V'} + \frac{\nu}{\beta} \|\nabla \mathbf{e}\| \quad (3.86)$$

and

$$\nu \|\nabla \mathbf{e}\|^2 \leq \frac{c_1}{\nu} \|R_1\|_{V'}^2 + \frac{c_2 \nu}{\beta^2} \|R_2\|_{Q'}^2 \quad (3.87)$$

with for instance $c_1 = c_2 = 3$, the value of these constants depending only on how we use Young's inequality. From the last two inequalities, we deduce that the scaling (3.84) should be used to get

$$\nu^k \|\nabla \mathbf{e}\|^2 + \nu^{k-2} \|E\|^2 \leq C \left(\nu^{k-2} \|R_1\|_{V'}^2 + \nu^k \|R_2\|_{Q'}^2 \right),$$

where C is a constant independent of ν (but which depends on the inf-sup constant β).

We mention that in a diffusion-dominating regime, the choice $k = 0$ yields a total error $\|\mathbf{e}, E\|_0$ which remains constant when ν varies. Indeed, in such a case the velocity error $\|\nabla \mathbf{e}\|$ is constant while the pressure error $\|E\|$ behaves as ν , i.e. $\frac{1}{\nu} \|E\|$ is constant.

3.C Proof of some properties

Proposition 3.C.1. *Let $A, B, C \in \mathbb{R}^{n \times n}$ be square matrices with coefficients denoted respectively by a_{ij} , b_{ij} and c_{ij} for $1 \leq i, j \leq n$, and let \mathbf{w} be any smooth function with value in \mathbb{R}^n . We then have*

$$AB : CB = ABB^T : C \quad (3.88)$$

and

$$(B^T \nabla) \mathbf{w} = \nabla \mathbf{w} B. \quad (3.89)$$

Proof. We first show (3.88). For the term on the left-hand side, we have

$$AB : CB = \sum_{i,j=1}^n (AB)_{ij} (CB)_{ij} = \sum_{i,j=1}^n \left(\sum_{l=1}^n a_{il} b_{lj} \right) \left(\sum_{k=1}^n c_{ik} b_{kj} \right) = \sum_{i,j,k,l=1}^n a_{il} b_{lj} c_{ik} b_{kj},$$

while for the right-hand side, we get

$$ABB^T : C = \sum_{i,k=1}^n (ABB^T)_{ik} (C)_{ik} = \sum_{i,k=1}^n \sum_{j=1}^n (AB)_{ij} (B^T)_{jk} (C)_{ik} = \sum_{i,j,k,l=1}^n a_{il} b_{lj} b_{kj} c_{ik}.$$

We now prove (3.89). From the *definition* of the gradient operator applied to a vector field, we have

$$(B^T \nabla) \mathbf{w} = \begin{pmatrix} ((B^T \nabla) w_1)^T \\ \vdots \\ ((B^T \nabla) w_n)^T \end{pmatrix} = \begin{pmatrix} (B^T \nabla)_1 w_1 & \cdots & (B^T \nabla)_n w_1 \\ \vdots & \ddots & \vdots \\ (B^T \nabla)_1 w_n & \cdots & (B^T \nabla)_n w_n \end{pmatrix}$$

where w_i denotes the i^{th} component of \mathbf{w} , and thus

$$[(B^T \nabla) \mathbf{w}]_{ij} = (B^T \nabla)_j (\mathbf{w})_i.$$

Therefore, the coefficient of the i^{th} -row and j^{th} -column of the $n \times n$ matrix $(B^T \nabla) \mathbf{w}$ is given by

$$[(B^T \nabla) \mathbf{w}]_{ij} = \sum_{k=1}^n (B^T)_{jk} (\nabla)_k w_i = \sum_{k,l=1}^n b_{kj} \frac{\partial w_i}{\partial \xi_k} = \sum_{k=1}^n (\nabla \mathbf{w})_{ik} (B)_{kj} = (\nabla \mathbf{w} B)_{ij}.$$

□

We now show the relation (3.34) used in Section 3.2.2 to write the strong formulation of the problem on D . It can be proven by an integration by part back on the random domain D_ω or using the Piola identity $\nabla \cdot (J_{\mathbf{x}} A^T) = \mathbf{0}$ (see [101] for instance). Indeed, we have

$$\int_D q |J_{\mathbf{x}}| (A^T \nabla) \cdot \mathbf{v} d\xi = \int_{D_\omega} \tilde{q} \nabla_{\mathbf{x}} \cdot \tilde{\mathbf{v}} d\mathbf{x} = - \int_{D_\omega} \nabla_{\mathbf{x}} \tilde{q} \cdot \tilde{\mathbf{v}} d\mathbf{x} = - \int_D |J_{\mathbf{x}}| (A^T \nabla q) \cdot \mathbf{v} d\xi,$$

which yields (3.34) since $J_{\mathbf{x}}$ is either positive or negative, depending if the orientation is preserved or not by the mapping. Using the second alternative, since $\nabla \cdot (J_{\mathbf{x}} A \mathbf{v}) = (\nabla \cdot (J_{\mathbf{x}} A^T)) \cdot \mathbf{v} + (J_{\mathbf{x}} A^T \nabla) \cdot \mathbf{v}$ we have

$$\begin{aligned} \int_D q J_{\mathbf{x}} (A^T \nabla) \cdot \mathbf{v} d\xi &= \int_D q \nabla \cdot (J_{\mathbf{x}} A \mathbf{v}) d\xi - \int_D \underbrace{(\nabla \cdot (J_{\mathbf{x}} A^T))}_{=0} \cdot (q \mathbf{v}) d\xi \\ &= - \int_D J_{\mathbf{x}} (A^T \nabla q) \cdot \mathbf{v} d\xi. \end{aligned}$$

Be aware that in [101], the divergence operator applied to a tensor field is defined as the divergence applied to its transposed according to the definition used here. Recall that here we

defined $[\nabla \cdot (J_{\mathbf{x}} A^T)]_i = \sum_{j=1}^d \frac{\partial}{\partial \xi_j} (J_{\mathbf{x}} (A^T)_{ij}) = \sum_{j=1}^d \frac{\partial}{\partial \xi_j} (J_{\mathbf{x}} \frac{\partial (\xi_{\omega})_j}{\partial x_i} \circ \mathbf{x}_{\omega})$ for $i = 1, \dots, d$. Moreover, we mention that the Piola identity, which is easily obtained for smooth functions, say C^2 functions, is still valid (in a weak sense) for less regular functions such as H^1 functions (see for instance [12, 47]).

Finally, we derive the bound for the term $\Pi_3 = b(\mathbf{v}, p_{0,h}; \mathbf{y}_0) - b(\mathbf{v}, p_{0,h}; \mathbf{y})$ that appear in the proof of Proposition 3.2.14. Writing $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and $\mathbf{v} = (v_1, v_2)^T$, the two terms in component form read

$$b(\mathbf{v}, p_{0,h}; \mathbf{y}_0) = - \int_D p_{0,h} \nabla \cdot \mathbf{v} d\boldsymbol{\xi} = - \int_D p_{0,h} \left(\frac{\partial v_1}{\partial \xi_1} + \frac{\partial v_2}{\partial \xi_2} \right) d\boldsymbol{\xi}$$

and

$$\begin{aligned} b(\mathbf{v}, p_{0,h}; \mathbf{y}) &= - \int_D p_{0,h} J_{\mathbf{x}} (A^T \nabla) \cdot \mathbf{v} d\boldsymbol{\xi} \\ &= - \int_D p_{0,h} J_{\mathbf{x}} \left(A_{11} \frac{\partial v_1}{\partial \xi_1} + A_{21} \frac{\partial v_1}{\partial \xi_2} + A_{12} \frac{\partial v_2}{\partial \xi_1} + A_{22} \frac{\partial v_2}{\partial \xi_2} \right) d\boldsymbol{\xi}. \end{aligned}$$

Subtracting these two terms and using (both *continuous* and *discrete* version of) Cauchy-Schwarz's inequality we finally obtain

$$\begin{aligned} \Pi_3 &= \int_D (J_{\mathbf{x}} A_{11} - 1) p_{0,h} \frac{\partial v_1}{\partial \xi_1} d\boldsymbol{\xi} + \int_D J_{\mathbf{x}} A_{21} p_{0,h} \frac{\partial v_1}{\partial \xi_2} d\boldsymbol{\xi} + \int_D J_{\mathbf{x}} A_{12} p_{0,h} \frac{\partial v_2}{\partial \xi_1} d\boldsymbol{\xi} \\ &\quad + \int_D (J_{\mathbf{x}} A_{22} - 1) p_{0,h} \frac{\partial v_2}{\partial \xi_2} d\boldsymbol{\xi} \\ &\leq \| (J_{\mathbf{x}} A_{11} - 1) p_{0,h} \| \left\| \frac{\partial v_1}{\partial \xi_1} \right\| + \| J_{\mathbf{x}} A_{21} p_{0,h} \| \left\| \frac{\partial v_1}{\partial \xi_2} \right\| + \| J_{\mathbf{x}} A_{12} p_{0,h} \| \left\| \frac{\partial v_2}{\partial \xi_1} \right\| \\ &\quad + \| (J_{\mathbf{x}} A_{22} - 1) p_{0,h} \| \left\| \frac{\partial v_2}{\partial \xi_2} \right\| \\ &\leq \left(\| (J_{\mathbf{x}} A_{11} - 1) p_{0,h} \|^2 + \| J_{\mathbf{x}} A_{21} p_{0,h} \|^2 + \| J_{\mathbf{x}} A_{12} p_{0,h} \|^2 \right. \\ &\quad \left. + \| (J_{\mathbf{x}} A_{22} - 1) p_{0,h} \|^2 \right)^{\frac{1}{2}} \left(\sum_{i,j=1}^2 \left\| \frac{\partial v_i}{\partial \xi_j} \right\|^2 \right)^{\frac{1}{2}} \\ &= \| (J_{\mathbf{x}} A^T - I) p_{0,h} \| \| \nabla \mathbf{v} \|. \end{aligned}$$

We could also proceed as follows:

$$\begin{aligned}
\Pi_3 &= b(\mathbf{v}, p_{0,h}; \mathbf{y}_0) - b(\mathbf{v}, p_{0,h}; \mathbf{y}) = - \int_D p_{0,h} \nabla \cdot \mathbf{v} d\boldsymbol{\xi} + \int_D p_{0,h} J_{\mathbf{x}}(A^T \nabla) \cdot \mathbf{v} d\boldsymbol{\xi} \\
&= \sum_{i=1}^d \left[- \int_D p_{0,h} (I \nabla)_i v_i d\boldsymbol{\xi} + \int_D p_{0,h} J_{\mathbf{x}}(A^T \nabla)_i v_i d\boldsymbol{\xi} \right] \\
&= \sum_{i,j=1}^d \left[- \int_D p_{0,h} \delta_{ij} \frac{\partial v_i}{\partial \xi_j} d\boldsymbol{\xi} + \int_D p_{0,h} J_{\mathbf{x}}(A^T)_{ij} \frac{\partial v_i}{\partial \xi_j} d\boldsymbol{\xi} \right] \\
&= \sum_{i,j=1}^d \left[\int_D p_{0,h} (J_{\mathbf{x}}(A^T)_{ij} - \delta_{ij}) \frac{\partial v_i}{\partial \xi_j} d\boldsymbol{\xi} \right] = \int_D p_{0,h} (J_{\mathbf{x}} A^T - I) : \nabla \mathbf{v} d\boldsymbol{\xi} \\
&\leq \int_D \|p_{0,h} (J_{\mathbf{x}} A^T - I)\|_F \|\nabla \mathbf{v}\|_F \\
&\leq \|p_{0,h} (J_{\mathbf{x}} A^T - I)\|_{L^2(D)} \|\nabla \mathbf{v}\|_{L^2(D)},
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Froebenius norm.

4 Time-dependent heat equation with random Robin boundary conditions

Introduction

In this chapter, we perform an *a posteriori* error analysis for a time-dependent PDE with random input data, namely the heat equation with random Robin boundary conditions. The analysis is very similar to what has been done in the previous chapters, except that we have to take into account the error due to time discretization. For instance, for the approximation $u \approx u_{0,h\tau}$, where $u_{0,h\tau}$ is a space-time approximation of the deterministic part u_0 in the expansion of the solution u , the *a posteriori* error estimate is constituted of three parts, see Proposition 4.3.1. Each part controls a different source of error, namely the error due to space discretization, time discretization and uncertainty (truncation in the expansion of u).

4.1 Problem statement

Let $D \subset \mathbb{R}^d$, $d = 2, 3$, be an open bounded domain with Lipschitz continuous boundary ∂D and let (Ω, \mathcal{F}, P) be a complete probability space. We consider the following heat problem with random Robin boundary conditions:

find $u : (0, T) \times D \times \Omega \rightarrow \mathbb{R}$ such that a.s. in Ω the following equations hold

$$\left\{ \begin{array}{lll} \frac{\partial u(t, \mathbf{x}, \omega)}{\partial t} - \nabla \cdot (k(\mathbf{x}) \nabla u(t, \mathbf{x}, \omega)) & = & f(t, \mathbf{x}) \quad \mathbf{x} \in D, t \in (0, T) \\ u(t, \mathbf{x}, \omega) & = & 0 \quad \mathbf{x} \in \Gamma_D, t \in (0, T) \\ k(\mathbf{x}) \frac{\partial u(t, \mathbf{x}, \omega)}{\partial \mathbf{n}} + \alpha(\mathbf{x}, \omega) u(t, \mathbf{x}, \omega) & = & g(t, \mathbf{x}) \quad \mathbf{x} \in \Gamma_R, t \in (0, T) \\ u(t, \mathbf{x}, \omega) & = & \varphi(\mathbf{x}) \quad \mathbf{x} \in D, t = 0 \end{array} \right. \quad (4.1)$$

with Γ_D and Γ_R the Dirichlet and Robin boundary parts such that $\Gamma_D \cup \Gamma_R = \partial D$ and $\Gamma_D \cap \Gamma_R = \emptyset$ and \mathbf{n} is the outward unit normal vector on Γ_R . Notice that the subsequent analysis can be quite easily extended to the cases $f = f(t, \mathbf{x}, \omega)$, $g = g(t, \mathbf{x}, \omega)$, $\varphi = \varphi(\mathbf{x}, \omega)$ or $k = k(\mathbf{x}, \omega)$. From a physical point of view, the Robin boundary conditions for the heat problem are used to model the Newton's law of cooling [123], namely that the rate of change of temperature is

Chapter 4. Time-dependent heat equation with random Robin boundary conditions

proportional to the temperature difference between the solid surface Γ_R and its surroundings. Mathematically, this results in imposing a linear combination of Dirichlet (impose the temperature) and Neumann (impose the heat flux) boundary conditions. The parameter α is the heat transfer coefficient and depends on the material, the geometry, the environment, etc. In practise, this coefficient is often determined from experiments and is therefore subject to uncertainty. Another similar problem arises for instance in glaciology, when modelling the motion of glaciers, see for instance [80, 104] and references therein. The boundary conditions prescribed on the sliding basal part are indeed affected by uncertainty, for instance due to a lack of knowledge of the shape of the mountain or the difficulty to get measurements of the velocity of the ice on the base of the glacier.

We make the following assumptions on the input data

$$f \in L^2(0, T; L^2(D)), \quad g \in L^2(0, T; L^2(\Gamma_R)), \quad k \in L^\infty(D; \mathbb{R}^{d \times d}), \quad \varphi \in L^2(D), \quad \alpha(\cdot, \omega) \in L^\infty(\Gamma_R) \text{ a.s.}$$

and

$$\exists k_{\min} > 0 \quad \text{such that} \quad \forall \xi \in \mathbb{R}^d, \quad k(\mathbf{x}) \xi \cdot \xi \geq k_{\min} |\xi|^2 \quad \text{a.e. in } D. \quad (4.2)$$

Moreover, we assume that the random field α depends on a finite number of random variables $\{Y_j\}_{j=1}^L$, namely

$$\alpha(\mathbf{x}, \omega) = \alpha(\mathbf{x}, \mathbf{Y}(\omega)) = \alpha(\mathbf{x}, Y_1(\omega), \dots, Y_L(\omega)).$$

Let $\Gamma = \Gamma_1 \times \dots \times \Gamma_L$, where $\Gamma_j = Y_j(\Omega)$, and let $\rho : \Gamma \rightarrow \mathbb{R}^+$ be the joint density function of the random vector \mathbf{Y} . Let

$$V = H_{\Gamma_D}^1 = \{v \in H^1(D) : v = 0 \text{ on } \Gamma_D\}$$

endowed with the norm

$$\|v\|_V := \begin{cases} |v|_{H^1(D)} = \|\nabla v\|_{L^2(D)} & \text{if } \Gamma_D \neq \emptyset \\ \|v\|_{H^1(D)} = \sqrt{\|v\|_{L^2(D)}^2 + \|\nabla v\|_{L^2(D)}^2} & \text{if } \Gamma_D = \emptyset. \end{cases}$$

The parametric (pointwise in \mathbf{y} and t) weak formulation of problem (4.1) reads:

find $u \in L_\rho^2(\Gamma; L^2(0, T; V) \cap C^0([0, T]; L^2(D)))$ such that

$$\begin{cases} u(0, \mathbf{x}, \mathbf{y}) &= \varphi(\mathbf{x}) \quad \mathbf{x} \in D, \rho\text{-a.e. } \mathbf{y} \in \Gamma \\ \frac{d}{dt} \int_D uv + a(u, v; \mathbf{y}) &= F(v) \quad \forall v \in V, \text{ a.e. } t \in (0, T), \rho\text{-a.e. } \mathbf{y} \in \Gamma \end{cases} \quad (4.3)$$

with

$$a(u, v; \mathbf{y}) := \int_D k \nabla u \cdot \nabla v + \int_{\Gamma_R} \alpha(\mathbf{y}) uv \quad (4.4)$$

$$F(v) := \int_D f v + \int_{\Gamma_R} g v. \quad (4.5)$$

We can easily show that problem (4.3) is well-posed under the assumption

$$\alpha(\mathbf{x}, \mathbf{y}) \geq \alpha_{\min} > 0 \quad \text{a.e. } \mathbf{x} \in \Gamma_R, \rho\text{-a.e. } \mathbf{y} \in \Gamma. \quad (4.6)$$

Indeed, the condition (4.6) ensures the (uniform) coercivity of the bilinear form a defined in (4.4), that is there exists a constant $C_a > 0$ such that

$$C_a \|v\|_V^2 \leq a(v, v; \mathbf{y}) \quad \forall v \in V \text{ and } \rho\text{-a.e. } \mathbf{y} \in \Gamma. \quad (4.7)$$

It is obvious that (4.6) implies (4.7) for the case $\Gamma_D \neq \emptyset$, i.e. when V is endowed with the gradient norm, while it can be proved proceeding ab absurdo for the case $\Gamma_D = \emptyset$.

Remark 4.1.1. *In the case $\Gamma_D \neq \emptyset$, the assumption (4.6) can be relaxed since the bilinear form a is also coercive under the condition*

$$\|\alpha(\cdot, \mathbf{y})\|_{L^\infty(\Gamma_R)} < \frac{k_{\min}}{C_T^2(1 + C_F^2)} \quad \rho\text{-a.e. } \mathbf{y} \in \Gamma, \quad (4.8)$$

where C_F and C_T denote the Friedrich-Poincaré and trace constants in (2.5) and (2.10), respectively. In particular, it is not necessary that α remains positive. Indeed, thanks to (4.2) and using

$$-\int_{\Gamma_R} \alpha v^2 \leq \|\alpha\|_{L^\infty(\Gamma_R)} \|v\|_{L^2(\Gamma_R)}^2 \leq C_T^2 \|\alpha\|_{L^\infty(\Gamma_R)} \|v\|_{H^1(D)}^2 \leq C_T^2(1 + C_F^2) \|\alpha\|_{L^\infty(\Gamma_R)} \|\nabla v\|_{L^2(D)}^2$$

we have

$$a(v, v; \mathbf{y}) = \int_D k |\nabla v|^2 + \int_{\Gamma_R} \alpha v^2 \geq (k_{\min} - C_T^2(1 + C_F^2) \|\alpha\|_{L^\infty(\Gamma_R)}) \|\nabla v\|_{L^2(D)}^2$$

for any $v \in V$ and ρ -a.e. in Γ . The coercivity constant $C_a > 0$ is then given by

$$C_a = \begin{cases} k_{\min} & \text{if (4.6) holds} \\ k_{\min} - \|\alpha\|_{L^\infty(\Gamma_R)} C_T^2(1 + C_F^2) & \text{if (4.8) holds.} \end{cases}$$

Specific form of α

We assume that the random coefficient α , which appears in the Robin boundary condition, depends in an affine way on the random variables, namely that it can be written

$$\alpha(\mathbf{x}, \mathbf{Y}(\omega)) = \alpha_0(\mathbf{x}) + \varepsilon \sum_{j=1}^L \alpha_j(\mathbf{x}) Y_j(\omega),$$

where $\{Y_j\}_{j=1}^L$ are independent random variables with zero mean and unit variance.

Example 4.1.2. *Let $D = (0, 1)^2$ with boundary Γ_D and $\Gamma_R = \Gamma_{R_1} \cup \Gamma_{R_2} \cup \Gamma_{R_3}$ as shown in Figure 4.1.*

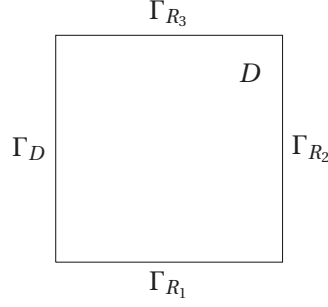


Figure 4.1: Geometry with label for each part of the boundary.

We take then $\alpha(\mathbf{x}, \mathbf{Y}(\omega)) = \alpha_0(\mathbf{x}) + \varepsilon \sum_{j=1}^3 \alpha_j(\mathbf{x}) Y_j(\omega)$ with

$$\alpha_0 = \begin{cases} \alpha_{0,1} & \text{if } \mathbf{x} \in \Gamma_{R_1} \\ \alpha_{0,2} & \text{if } \mathbf{x} \in \Gamma_{R_2} \\ \alpha_{0,3} & \text{if } \mathbf{x} \in \Gamma_{R_3} \end{cases}, \quad \alpha_j = \begin{cases} a_j & \text{if } \mathbf{x} \in \Gamma_{R_j} \\ 0 & \text{if } \mathbf{x} \in \Gamma_R \setminus \Gamma_{R_j} \end{cases}, \quad g = \begin{cases} g_1 & \text{if } \mathbf{x} \in \Gamma_{R_1} \\ g_2 & \text{if } \mathbf{x} \in \Gamma_{R_2} \\ g_3 & \text{if } \mathbf{x} \in \Gamma_{R_3} \end{cases}$$

and $\alpha_{0,j}, a_j \in L^\infty(\Gamma_{R_j})$, $j = 1, 2, 3$, such that (4.6) holds, i.e. $\alpha \geq \alpha_{\min} > 0$. For instance, in the case $\Gamma = [-1, 1]^3$, it is then required that $\varepsilon |a_j| < \alpha_{0,j}$ for $j = 1, 2, 3$.

Methodology

As in the previous chapters, we use a perturbation technique expanding the (random) solution u with respect to ε as:

$$u(t, \mathbf{x}, \mathbf{Y}(\omega)) = u_0(t, \mathbf{x}) + \varepsilon u_1(t, \mathbf{x}, \mathbf{Y}(\omega)) + \varepsilon^2 u_2(t, \mathbf{x}, \mathbf{Y}(\omega)) + \dots$$

The problem for the first term u_0 in the expansion simply reads:

find $u_0 : (0, T) \times D \rightarrow \mathbb{R}$ such that

$$\begin{cases} \frac{\partial u_0(t, \mathbf{x})}{\partial t} - \nabla \cdot (k(\mathbf{x}) \nabla u_0(t, \mathbf{x})) &= f(t, \mathbf{x}) & \mathbf{x} \in D, t \in (0, T) \\ u_0(t, \mathbf{x}) &= 0 & \mathbf{x} \in \Gamma_D, t \in (0, T) \\ k(\mathbf{x}) \frac{\partial u_0(t, \mathbf{x})}{\partial \mathbf{n}} + \alpha_0(\mathbf{x}) u_0(t, \mathbf{x}) &= g(t, \mathbf{x}) & \mathbf{x} \in \Gamma_R, t \in (0, T) \\ u_0(t, \mathbf{x}) &= \varphi(\mathbf{x}) & \mathbf{x} \in D, t = 0, \end{cases} \quad (4.9)$$

whose weak formulation can be written:

find $u_0 \in L^2(0, T; V) \cap C^0([0, T]; L^2(D))$ such that

$$\begin{cases} u_0(0, \mathbf{x}) &= \varphi(\mathbf{x}) & \mathbf{x} \in D \\ \frac{d}{dt} \int_D u_0 v + \int_D k \nabla u_0 \cdot \nabla v + \int_{\Gamma_R} \alpha_0 u_0 v &= \int_D f v + \int_{\Gamma_R} g v & \forall v \in V, \text{ a.e. } t \in (0, T). \end{cases} \quad (4.10)$$

Notice that problem (4.10) is nothing else than problem (4.3) with $\mathbf{y} = \mathbb{E}[\mathbf{Y}] = \mathbf{0}$. Writing $u_1(t, \mathbf{x}, \mathbf{Y}(\omega)) = \sum_{j=1}^L U_j(t, \mathbf{x}) Y_j(\omega)$, the second term in the expansion can be obtained by solving the L problems:

find $U_j : (0, T) \times D \rightarrow \mathbb{R}$ such that

$$\begin{cases} \frac{\partial U_j(t, \mathbf{x})}{\partial t} - \nabla \cdot (k(\mathbf{x}) \nabla U_j(t, \mathbf{x})) &= 0 & \mathbf{x} \in D, t \in (0, T) \\ U_j(t, \mathbf{x}) &= 0 & \mathbf{x} \in \Gamma_D, t \in (0, T) \\ k(\mathbf{x}) \frac{\partial U_j(t, \mathbf{x})}{\partial \mathbf{n}} + \alpha_0(\mathbf{x}) U_j(t, \mathbf{x}) &= -\alpha_j(\mathbf{x}) u_0(\mathbf{x}) & \mathbf{x} \in \Gamma_R, t \in (0, T) \\ U_j(t, \mathbf{x}) &= 0 & \mathbf{x} \in D, t = 0. \end{cases} \quad (4.11)$$

4.2 Numerical approximation

We assume from now on that $f \in C^0([0, T]; L^2(D))$, $g \in C^0([0, T]; L^2(\Gamma_R))$ and $\varphi \in C^0(\bar{D})$.

We approximate the solution u_0 of problem (4.10) using the (implicit) Backward Euler scheme in time and (\mathbb{P}_k) finite elements in space. For any $\tau > 0$, let $0 = t_0 < t_1 < \dots < t_M = T$ be a discretization of the time interval $[0, T]$ into M subintervals $I_n = [t_{n-1}, t_n]$ of length $\tau_n = t_n - t_{n-1} \leq \tau$, $n = 1, \dots, M$. Moreover, for any $h > 0$, let \mathcal{T}_h be a shape regular (in the sense of [49]) partition of D into d -simplices K of diameter $h_K \leq h$ and let

$$V_h = \{v \in C^0(\bar{D}) : v|_K \in \mathbb{P}_k, \forall K \in \mathcal{T}_h\} \cap V$$

be the subspace of V constituted of continuous, piecewise polynomial functions on \mathcal{T}_h .

Remark 4.2.1. Notice that a different mesh could be used for each time step, see e.g. [103], in which case we would write \mathcal{T}_h^n and V_h^n the mesh and FE space at time t_n . This functionality would be needed for instance when using adaptive algorithms, to allow the spatial meshes to vary in time. The introduction of an (interpolant) operator between two successive meshes is then required.

The fully discretized problem reads:

1. Initialization: $u_{0,h}^0 = r_h \varphi$
2. For $n = 1, \dots, M$: find $u_{0,h}^n \in V_h$ such that:

$$\int_D \frac{u_{0,h}^n - u_{0,h}^{n-1}}{\tau_n} v_h + \int_D k \nabla u_{0,h}^n \cdot \nabla v_h + \int_{\Gamma_R} \alpha_0 u_{0,h}^n v_h = \int_D f^n v_h + \int_{\Gamma_R} g^n v_h \quad \forall v_h \in V_h, \quad (4.12)$$

where $f^n = f(\cdot, t_n)$ and $g^n = g(\cdot, t_n)$. Finally, we define the *global* approximation $u_{0,hT}$, linear

on each subinterval I_n , by

$$u_{0,h\tau}(t, \mathbf{x}) := \frac{t - t_{n-1}}{\tau_n} u_{0,h}^n(\mathbf{x}) + \frac{t_n - t}{\tau_n} u_{0,h}^{n-1}(\mathbf{x}) \quad \text{for } t \in [t_{n-1}, t_n], n = 1, \dots, M. \quad (4.13)$$

4.3 *A posteriori* error analysis

For ease of notation, we introduce the element and edge residuals R and J defined on each element K and each edge e by, respectively,

$$R(u_{0,h\tau})|_K := f - \frac{\partial u_{0,h\tau}}{\partial t} + \nabla \cdot (k \nabla u_{0,h\tau}) \quad (4.14)$$

and

$$J(u_{0,h\tau})|_e := \begin{cases} \frac{1}{2} [k \nabla u_{0,h\tau} \cdot \mathbf{n}_e]_{\mathbf{n}_e} & \text{if } e \subset D \\ g - \alpha_0 u_{0,h\tau} - k \nabla u_{0,h\tau} \cdot \mathbf{n}_e & \text{if } e \subset \Gamma_R \\ 0 & \text{if } e \subset \Gamma_D. \end{cases} \quad (4.15)$$

We have denoted by $[\cdot]_{\mathbf{n}_e}$ the jump across an interior edge e , defined by

$$[\varphi]_{\mathbf{n}_e}(\mathbf{x}) := \lim_{t \rightarrow 0^+} (\varphi(\mathbf{x} + t\mathbf{n}_e) - \varphi(\mathbf{x} - t\mathbf{n}_e)).$$

Here, \mathbf{n}_e is the outer unit normal vector to the edge e if $e \subset \Gamma_R$ while for interior edges $e \subset D$, it is a unit normal vector to e of arbitrary (but fixed) direction. Notice that the choice of direction is irrelevant since quantities of the type $[\nabla \varphi \cdot \mathbf{n}_e]_{\mathbf{n}_e}$ is not affected by this choice, while $[\varphi]_{\mathbf{n}_e}$ is.

We have now introduced all the ingredients necessary to derive our *a posteriori* error estimate for the error $e := u - u_{0,h\tau}$ given in the following proposition.

Proposition 4.3.1. *Let u be the weak solution of problem (4.1) and let $u_{0,h\tau}$ be defined in (4.13). Then there exists a constant $C > 0$ depending only on the trace constant and the mesh aspect ratio such that*

$$\begin{aligned} \mathbb{E} \left[\|(u - u_{0,h\tau})(T)\|_{L^2(D)}^2 \right] &+ C_a \int_0^T \mathbb{E} [\|u - u_{0,h\tau}\|_V^2] dt \leq \\ &\|\varphi - r_h \varphi\|_{L^2(D)}^2 + \frac{C}{C_a} \sum_{n=1}^M \sum_{K \in \mathcal{T}_h} \left[\int_{t_{n-1}}^{t_n} ((\eta_K^n)^2 + (\gamma_K^n)^2 + (\theta_K^n)^2) dt \right], \end{aligned} \quad (4.16)$$

where C_a is the constant in (4.7) and

$$(\eta_K^n)^2 := h_K^2 \|R(u_{0,h\tau})\|_{L^2(K)}^2 + \sum_{e \in \partial K} h_e \|J(u_{0,h\tau})\|_{L^2(e)}^2 \quad (4.17)$$

$$(\gamma_K^n)^2 := \|k \nabla(u_{0,h\tau} - u_{0,h}^n)\|_{L^2(K)}^2 + \|f - f^n\|_{L^2(K)}^2 + \sum_{e \in \partial K \cap \Gamma_R} \|g - g^n - \alpha_0(u_{0,h\tau} - u_{0,h}^n)\|_{L^2(e)}^2 \quad (4.18)$$

$$(\theta_K^n)^2 := \varepsilon^2 \sum_{j=1}^L \|\alpha_j u_{0,h\tau}\|_{L^2(\partial K \cap \Gamma_R)}^2. \quad (4.19)$$

Proof. Let us write $e = u - u_{0,h\tau}$. In what follows, all equations are valid for a.e. t and a.s. in Ω without necessarily mentioning it. Moreover, C will denote a generic constant, whose value might change from one occurrence to another, that depends only on the interpolation constants in (1.26), (1.27) and (1.28), the trace constant in (2.10) and, if $\Gamma_D \neq \emptyset$, the Friedrich-Poincaré constant in (2.5). Thanks to equations (4.12) and (4.13), we have for each $v_h \in V_h$ and each $n \in \{1, \dots, M\}$

$$\begin{aligned} \int_D \frac{\partial u_{0,h\tau}}{\partial t} v_h + \int_D k \nabla u_{0,h\tau} \cdot \nabla v_h + \int_{\Gamma_R} \alpha_0 u_{0,h\tau} v_h &= \int_D f v_h + \int_{\Gamma_R} g v_h + \int_D k \nabla(u_{0,h\tau} - u_{0,h}^n) \cdot \nabla v_h \\ &+ \int_{\Gamma_R} \alpha_0(u_{0,h\tau} - u_{0,h}^n) v_h + \int_D (f^n - f) v_h \\ &+ \int_{\Gamma_R} (g^n - g) v_h \end{aligned} \quad (4.20)$$

using the fact that $\frac{\partial u_{0,h\tau}}{\partial t} = \frac{u_{0,h}^n - u_{0,h}^{n-1}}{\tau_n}$ on each time subinterval I_n , $n = 1, \dots, M$. Thanks to the coercivity of a , see (4.7), we have

$$C_a \|e\|_V^2 \leq \int_D k |\nabla e|^2 + \int_{\Gamma_R} \alpha e^2.$$

We now let n be any value in $\{1, \dots, M\}$. Then, for all $v \in V$ we have

$$\begin{aligned} \frac{d}{dt} \int_D e v + \int_D k \nabla e \cdot \nabla v + \int_{\Gamma_R} \alpha e v &= \int_D f v + \int_{\Gamma_R} g v - \int_D \frac{\partial u_{0,h\tau}}{\partial t} v - \int_D k \nabla u_{0,h\tau} \cdot \nabla v \\ &- \int_{\Gamma_R} \alpha_0 u_{0,h\tau} v - \int_{\Gamma_R} (\alpha - \alpha_0) u_{0,h\tau} v \\ &\stackrel{(4.20)}{=} \int_D f(v - I_h v) + \int_{\Gamma_R} g(v - I_h v) - \int_D \frac{\partial u_{0,h\tau}}{\partial t} (v - I_h v) \\ &- \int_D k \nabla u_{0,h\tau} \cdot \nabla (v - I_h v) - \int_{\Gamma_R} \alpha_0 u_{0,h\tau} (v - I_h v) \\ &- \int_D k \nabla(u_{0,h\tau} - u_{0,h}^n) \cdot \nabla I_h v - \int_{\Gamma_R} \alpha_0(u_{0,h\tau} - u_{0,h}^n) I_h v \\ &- \int_D (f^n - f) I_h v - \int_{\Gamma_R} (g^n - g) I_h v - \int_{\Gamma_R} (\alpha - \alpha_0) u_{0,h\tau} v, \end{aligned}$$

where I_h denotes the Clément interpolant of v . Taking then $v = e(t, \cdot, \mathbf{Y}(\omega))$ a.e. $t \in I_n$ and a.s.

in Ω in the last inequality, we get

$$\frac{1}{2} \frac{d}{dt} \|e\|_{L^2(D)}^2 + C_a \|e\|_V^2 \leq \text{I} + \text{II} + \text{III} \quad (4.21)$$

with

$$\begin{aligned} \text{I} &:= \sum_{K \in \mathcal{T}_h} \left\{ \int_K R(u_{0,h\tau})(e - I_h e) + \int_{\partial K} J(u_{0,h\tau})(e - I_h e) \right\} \\ \text{II} &:= - \int_D k \nabla(u_{0,h\tau} - u_{0,h}^n) \cdot \nabla I_h e - \int_{\Gamma_R} \alpha_0(u_{0,h\tau} - u_{0,h}^n) I_h e + \int_D (f - f^n) I_h e + \int_{\Gamma_R} (g - g^n) I_h e \\ \text{III} &:= - \int_{\Gamma_R} (\alpha - \alpha_0) u_{0,h\tau} e \end{aligned}$$

and R and J defined in (4.14) and (4.15), respectively. Notice that the terms I, II and III control the error due to space discretization, time discretization and truncation in the expansion of u , respectively. We now bound each of these terms separately.

bound for I: recalling the definition of η_K^n in (4.17), we obtain using a standard procedure the bound

$$\text{I} \leq C_1 \left(\sum_{K \in \mathcal{T}_h} (\eta_K^n)^2 \right)^{\frac{1}{2}} \|e\|_V, \quad (4.22)$$

where C_1 is a positive constant that depends only on the interpolation constants in (1.26) and (1.28).

bound for II: thanks to the triangle inequality, the interpolation error bounds (1.27) and (1.28) and the trace inequality (2.10), the following inequalities hold true

$$\begin{aligned} \|\nabla I_h e\|_{L^2(K)} &\leq \|\nabla e\|_{L^2(K)} + \|\nabla(e - I_h e)\|_{L^2(K)} \leq C|e|_{H^1(N(K))}, \\ \|I_h e\|_{L^2(K)} &\leq \|e\|_{L^2(K)} + \|e - I_h e\|_{L^2(K)} \leq C(1 + h_K) \|e\|_{H^1(N(K))} \leq C\|e\|_{H^1(N(K))}, \\ \|I_h e\|_{L^2(\Gamma_R)} &\leq C_T \|I_h e\|_{H^1(D)} = C_T \left[\sum_{K \in \mathcal{T}_h} \left(\|I_h e\|_{L^2(K)}^2 + \|\nabla I_h e\|_{L^2(K)}^2 \right) \right]^{\frac{1}{2}} \leq C\|e\|_V. \end{aligned}$$

Therefore, regrouping the integrals over the boundary Γ_R , we obtain the bound

$$\begin{aligned} \text{II} &\leq C_2 \left[\sum_{K \in \mathcal{T}_h} \left(\|k \nabla(u_{0,h\tau} - u_{0,h}^n)\|_{L^2(K)}^2 + \|f - f^n\|_{L^2(K)}^2 \right) \right. \\ &\quad \left. + \sum_{e \in \Gamma_R} \|g - g^n - \alpha_0(u_{0,h\tau} - u_{0,h}^n)\|_{L^2(e)}^2 \right]^{\frac{1}{2}} \|e\|_V \\ &= C_2 \left(\sum_{K \in \mathcal{T}_h} (\gamma_K^n)^2 \right)^{\frac{1}{2}} \|e\|_V \end{aligned} \quad (4.23)$$

with γ_K^n given in (4.18) and where C_2 is a positive constant that depends only on the constants in (1.27), (1.28) and (2.10). It additionally depends on the Friedrich-Poincaré constant in (2.5)

in the case $\Gamma_D \neq \emptyset$.

bound for III: for the last term, we easily get

$$\text{III} \leq \|(\alpha - \alpha_0)u_{0,h\tau}\|_{L^2(\Gamma_R)} \|e\|_{L^2(\Gamma_R)} \leq C_3 \|(\alpha - \alpha_0)u_{0,h\tau}\|_{L^2(\Gamma_R)} \|e\|_V, \quad (4.24)$$

where $C_3 = C_T$ if $\Gamma_D = \emptyset$ and $C_3 = C_T \sqrt{1 + C_F^2}$ otherwise, with C_T and C_F given in (2.10) and (2.5), respectively.

Using the bounds (4.22), (4.23) and (4.24) in (4.21) yields

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|e\|_{L^2(D)}^2 + C_a \|e\|_V^2 &\leq C \left[\sum_{K \in \mathcal{T}_h} ((\eta_K^n)^2 + (\gamma_K^n)^2) + \|(\alpha - \alpha_0)u_{0,h\tau}\|_{L^2(\Gamma_R)}^2 \right]^{\frac{1}{2}} \|e\|_V \\ &\leq \frac{C}{2C_a} \left[\sum_{K \in \mathcal{T}_h} ((\eta_K^n)^2 + (\gamma_K^n)^2) + \|(\alpha - \alpha_0)u_{0,h\tau}\|_{L^2(\Gamma_R)}^2 \right] + \frac{C_a}{2} \|e\|_V^2 \end{aligned}$$

and thus, splitting the integral of the last term of the right-hand side over the elements K we get

$$\frac{d}{dt} \|e\|_{L^2(D)}^2 + C_a \|e\|_V^2 \leq \frac{C}{C_a} \sum_{K \in \mathcal{T}_h} \left[(\eta_K^n)^2 + (\gamma_K^n)^2 + \|(\alpha - \alpha_0)u_{0,h\tau}\|_{L^2(\partial K \cap \Gamma_R)}^2 \right].$$

To conclude the proof, we integrate the last inequality over the time subinterval I_n , we sum then over n ranging from 1 to M and finally, we take the expected value on both sides recalling that $\mathbb{E}[Y_i Y_j] = \delta_{ij}$. \square

4.4 Numerical results

We give here two numerical examples to test the *a posteriori* error estimate derived in Section 4.3, see Proposition 4.3.1. We use \mathbb{P}_1 finite elements for the physical space approximation. In both examples, we set $k = I$ and we consider the case $\Gamma_D \neq \emptyset$. Therefore, the error $e = u - u_{0,h\tau}$ with $u_{0,h\tau}$ defined in (4.13) is measured with the norm

$$err := \mathbb{E} \left[\int_0^T \|\nabla e(t, \cdot, \cdot)\|_{L^2(D)}^2 dt \right]^{\frac{1}{2}} = \left(\int_{\Omega} \int_0^T \int_D |\nabla e(t, \mathbf{x}, \omega)|^2 d\mathbf{x} dt dP(\omega) \right)^{\frac{1}{2}}. \quad (4.25)$$

Similarly to [103], we define then the error estimator

$$est := \left(w_{\eta}^2 \eta^2 + w_{\gamma}^2 \gamma^2 + w_{\theta}^2 \theta^2 \right)^{\frac{1}{2}} \quad (4.26)$$

with weights w_{η} , w_{γ} and w_{θ} to be defined and

$$\eta^2 = \sum_{n=1}^M \sum_{K \in \mathcal{T}_h} \int_{t_{n-1}}^{t_n} (\eta_K^n)^2 dt, \quad \gamma^2 = \sum_{n=1}^M \sum_{K \in \mathcal{T}_h} \int_{t_{n-1}}^{t_n} (\gamma_K^n)^2 dt, \quad \theta^2 = \sum_{n=1}^M \sum_{K \in \mathcal{T}_h} \int_{t_{n-1}}^{t_n} (\theta_K^n)^2 dt$$

Chapter 4. Time-dependent heat equation with random Robin boundary conditions

where η_K^n , γ_K^n and θ_K^n are defined in (4.17), (4.18) and (4.19), respectively. Notice that η controls the space discretization, γ the time discretization and θ the truncation in the expansion of u with respect to ε .

Let $D = (0, 1)^2$ with boundary $\partial D = \Gamma_D \cup \Gamma_R$ as in Figure 4.1, let $T = 1$ and let Y_j , $j = 1, 2, 3$, be independent uniform random variables in $[-\sqrt{3}, \sqrt{3}]$. For the first case¹, we consider

$$u_0(t, x_1, x_2) = \sin\left(\frac{10\pi t}{2}\right) \sin\left(\frac{\pi x_1}{2}\right) \sin\left(\frac{\pi x_2}{2}\right) \quad \text{and} \quad \alpha(\mathbf{x}, \mathbf{Y}(\omega)) = \alpha_0(\mathbf{x}) + \varepsilon \sum_{j=1}^3 \alpha_j(\mathbf{x}) Y_j(\omega) \quad (4.27)$$

with $\alpha_0(\mathbf{x}) = 1$ and $\alpha_j(\mathbf{x}) = \chi_{\Gamma_{R_j}}(\mathbf{x})$, χ being the indicator function. We plug then u_0 and α_0 in (4.9) and compute the corresponding (deterministic) right-hand side f , boundary data g and initial condition φ . For the second case, using the same notation as in Example 4.1.2, we choose

$$f = \sin(2\pi x_1) t, \quad \varphi = 0, \quad g_1 = g_2 = g_3 = 0, \quad \alpha_0 = \begin{cases} 1 & \text{if } \mathbf{x} \in \Gamma_{R_1} \\ 2 & \text{if } \mathbf{x} \in \Gamma_{R_2} \\ 1.4 & \text{if } \mathbf{x} \in \Gamma_{R_3} \end{cases} \quad \text{and} \quad a_1 = 0.9, a_2 = 1.2, a_3 = 1. \quad (4.28)$$

We use a Delaunay triangulation with N equispaced vertices on each side of D for the space discretization and a uniform time step τ for the time discretization.

Deterministic case

We start considering the case $\varepsilon = 0$. For the first problem, the error is computed with respect to the exact solution u_0 in (4.27) while for the second case (4.28), we use a reference solution computed with $N_{ref} = 80$ and $\tau_{ref} = 2^{-7}$. Moreover, the constants w_η and w_γ in (4.26) have been tuned considering two test problems with exact solutions for (4.9), namely $u_0 = \sin(\pi x_1/2) \sin(\pi x_2/2)$ (mainly space error) and $u_0 = \sin(\pi t/2)$ (mainly time error), leading to $w_\eta = 1/5$ and $w_\gamma = 1/13$.

We give in Table 4.1 the results we get for the first case described in (4.27), considering various meshes with $N = 10, 20, 30, 40$ and various time steps $\tau = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$. The results obtained when computing the error with respect to a reference solution obtained with $N_{ref} = 80$ and $\tau_{ref} = 2^{-9}$ are also provided, for comparison with the random case below where such reference discretization parameters are used. The results for the case (4.28) with $N = 10, 20, 40$ and $\tau = 2^{-4}, 2^{-5}, 2^{-6}$ are provided in Table 4.2.

We see that for the first case (4.27), the error due to time discretization dominates the one due to the space approximation. The contrary holds for the second case (4.28) where the FE error is dominant. In both cases, the error estimator that contains the weights w_η and w_γ provides

¹This first example is similar to the case (3a) considered in [103]. The difference is that here we impose Robin (random) boundary conditions on a part of the boundary.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	est	e.i.	err ref	e.i. ref
10	2^{-4}	3.0665e-1	1.0463e-1	3.0351e-1	3.2104e-1	1.0469	2.9537e-1	1.0869
10	2^{-5}	1.6313e-1	7.5454e-2	1.5649e-1	1.7373e-1	1.0650	1.5339e-1	1.1326
10	2^{-6}	9.2626e-2	6.0711e-2	7.8871e-2	9.9531e-2	1.0745	8.415e-2	1.1827
10	2^{-7}	6.2591e-2	5.4026e-2	3.9521e-2	6.6938e-2	1.0695	5.6666e-2	1.1813
20	2^{-4}	3.0436e-1	5.2280e-2	3.0351e-1	3.0798e-1	1.0119	2.9298e-1	1.0512
20	2^{-5}	1.5801e-1	3.7632e-2	1.5649e-1	1.6095e-1	1.0186	1.4795e-1	1.0879
20	2^{-6}	8.2734e-2	3.0209e-2	7.8869e-2	8.4457e-2	1.0208	7.3232e-2	1.1533
20	2^{-7}	4.6377e-2	2.6833e-2	3.9520e-2	4.7768e-2	1.0300	3.8267e-2	1.2483
40	2^{-4}	3.0383e-1	2.5771e-2	3.0351e-1	3.0460e-1	1.0025	2.9242e-1	1.0416
40	2^{-5}	1.5679e-1	1.8571e-2	1.5649e-1	1.5759e-1	1.0051	1.4665e-1	1.0746
40	2^{-6}	8.0254e-2	1.4930e-2	7.8869e-2	8.0270e-2	1.0002	7.0421e-2	1.1399
40	2^{-7}	4.1706e-2	1.3278e-2	3.9520e-2	4.1691e-2	0.9996	3.2461e-2	1.2843
80	2^{-4}	3.0369e-1	1.2951e-2	3.0351e-1	3.0378e-1	1.0003	2.9222e-1	1.0396
80	2^{-5}	1.5648e-1	9.3271e-3	1.5649e-1	1.5677e-1	1.0019	1.4616e-1	1.0726
80	2^{-6}	7.9614e-2	7.4936e-3	7.8869e-2	7.9224e-2	0.9951	6.9327e-2	1.1428
80	2^{-7}	4.0438e-2	6.6615e-3	3.9520e-2	4.0077e-2	0.9911	2.9953e-2	1.3380

Table 4.1: Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0$.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	est	e.i.
10	2^{-4}	9.8673e-3	1.0500e-2	2.4507e-3	1.0782e-2	1.0928
10	2^{-5}	9.8634e-3	1.0491e-2	1.2254e-3	1.0562e-2	1.0708
10	2^{-6}	9.8624e-3	1.0488e-2	6.1275e-4	1.0506e-2	1.0653
20	2^{-4}	5.1306e-3	5.2838e-3	2.4512e-3	5.8247e-3	1.1353
20	2^{-5}	5.1233e-3	5.2790e-3	1.2257e-3	5.4194e-3	1.0578
20	2^{-6}	5.1217e-3	5.2777e-3	6.1287e-4	5.3131e-3	1.0374
40	2^{-4}	2.7265e-3	2.6335e-3	2.4513e-3	3.5978e-3	1.3196
40	2^{-5}	2.7129e-3	2.6311e-3	1.2257e-3	2.9026e-3	1.0699
40	2^{-6}	2.7099e-3	2.6304e-3	6.1290e-4	2.7009e-3	0.9967

Table 4.2: Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0$.

an efficient estimation of the error, the effectivity index being close to 1.

Random case

Let us now analyse the random case. The true error err in (4.25) is computed using the standard Monte-Carlo method with sample size $K = 100$. Moreover, for the first case (4.27), the reference solution is computed using $N_{ref} = 80$ and $\tau_{ref} = 2^{-9}$ while we use again $N_{ref} = 80$ and $\tau_{ref} = 2^{-7}$ for the second case (4.28). We choose $w_\theta = 1/3$ in (4.26), value obtained by considering either case with the same mesh for the approximation and the reference solution, for instance with the coarsest mesh parameters $N = 10$ and $\tau = 2^{-4}$. Notice that we get similar value for the case $N = N_{ref}$ and $\tau = \tau_{ref}$. We report in Tables 4.3, 4.4 and 4.5 the results we get for the first example (4.27) with $\varepsilon = 0.4$, $\varepsilon = 0.2$ and $\varepsilon = 0.1$, respectively.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	$w_\theta\theta$	est	e.i.
10	2^{-4}	3.0838e-1	1.0463e-1	3.0351e-1	8.1729e-2	3.3128e-1	1.0742
10	2^{-5}	1.8362e-1	7.5454e-2	1.5649e-1	8.6368e-2	1.9402e-1	1.0566
10	2^{-6}	1.3418e-1	6.0711e-2	7.8871e-2	8.9909e-2	1.3413e-1	0.9996
10	2^{-7}	1.1464e-1	5.4026e-2	3.9521e-2	9.2003e-2	1.1378e-1	0.9925
20	2^{-4}	3.1287e-1	5.2280e-2	3.0351e-1	8.1727e-2	3.1864e-1	1.0184
20	2^{-5}	1.8145e-1	3.7632e-2	1.5649e-1	8.6356e-2	1.8265e-1	1.0067
20	2^{-6}	1.2883e-1	3.0209e-2	7.8869e-2	8.9889e-2	1.2334e-1	0.9574
20	2^{-7}	1.0510e-1	2.6833e-2	3.9520e-2	9.1978e-2	1.0364e-1	0.9861
40	2^{-4}	3.1236e-1	2.5771e-2	3.0351e-1	8.1726e-2	3.1537e-1	1.0097
40	2^{-5}	1.7917e-1	1.8571e-2	1.5649e-1	8.6352e-2	1.7970e-1	1.0029
40	2^{-6}	1.2198e-1	1.4930e-2	7.8869e-2	8.9884e-2	1.2051e-1	0.9880
40	2^{-7}	1.0494e-1	1.3278e-2	3.9520e-2	9.1971e-2	1.0098e-1	0.9622
80	2^{-4}	3.0781e-1	1.2951e-2	3.0351e-1	8.1726e-2	3.1458e-1	1.0220
80	2^{-5}	1.8436e-1	9.3271e-3	1.5649e-1	8.6352e-2	1.7898e-1	0.9708
80	2^{-6}	1.1655e-1	7.4936e-3	7.8869e-2	8.9882e-2	1.1981e-1	1.0280
80	2^{-7}	1.0339e-1	6.6615e-3	3.9520e-2	9.1970e-2	1.0032e-1	0.9703

Table 4.3: Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.4$.

By analysing the results for this first case, we see that the (weighted) error estimator defined in (4.26) provides a good control of the error. Indeed, the effectivity index remains close to one for any value of N , τ and ε . Moreover, examining the behaviour of the error into more details, we see that each contribution $w_\eta\eta$, $w_\gamma\gamma$ and $w_\theta\theta$ efficiently controls the error. For instance, let us consider the case $N = 80$ for which the FE error is negligible. When $\varepsilon = 0.1$, the time estimator is dominant for any value of τ and the error is indeed divided by two when τ is halved. On the contrary, for $\varepsilon = 0.4$, the stochastic estimator is dominant for $\tau = 2^{-6}$ and $\tau = 2^{-7}$ and we can indeed observe it on the error: for the various time steps, the error decreases by a factor 1.67,

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	$w_\theta\theta$	est	e.i.
10	2^{-4}	3.0097e-1	1.0463e-1	3.0351e-1	4.0865e-2	3.2363e-1	1.0753
10	2^{-5}	1.6461e-1	7.5454e-2	1.5649e-1	4.3184e-2	1.7902e-1	1.0875
10	2^{-6}	9.8203e-2	6.0711e-2	7.8871e-2	4.4955e-2	1.0921e-1	1.1121
10	2^{-7}	7.3308e-2	5.4026e-2	3.9521e-2	4.6002e-2	8.1221e-2	1.1079
20	2^{-4}	2.9975e-1	5.2280e-2	3.0351e-1	4.0863e-2	3.1068e-1	1.0365
20	2^{-5}	1.5843e-1	3.7632e-2	1.5649e-1	4.3178e-2	1.6664e-1	1.0518
20	2^{-6}	8.6561e-2	3.0209e-2	7.8869e-2	4.4945e-2	9.5671e-2	1.1052
20	2^{-7}	6.2790e-2	2.6833e-2	3.9520e-2	4.5989e-2	6.6308e-2	1.0560
40	2^{-4}	2.9500e-1	2.5771e-2	3.0351e-1	4.0863e-2	3.0733e-1	1.0418
40	2^{-5}	1.5450e-1	1.8571e-2	1.5649e-1	4.3176e-2	1.6340e-1	1.0576
40	2^{-6}	8.8589e-2	1.4930e-2	7.8869e-2	4.4942e-2	9.1995e-2	1.0384
40	2^{-7}	5.9959e-2	1.3278e-2	3.9520e-2	4.5986e-2	6.2071e-2	1.0352
80	2^{-4}	2.9687e-1	1.2951e-2	3.0351e-1	4.0863e-2	3.0652e-1	1.0325
80	2^{-5}	1.5454e-1	9.3271e-3	1.5649e-1	4.3176e-2	1.6260e-1	1.0522
80	2^{-6}	8.6499e-2	7.4936e-3	7.8869e-2	4.4941e-2	9.1084e-2	1.0530
80	2^{-7}	5.5422e-2	6.6615e-3	3.9520e-2	4.5985e-2	6.0998e-2	1.1006

Table 4.4: Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.2$.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	$w_\theta\theta$	est	e.i.
10	2^{-4}	2.9570e-1	1.0463e-1	3.0351e-1	2.0432e-2	3.2169e-1	1.0879
10	2^{-5}	1.5506e-1	7.5454e-2	1.5649e-1	2.1592e-2	1.7507e-1	1.1291
10	2^{-6}	8.7940e-2	6.0711e-2	7.8871e-2	2.2477e-2	1.0204e-1	1.1603
10	2^{-7}	6.1673e-2	5.4026e-2	3.9521e-2	2.3001e-2	7.0780e-2	1.1477
20	2^{-4}	2.9410e-1	5.2280e-2	3.0351e-1	2.0432e-2	3.0865e-1	1.0495
20	2^{-5}	1.5116e-1	3.7632e-2	1.5649e-1	2.1589e-2	1.6239e-1	1.0743
20	2^{-6}	7.7490e-2	3.0209e-2	7.8869e-2	2.2472e-2	8.7395e-2	1.1278
20	2^{-7}	4.6148e-2	2.6833e-2	3.9520e-2	2.2995e-2	5.3015e-2	1.1488
40	2^{-4}	2.9313e-1	2.5771e-2	3.0351e-1	2.0431e-2	3.0528e-1	1.0415
40	2^{-5}	1.4913e-1	1.8571e-2	1.5649e-1	2.1588e-2	1.5906e-1	1.0666
40	2^{-6}	7.4518e-2	1.4930e-2	7.8869e-2	2.2471e-2	8.3356e-2	1.1186
40	2^{-7}	4.1056e-2	1.3278e-2	3.9520e-2	2.2993e-2	4.7611e-2	1.1596
80	2^{-4}	2.9480e-1	1.2951e-2	3.0351e-1	2.0431e-2	3.0447e-1	1.0328
80	2^{-5}	1.4910e-1	9.3271e-3	1.5649e-1	2.1588e-2	1.5825e-1	1.0613
80	2^{-6}	7.4526e-2	7.4936e-3	7.8869e-2	2.2471e-2	8.2349e-2	1.1050
80	2^{-7}	3.8214e-2	6.6615e-3	3.9520e-2	2.2992e-2	4.6204e-2	1.2091

Table 4.5: Error, estimators and effectivity index for the first case (4.27) with $\varepsilon = 0.1$.

Chapter 4. Time-dependent heat equation with random Robin boundary conditions

1.58 and 1.13. The case $\varepsilon = 0.2$ presents an intermediate stage with ratios 1.92, 1.79 and 1.56. Similar reasoning can be made for any other cases, namely that the saturation of the error is well explained by the domination of one of the error estimators. To conclude on this example, we finally mention that the slight increase of e.i. when τ decreases in Table 4.5 is due to the fact that the error is computed with respect to a reference solution. Indeed, if we consider the deterministic case $\varepsilon = 0$ with $N = 80$ and $\tau = 2^{-7}$, the error with respect to the reference solution is 0.0299529 yielding an effectivity index of about 1.36, see also Table 4.1.

The results for the second case with $\varepsilon = 0.5$ and $\varepsilon = 0.25$ are provided in Tables 4.6 and 4.7, respectively.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	$w_\theta\theta$	est	e.i.
10	2^{-4}	1.0989e-2	1.0500e-2	2.4507e-3	5.2493e-3	1.1992e-2	1.0913
10	2^{-5}	1.1020e-2	1.0491e-2	1.2254e-3	5.2393e-3	1.1790e-2	1.0699
10	2^{-6}	1.1140e-2	1.0488e-2	6.1275e-4	5.2356e-3	1.1738e-2	1.0537
20	2^{-4}	7.2634e-3	5.2838e-3	2.4512e-3	5.2568e-3	7.8461e-3	1.0802
20	2^{-5}	7.1864e-3	5.2790e-3	1.2257e-3	5.2469e-3	7.5432e-3	1.0496
20	2^{-6}	6.8839e-3	5.2777e-3	6.1287e-4	5.2431e-3	7.4646e-3	1.0844
40	2^{-4}	5.7040e-3	2.6335e-3	2.4513e-3	5.2591e-3	6.3720e-3	1.1171
40	2^{-5}	5.3548e-3	2.6311e-3	1.2257e-3	5.2491e-3	5.9982e-3	1.1202
40	2^{-6}	5.5691e-3	2.6304e-3	6.1290e-4	5.2454e-3	5.8999e-3	1.0594

Table 4.6: Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0.5$.

N	τ	err	$w_\eta\eta$	$w_\gamma\gamma$	$w_\theta\theta$	est	e.i.
10	2^{-4}	1.0142e-2	1.0500e-2	2.4507e-3	2.6247e-3	1.1097e-2	1.0942
10	2^{-5}	1.0155e-2	1.0491e-2	1.2254e-3	2.6197e-3	1.0882e-2	1.0717
10	2^{-6}	1.0167e-2	1.0488e-2	6.1275e-4	2.6178e-3	1.0827e-2	1.0650
20	2^{-4}	5.7001e-3	5.2838e-3	2.4512e-3	2.6284e-3	6.3903e-3	1.1211
20	2^{-5}	5.6392e-3	5.2790e-3	1.2257e-3	2.6234e-3	6.0210e-3	1.0677
20	2^{-6}	5.6824e-3	5.2777e-3	6.1287e-4	2.6216e-3	5.9247e-3	1.0426
40	2^{-4}	3.6562e-3	2.6335e-3	2.4513e-3	2.6296e-3	4.4563e-3	1.2188
40	2^{-5}	3.6174e-3	2.6311e-3	1.2257e-3	2.6246e-3	3.9132e-3	1.0818
40	2^{-6}	3.6337e-3	2.6304e-3	6.1290e-4	2.6227e-3	3.7647e-3	1.0361

Table 4.7: Error, estimators and effectivity index for the second case (4.28) with $\varepsilon = 0.25$.

Looking at the estimators for the case $\varepsilon = 0.5$, we see that the FE error is dominant when $N = 10$, the FE and stochastic errors balanced for $N = 20$ and the stochastic error is dominant when $N = 40$. We indeed observe this behaviour for the error. First, it remains more or less constant when changing the time step. Moreover, it decreases by a factor about 1.6 when

doubling N from 10 to 20, while the reduction of the error is only about 1.2 from $N = 20$ to $N = 40$. When diminishing the level of uncertainty, taking $\varepsilon = 0.25$, the stochastic error is lower and the error decreases by a factor 1.8 when increasing N from 10 to 20 and a factor 1.6 comparing the error for $N = 20$ and $N = 40$. Finally, the FE and stochastic error estimators are balanced when $N = 40$.

Conclusions

We have considered in this chapter the heat equation with random Robin boundary conditions. Under the assumption of small uncertainty, we have used a perturbation technique for the stochastic space approximation. In addition, the finite element method and the (implicit) backward Euler scheme have been used for the space and time discretizations, respectively. The *a posteriori* error estimator we have obtained for the approximation of the first term in the expansion consists in three distinct terms controlling each source of error. In the numerical experiments, we have introduced a *weighted* error estimator, with weights tuned numerically, and we have tested its efficiency on two different examples.

5 Error analysis for the stochastic collocation method

Introduction

In the previous chapters, we have used a perturbation approach for the stochastic space approximation. Such technique is no longer appropriate for problems with large variability. An alternative is to use the stochastic Galerkin or the stochastic collocation methods that present potentially much faster convergence rate than Monte-Carlo type methods and can handle large uncertainties. The advantage of the stochastic collocation method is that, as sampling methods, it requires only the solution of decoupled deterministic problems and thus allows the re-usability of deterministic solvers. However, it suffers from the so-called *curse of dimensionality* when tensor grids are used, namely the performance of the method deteriorates as the number of random variables increases. A remedy is then to exploit the possible anisotropy of the solution, in the sense that the different random variables might not have the same influence on the solution. Example of works in this direction are the anisotropic sparse grid method proposed in [96] or the quasi-optimal sparse grids method introduced in [20]. In the latter, the adaptive algorithm is based on *a priori* error estimates whose constants are numerically tuned during the process, yielding what the authors called an *a priori/a posteriori* strategy for which the proof of convergence has been obtained in [94]. An *a posteriori* sparse grid algorithm has been proposed in [95], where the adaptive process is driven by profit indicators obtained by solving additional PDEs. The method is applicable to a wide range of problems, including for instance the case of unbounded random variables or non-nested grids and can be combined with a Monte Carlo sampling, using a control variate technique, to handle rough random field [98]. However, the error indicators proposed so far are heuristic and do not provide a certified control of the error. The goal here is to derive a guaranteed upper bound of the error and use the stochastic error estimator to steer an adaptive process yielding an approximate solution with prescribed accuracy.

In this chapter, we thus present a residual-based *a posteriori* error estimate accounting both the stochastic collocation and the Finite Element error. We consider again the model problem of Chapter 1, namely a diffusion equation with a random diffusion coefficient that depends

in an affine manner of a finite number of random variables. We start by briefly recalling the SC method before presenting the error estimate. We give then possible adaptive algorithms, focusing on the stochastic space adaptation since the physical space adaptation can be done following a standard procedure. Finally, we give some preliminary numerical results to test the efficiency of a simple version of our sparse grid adaptive strategy.

5.1 Problem statement

Let $D \subset \mathbb{R}^d$ be an open bounded domain with Lipschitz continuous boundary ∂D and let (Ω, \mathcal{F}, P) be a complete probability space. We consider the diffusion problem:

find $u : D \times \Omega \rightarrow \mathbb{R}$ such that P -a.e. in Ω , in other words a.s. in Ω , the following equation holds

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) &= f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x}, \omega) &= 0 & \mathbf{x} \in \partial D \end{cases} \quad (5.1)$$

with deterministic forcing term $f \in L^2(D)$ and random field a on (Ω, \mathcal{F}, P) over $L^\infty(D)$. We assume that the random diffusion coefficient a is uniformly bounded from below and above and that it depends affinely on a finite number of random variables. More precisely, we assume that a satisfies the two following properties:

$$\exists 0 < a_{\min} \leq a_{\max} < \infty : P(\omega \in \Omega : a_{\min} \leq a(\mathbf{x}, \omega) \leq a_{\max} \forall \mathbf{x} \in \bar{D}) = 1. \quad (5.2)$$

and

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \sum_{n=1}^N a_n(\mathbf{x}) Y_n(\omega), \quad (5.3)$$

where $\{Y_n\}_{n=1}^N$ are independent random variables. Thanks to the Doob-Dynkin Lemma, the solution u depends on the same random variables as the diffusion coefficient a , i.e. we have $u(\mathbf{x}, \omega) = u(\mathbf{x}, Y_1(\omega), \dots, Y_N(\omega))$. Let us introduce $\Gamma = \Gamma_1 \times \dots \times \Gamma_N$ with $\Gamma_n = Y_n(\Omega)$ for $n = 1, \dots, N$. Moreover, let $\rho : \Gamma \rightarrow \mathbb{R}_+$ be the joint probability density function of the random vector $\mathbf{Y} = (Y_1, \dots, Y_N)$, which factorizes as $\rho(\mathbf{y}) = \prod_{n=1}^N \rho_n(y_n)$ for all $\mathbf{y} = (y_1, \dots, y_N) \in \Gamma$. We can then replace the probability space (Ω, \mathcal{F}, P) by $(\Gamma, B(\Gamma), \rho(\mathbf{y}) d\mathbf{y})$, where $B(\Gamma)$ denotes the Borel σ -algebra defined on Γ and $\rho(\mathbf{y}) d\mathbf{y}$ the probability measure of \mathbf{Y} . Finally, we define the Bochner space

$$L_\rho^2(\Gamma; H_0^1(D)) := \{v : \Gamma \rightarrow H_0^1(D) \mid v \text{ is strongly measurable and } \|v\|_{L_\rho^2(\Gamma; H_0^1(D))} < \infty\} \quad (5.4)$$

with

$$\|v\|_{L_\rho^2(\Gamma; H_0^1(D))} := \left(\int_\Gamma \|\nabla v(\mathbf{y})\|_{L^2(D)}^2 \rho(\mathbf{y}) d\mathbf{y} \right)^{\frac{1}{2}}.$$

The (parametric, pointwise) weak formulation of problem (5.1) reads:

find $u : \Gamma \rightarrow H_0^1(D)$ such that

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (5.5)$$

By a straightforward application of Lax-Milgram's lemma, assumption (5.2) ensures the well-posedness of problem (5.5), namely that there exists a unique solution $u \in L_\rho^2(\Gamma; H_0^1(D))$ which satisfies the *a priori* estimate

$$\|u\|_{L_\rho^2(\Gamma; H_0^1(D))} \leq \frac{C_P}{a_{\min}} \|f\|_{L^2(D)}.$$

Moreover, it has been shown (see for instance [7]) that the parametric solution u of problem (5.5) is analytic with respect to each parameter $y_n \in \Gamma_n$, $n = 1, \dots, N$.

5.2 Stochastic collocation finite element method

In this section, we briefly present the stochastic collocation finite element method (SC-FEM for short) for solving numerically PDEs with random input data, following closely [115] and focusing on the model problem (5.1). We also refer to [7, 124] for a complete discussion on this method. The idea is to proceed in two steps: first a semi-discretization of problem (5.5) using the FEM for the physical space approximation and then the application a collocation method for the stochastic space approximation using global polynomials in \mathbf{y} . We thus seek for an approximate solution in a space $\mathbb{P}(\Gamma) \otimes V_h$, with $\mathbb{P}(\Gamma) \subset L_\rho^2(\Gamma)$ a polynomial space on Γ and V_h a FE subspace of $V = H_0^1(D)$.

More precisely, for any $h > 0$, let \mathcal{T}_h be a regular triangulation of D with elements T of diameter $h_T \leq h$. We assume that there exists a constant $c > 0$ satisfying

$$\frac{h_T}{\rho_T} \leq c \quad \forall T \in \mathcal{T}_h, \forall h > 0 \quad (5.6)$$

where $\rho_T = \sup\{\text{diam}(B) : B \text{ is a ball contained in } T\}$. We consider $V_h \subset V$ a finite element space of dimension N_h constituted of continuous piecewise polynomials on \mathcal{T}_h . The semi-discretized problem is therefore given by:

find $u_h : \Gamma \rightarrow V_h$ such that

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla u_h(\mathbf{x}, \mathbf{y}) \cdot \nabla v_h(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v_h(\mathbf{x}) d\mathbf{x} \quad \forall v_h \in V_h, \rho\text{-a.e. in } \Gamma. \quad (5.7)$$

The problem (5.7) is then further discretized by considering a set $\{\mathbf{y}_1, \dots, \mathbf{y}_{N_c}\}$ of N_c collocation

points in Γ and building the global polynomial approximation

$$u_{h,N_c}(\mathbf{y}) = \sum_{k=1}^{N_c} u_h(\mathbf{y}_k) L_k(\mathbf{y}) \quad (5.8)$$

for appropriate multivariate (for instance Lagrange) polynomials L_k , where $u_h(\mathbf{y}_k)$ is the solution of problem (5.7) with $\mathbf{y} = \mathbf{y}_k$. Notice that if L_k satisfies $L_k(\mathbf{y}_l) = \delta_{kl}$, then the method presented above for the stochastic space approximation is a collocation method in the sense of [109], see [124].

A possible choice for the collocation points $\mathbf{y}_k \in \Gamma$ is to take the Cartesian product of the abscissas in each direction. However, using such tensor grid would rapidly become computationally unaffordable due to the *curse of dimensionality*: the number of nodes increases exponentially with N . To alleviate this drawback, the idea is to use a so-called *sparse grid*, first introduced by Smolyak in [113]. Let us define

$$\mathcal{U}_n^{m(i_n)} : C^0(\Gamma_n) \rightarrow \mathbb{P}_{m(i_n)-1}(\Gamma_n) \quad (5.9)$$

a sequence of univariate polynomial interpolant operators along each direction Γ_n for $n = 1, \dots, N$. Here, $m(i_n)$ denotes the number of collocation points used to build the interpolant of level i_n and $\mathbb{P}_q(\Gamma_n)$ is the space of polynomials in y_n of degree at most q . The function m should satisfy $m(0) = 0$, $m(1) = 1$ and $m(i) < m(i+1)$ for any $i \geq 1$. Moreover, let $I \subset \mathbb{N}_+^N$ be a multi-index set, where $\mathbb{N}_+ = \{1, 2, \dots\}$ denotes the positive integers. Setting $\mathcal{U}_n^0 = 0$ for $n = 1, \dots, N$, we define then the sparse grid interpolant S_I by

$$u_{h,I}(\mathbf{y}) = S_I[u_h](\mathbf{y}) = \sum_{\mathbf{i} \in I} \Delta^{\mathbf{m}(\mathbf{i})}(u_h)(\mathbf{y}) \quad (5.10)$$

where

$$\Delta^{\mathbf{m}(\mathbf{i})} = \bigotimes_{n=1}^N \Delta_n^{m(i_n)} = \bigotimes_{n=1}^N \left(\mathcal{U}_n^{m(i_n)} - \mathcal{U}_n^{m(i_n-1)} \right)$$

and $\mathbf{m}(\mathbf{i}) = (m(i_1), \dots, m(i_N))$. The operators $\Delta_n^{m(i_n)}$ and $\Delta^{\mathbf{m}(\mathbf{i})}$ are often referred to as *difference* (or *detail*) and *hierarchical surplus* operators, respectively. In what follows, we assume that

$$u_h(\mathbf{y}) = \sum_{\mathbf{i} \in \mathbb{N}_+^N} \Delta^{\mathbf{m}(\mathbf{i})}(u_h)(\mathbf{y}) \quad \rho\text{-a.e. in } \Gamma, \quad (5.11)$$

which holds if u is sufficiently smooth in \mathbf{y} and if the operators $\mathcal{U}_n^{m(i_n)}$ in (5.9) are such that $\bigotimes_{n=1}^N \mathcal{U}_n^{m(i_n)} u \rightarrow u$ in V as $\mathbf{i} \rightarrow \infty$. Finally, we mention that the operator S_I in (5.10) can be equivalently written as a linear combination of tensor grid interpolations, see for instance [122], as

$$S_I[u_h](\mathbf{y}) = \sum_{\mathbf{i} \in I} c_{\mathbf{i}} \bigotimes_{n=1}^N \mathcal{U}_n^{m(i_n)}(u_h)(\mathbf{y}), \quad c_{\mathbf{i}} = \sum_{\substack{\mathbf{j} \in \{0,1\}^N \\ (\mathbf{i}+\mathbf{j}) \in I}} (-1)^{|\mathbf{j}|} \quad (5.12)$$

in which many of the coefficients $c_{\mathbf{i}}$ are actually zero, namely whenever $\mathbf{i} + \mathbf{1} \in I$. We then call

sparse grid the set of N_c collocation points needed by (5.12) to compute $S_I[u_h]$. To summarize, the sparse grid interpolant S_I is characterized by the multi-index set I , the function m defining the number of collocation points on each level and the type of univariate nodes. One example, see for instance [18], is to consider

$$I(l) = \{\mathbf{i} \in \mathbb{N}_+^N : \sum_{n=1}^N (i_n - 1) \leq l\}$$

with

$$m(i) = \begin{cases} 0 & \text{if } i = 0 \\ 1 & \text{if } i = 1 \\ 2^{i-1} + 1 & \text{if } i > 1 \end{cases} \quad (5.13)$$

and Clenshaw-Curtis nodes, yielding nested grids. Here l denotes the level of the sparse grid. Remark that I must contain the multi-index $\mathbf{1}$, which allows to approximate constant functions.

In what follows, the only restriction on I will be that it is a downward closed set (a.k.a. lower set), i.e. it satisfies

$$\forall \mathbf{i} \in I, \quad \mathbf{i} - \mathbf{e}_j \in I \quad \forall j = 1, \dots, N \text{ such that } i_j > 1. \quad (5.14)$$

We give in Figure 5.1 an example of two multi-index sets satisfying or not this condition. The set on the left does not satisfy (5.14) because $(3, 2)$ is in the set while $(2, 2)$ is not. This condition is necessary to get good approximation properties, see for instance [66]. Moreover, our error estimate will only be valid in the case S_I is interpolatory, i.e. it satisfies $S_I[f](\mathbf{y}_k) = f(\mathbf{y}_k)$ for $k = 1, \dots, N_c$ where $\{\mathbf{y}_1, \dots, \mathbf{y}_{N_c}\}$ are the collocation points in the sparse grid underlying the multi-index set I and function m . Notice that such property requires the use of nested nodes.

Finally, we introduce the notion of margin M_I , reduced margin R_I and boundary ∂I of a multi-index set I , see Figure 5.1-right for an illustration, defined respectively by

$$\begin{aligned} M_I &= \{\mathbf{i} \in \mathbb{N}_+^N \setminus I : \mathbf{i} - \mathbf{e}_n \in I \text{ for some } n \in \{1, \dots, N\}\} \\ R_I &= \{\mathbf{i} \in M_I : \mathbf{i} - \mathbf{e}_n \in I \text{ for all } n = 1, \dots, N \text{ with } i_n > 1\} \\ \partial I &= \{\mathbf{i} \in I : \mathbf{i} + \mathbf{e}_n \notin I \text{ for some } 1 \leq n \leq N\}. \end{aligned}$$

Notice that for a downward closed multi-index set I and $\mathbf{j} \notin I$, then $I \cup \{\mathbf{j}\}$ is downward closed if and only if $\mathbf{j} \in R_I$.

From now on, unless otherwise clearly stated, we assume that I is downward closed and that the operator S_I is interpolatory.

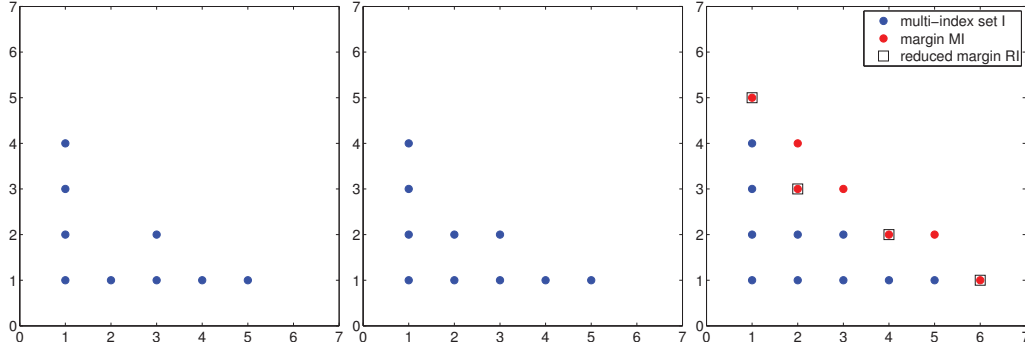


Figure 5.1: Non-downward closed set (left), downward closed set (middle) and multi-index set with its margin and reduced margin (right).

5.3 Residual-based *a posteriori* error estimate

We will now derive an *a posteriori* error estimate for the error $u - S_I[u_h]$ which consists of two parts controlling the finite element and stochastic collocation errors, respectively. We first give two results that we will use in the derivation of the error estimate.

Proposition 5.3.1. *Let S_I be the operator defined in (5.10). Then for any $f, g \in C^0(\Gamma)$ we have*

$$S_I[fg] = S_I[fS_I[g]].$$

Proof. Since S_I is assumed to be interpolatory, we have $S_I[g](\mathbf{y}_k) = g(\mathbf{y}_k)$ for all $k = 1, \dots, N_c$. By the definition of S_I , we get then for any $\mathbf{y} \in \Gamma$

$$\begin{aligned} S_I[fS_I[g]](\mathbf{y}) &= \sum_{k=1}^{N_c} (fS_I[g])(\mathbf{y}_k) L_k(\mathbf{y}) = \sum_{k=1}^{N_c} f(\mathbf{y}_k) S_I[g](\mathbf{y}_k) L_k(\mathbf{y}) \\ &= \sum_{k=1}^{N_c} f(\mathbf{y}_k) g(\mathbf{y}_k) L_k(\mathbf{y}) = S_I[fg](\mathbf{y}). \end{aligned}$$

□

For any multi-index set I , let us define the polynomial space \mathbb{P}_I by

$$\mathbb{P}_I = \sum_{\mathbf{i} \in I} \mathbb{P}_{m(i_1)-1} \otimes \dots \otimes \mathbb{P}_{m(i_N)-1}.$$

Notice that since we are using nested points, we have $N_c = \dim(\mathbb{P}_I)$ with N_c the number of collocation points in the sparse grid. Moreover, we have the following crucial approximation properties.

Proposition 5.3.2. *Let S_I be the operator defined in (5.10). Then*

1. $S_I[f] \in \mathbb{P}_I \quad \forall f \in C^0(\Gamma)$
2. S_I is exact on \mathbb{P}_I , i.e. $S_I[f] = f \quad \forall f \in \mathbb{P}_I$.

Proof. See [11]. □

Finally, we introduce the (generalized) jump of a function φ across an edge $e \in \mathcal{T}_h$ in the direction of \mathbf{n}_e as in Chapter 1 by

$$[\varphi]_{\mathbf{n}_e}(\mathbf{x}) := \begin{cases} \lim_{t \rightarrow 0^+} (\varphi(\mathbf{x} + t\mathbf{n}_e) - \varphi(\mathbf{x} - t\mathbf{n}_e)) & \text{if } e \not\subset \partial D \\ 0 & \text{if } e \subset \partial D. \end{cases}$$

We can now state our residual-based *a posteriori* error estimate.

Proposition 5.3.3. *Let u and u_h be the solutions of (5.5) and (5.7), respectively and let $S_I[u_h]$ be the sparse grid approximation of u_h computed using the multi-index set I . There exists a constant $C > 0$ depending only on the mesh aspect ratio such that for any $p \in [1, \infty]$ we have*

$$\|u - S_I[u_h]\|_{L^p_\rho(\Gamma; V)} \leq \frac{1}{a_{\min}} [C\eta_I + \zeta_I], \quad (5.15)$$

where

$$\eta_I = \sum_{k=1}^{N_c} \eta_{I,k} \|L_k\|_{L^p_\rho(\Gamma)}, \quad \eta_{I,k} := \left(\sum_{T \in \mathcal{T}_h} \eta_{I,k,T}^2 \right)^{\frac{1}{2}} \quad (5.16)$$

with

$$\eta_{I,k,T} := h_T^2 \|f + \nabla \cdot (a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k))\|_{L^2(D)}^2 + \sum_{e \subset \partial T} h_e \left\| \frac{1}{2} [a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k) \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2 \quad (5.17)$$

and

$$\zeta_I = \sum_{\mathbf{i} \in M_I} \zeta_{I,\mathbf{i}}, \quad \zeta_{I,\mathbf{i}} := \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h])\|_{L^p_\rho(\Gamma; L^2(D))}. \quad (5.18)$$

Proof. In what follows, all equations hold ρ -a.e. in Γ without specifically mentioning it. Moreover, the dependence of each function on variables will not necessarily be indicated, unless ambiguity arises. For any $v \in V$ we have

$$\begin{aligned} \int_D a \nabla(u - S_I[u_h]) \cdot \nabla v &= \int_D f v - \int_D a \nabla S_I[u_h] \cdot \nabla v \\ &= \underbrace{S_I \left[\int_D f v - \int_D a \nabla u_h \cdot \nabla v \right]}_{=: I} \\ &\quad + \underbrace{S_I \left[\int_D a \nabla u_h \cdot \nabla v \right] - \int_D a \nabla S_I[u_h] \cdot \nabla v}_{=: II}. \end{aligned} \quad (5.19)$$

For the second equality, we have used that f is assumed to be deterministic and thus $S_I[f] = f$ for any multi-index set I . We analyse the terms I and II separately. For the first term, thanks to the Galerkin orthogonality we have

$$\begin{aligned} \text{I} &= \sum_{k=1}^{N_c} \left[\int_D f v - \int_D a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k) \cdot \nabla v \right] L_k(\mathbf{y}) \\ &= \sum_{k=1}^{N_c} \left[\int_D f(v - v_h) - \int_D a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k) \cdot \nabla(v - v_h) \right] L_k(\mathbf{y}) \end{aligned} \quad (5.20)$$

for any $v_h \in V_h$. We take $v_h = I_h v$ the Cl  ment interpolant of v for which we have the following interpolation error bounds, see also (1.26) and (1.28)

$$\|v - I_h v\|_{L^2(T)} \leq Ch_T \|\nabla v\|_{L^2(N(T))} \quad \text{and} \quad \|v - I_h v\|_{L^2(e)} \leq Ch_e^{\frac{1}{2}} \|\nabla v\|_{L^2(N(T_e))} \quad (5.21)$$

for any element T and any edge e . Here, for an internal edge e , T_e is the union of the two elements touching e and $N(T)$ (resp. $N(T_e)$) denotes the patch of elements associated to T (resp. T_e). After splitting the integral in (5.20) over each element T and integrating by part, we obtain

$$\text{I} \leq C \sum_{k=1}^{N_c} |L_k(\mathbf{y})| \eta_{I,k} \|\nabla v\|_{L^2(D)} \quad (5.22)$$

with $\eta_{I,k}$ defined in (5.16). Notice that this term $\eta_{I,k}$ is *deterministic*, namely it does not depend on \mathbf{y} . It controls the FE error made when solving approximately the problem for the collocation point \mathbf{y}_k . We now bound the second term II. We first notice that, thanks to Proposition 5.3.1, we have $S_I[a \nabla u_h] = S_I[a \nabla S_I[u_h]]$ since S_I is assumed to be interpolatory. Therefore, using relation (5.11) we get

$$\begin{aligned} \text{II} &= \int_D (S_I[a \nabla S_I[u_h]] - a \nabla S_I[u_h]) \cdot \nabla v = - \int_D \sum_{\mathbf{i} \notin I} \Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h]) \cdot \nabla v \\ &= - \int_D \sum_{\mathbf{i} \in M_I} \Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h]) \cdot \nabla v \\ &\leq \left\| \sum_{\mathbf{i} \in M_I} \Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h]) \right\|_{L^2(D)} \|\nabla v\|_{L^2(D)}. \end{aligned} \quad (5.23)$$

We have used the fact that a depends in an affine way on the random variables, see (5.3), to restrict the summation over the multi-indices of the margin M_I of I . Indeed, by Proposition 5.3.2 we have

$$S_I[u_h] \in \mathbb{P}_I, \quad \text{where} \quad \mathbb{P}_I = \sum_{\mathbf{i} \in I} \mathbb{P}_{\mathbf{m}(\mathbf{i})-1} \quad \text{with} \quad \mathbb{P}_{\mathbf{m}(\mathbf{i})-1} = \mathbb{P}_{m(i_1)-1} \otimes \dots \otimes \mathbb{P}_{m(i_N)-1}$$

and by assumption

$$a \in \mathbb{P}_0 + \sum_{n=1}^N \mathbb{P}_{\mathbf{e}_n}, \quad \text{with} \quad \mathbb{P}_{\mathbf{e}_n} = \mathbb{P}_0 \otimes \dots \otimes \mathbb{P}_0 \otimes \underbrace{\mathbb{P}_1}_{n^{th} \text{ index}} \otimes \mathbb{P}_0 \dots \otimes \mathbb{P}_0.$$

Therefore, we have $a \nabla S_I[u_h] \in \sum_{n=1}^N \sum_{\mathbf{i} \in I} \mathbb{P}_{\mathbf{m}(\mathbf{i}) - \mathbf{1} + \mathbf{e}_n} \subset \mathbb{P}_{I \cup M_I}$ and thus

$$\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h]) = 0 \quad \forall \mathbf{i} \notin I \cup M_I \quad (5.24)$$

using again Proposition 5.3.2, namely that $S_{I \cup M_I}$ is exact on $\mathbb{P}_{I \cup M_I}$. Thanks to the uniform lower bound a_{\min} on a , taking then $v = u(\mathbf{y}) - S_I[u_h](\mathbf{y})$ in (5.19) and using the bounds (5.22) and (5.23) for the terms I and II, respectively, yields

$$\|\nabla(u(\mathbf{y}) - S_I[u_h](\mathbf{y}))\|_{L^2(D)} \leq \frac{1}{a_{\min}} \left(C \sum_{k=1}^{N_c} |L_k(\mathbf{y})| \eta_{I,k} + \left\| \sum_{\mathbf{i} \in M_I} \Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h])(\mathbf{y}) \right\|_{L^2(D)} \right). \quad (5.25)$$

To conclude the proof, it only remains to take the $L^p_\rho(\Gamma)$ norm on both sides of the last inequality and to use the triangle inequality for the norm $L^p_\rho(\Gamma; L^2(D))$ to take out the sum over the multi-indices $\mathbf{i} \in M_I$. \square

Notice that in this proof, we have strongly used the fact that S_I is interpolatory and that a depends in an affine way on the random variables. The latter allows us to restrict the summation over all the multi-indices outside I in the bound of II to the multi-indices belonging to the margin M_I . Moreover, it is worth mentioning that equation (5.25) yields a pointwise (in \mathbf{y}) error estimate.

Remark 5.3.4. The spatial error estimate η_I in (5.16) depends on $\|L_k(\mathbf{y})\|_{L^p_\rho(\Gamma)}$, $k = 1, \dots, N_c$, i.e. on the stability constant of the operator S_I . These quantities can be bounded using the Lebesgue constant for S_I , whose growth depends on the choice of the function m and the family of interpolation points used by $\mathcal{U}_n^{m(i)}$, $n = 1, \dots, N$. For instance, when using a doubling rule for m as in (5.13) and Clenshaw-Curtis nodes, the Lebesgue constant associated with the operator S_I can be bounded by $|I|^2$ [45]. As an alternative, we could bound the term I in (5.20) as follows

$$\begin{aligned} I &= \sum_{T \in \mathcal{T}_h} \left[\int_T \sum_{k=1}^{N_c} L_k(\mathbf{y}) (f + \nabla \cdot (a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k))) (v - v_h) + \right. \\ &\quad \left. \frac{1}{2} \sum_{e \in \partial T} \int_e \sum_{k=1}^{N_c} L_k(\mathbf{y}) [a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k) \cdot \mathbf{n}_e]_{\mathbf{n}_e} (v - v_h) \right] \\ &\leq C \left(\sum_{T \in \mathcal{T}_h} \eta_{I,T}^2 \right)^{\frac{1}{2}} \|\nabla v\|_{L^2(D)} \end{aligned}$$

with

$$\eta_{I,T}(\mathbf{y})^2 := h_T^2 \left\| \sum_{k=1}^{N_c} L_k(\mathbf{y}) (f + \nabla \cdot (a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k))) \right\|_{L^2(T)}^2 + \sum_{e \in \partial T} h_e \left\| \frac{1}{2} \sum_{k=1}^{N_c} [a(\mathbf{y}_k) \nabla u_h(\mathbf{y}_k) \cdot \mathbf{n}_e]_{\mathbf{n}_e} \right\|_{L^2(e)}^2. \quad (5.26)$$

Since $\left(\sum_{T \in \mathcal{T}_h} \eta_{I,T}^2\right)^{\frac{1}{2}} \leq \sum_{T \in \mathcal{T}_h} \eta_{I,T}$, we can then replace (5.15) by

$$\|u - S_I[u_h]\|_{L_\rho^p(\Gamma;V)} \leq \frac{1}{a_{\min}} \left[C \sum_{T \in \mathcal{T}_h} \|\eta_{I,T}\|_{L_\rho^p(\Gamma)} + \zeta_I \right]. \quad (5.27)$$

Mesh refinement, using the error estimate of Proposition 5.3.3 or the one proposed here would lead to different adaptive strategies. The estimator in (5.16) gives an estimation of the spatial error for each collocation point, that is further localized on each element $T \in \mathcal{T}_h$. Indeed, the estimator $\eta_{I,k,T}$ in (5.17) is an indicator of the FE error for element T and collocation point \mathbf{y}_k . Therefore, different spatial meshes could be considered for each collocation point. On the contrary, the estimator in (5.26) gives an estimation of the spatial error for each element $T \in \mathcal{T}_h$ and contains the contribution of all the collocation point. In this case, the same spatial mesh would then be used for all the collocation points.

5.3.1 An abstract reformulation of the problem

We consider the (pointwise in \mathbf{y}) abstract problem:

$$\text{find: } u(\mathbf{y}) \in V \quad \text{such that} \quad \mathcal{A}(u, v; \mathbf{y}) = \mathcal{F}(v; \mathbf{y}) \quad \forall v \in V, \rho\text{-a.e. in } \Gamma. \quad (5.28)$$

Using the finite element method for the physical space approximation, we get the following semi-discretized problem:

$$\text{find } u_h(\mathbf{y}) \in V_h \quad \text{such that} \quad \mathcal{A}(u_h, v_h; \mathbf{y}) = \mathcal{F}(v_h; \mathbf{y}) \quad \forall v_h \in V_h, \rho\text{-a.e. in } \Gamma. \quad (5.29)$$

Lax-Milgram's lemma ensures the well-posedness of problems (5.28) and (5.29) under the assumptions that the bilinear form \mathcal{A} is (uniformly in \mathbf{y}) continuous and coercive and that the linear functional \mathcal{F} is continuous. In particular, we assume that there exist two constants $\underline{\alpha}, \bar{\alpha} > 0$ such that ρ -a.e. in Γ

$$\underline{\alpha} \|v\|_V^2 \leq \mathcal{A}(v, v; \mathbf{y}) \quad \text{and} \quad |\mathcal{A}(u, v; \mathbf{y})| \leq \bar{\alpha} \|u\|_V \|v\|_V \quad \forall u, v \in V.$$

We can then derive the following *a posteriori* error estimate.

Proposition 5.3.5. *Let u and u_h be the solutions of (5.28) and (5.29), respectively and let $u_{h,I} = S_I[u_h]$ be the sparse grid approximation of u_h computed using the multi-index set I . If the series in (5.11) converge absolutely, then*

$$\|u - u_{h,I}\|_{L_\rho^p(\Gamma;V)} \leq \frac{1}{\underline{\alpha}} \left[\|R(u_h; \cdot)\|_{L_\rho^p(\Gamma;V')} + \bar{\alpha} \sum_{\mathbf{i} \notin I} \|\Delta^{\mathbf{m}(\mathbf{i})}[u_h]\|_{L_\rho^p(\Gamma;V)} \right]$$

where the residual R is defined for any $w, v \in V$ and any $\mathbf{y} \in \Gamma$ by

$$\langle R(w; \mathbf{y}), v \rangle := F(v; \mathbf{y}) - \mathcal{A}(w, v; \mathbf{y})$$

with $\langle \cdot, \cdot \rangle$ the duality pairing bracket between V and V' .

We highlight that in this proposition, S_I is not assumed to be interpolatory and the dependence on \mathbf{y} of the coefficients in \mathcal{A} and \mathcal{F} is not specified. In particular, we do not assume an affine dependency. However, the absolute convergence of the series in (5.11) is required and the estimator is not computable as is since it contains an infinite series. A computable estimator can however be obtained if we are able to provide estimation of the tail of the series.

Proof. For any $v \in V$ and ρ -a.e. in Γ we have

$$\begin{aligned} \mathcal{A}(u(\mathbf{y}) - u_{h,I}(\mathbf{y}), v; \mathbf{y}) &= F(v; \mathbf{y}) - \mathcal{A}(u_{h,I}(\mathbf{y}), v; \mathbf{y}) \\ &= \underbrace{F(v; \mathbf{y}) - A(u_h(\mathbf{y}), v; \mathbf{y})}_{=:I} + \underbrace{A(u_h(\mathbf{y}) - u_{h,I}(\mathbf{y}), v; \mathbf{y})}_{=:II}. \end{aligned}$$

Bounding each term separately, we easily obtain

$$I = \langle R(u_h(\mathbf{y}); \mathbf{y}), v \rangle \leq \|R(u_h(\mathbf{y}); \mathbf{y})\|_{V'} \|v\|_V$$

and

$$II \leq \bar{\alpha} \|u_h(\mathbf{y}) - u_{h,I}(\mathbf{y})\|_V \|v\|_V = \bar{\alpha} \|(id - S_I)u_h(\mathbf{y})\|_V \|v\|_V \leq \bar{\alpha} \sum_{\mathbf{i} \notin I} \|\Delta^{\mathbf{m}(\mathbf{i})}[u_h](\mathbf{y})\|_V \|v\|_V$$

where id denotes the identity operator. For the second term, we have used the relation (5.11), namely that the sparse grid approximation converges ρ -a.e. in Γ . Therefore, thanks to the coercivity of \mathcal{A} , taking $v = u(\mathbf{y}) - u_{h,I}(\mathbf{y})$ ρ -a.e. in Γ we get

$$\|u(\mathbf{y}) - u_{h,I}(\mathbf{y})\|_V \leq \frac{1}{\underline{\alpha}} \left[\|R(u_h(\mathbf{y}); \mathbf{y})\|_{V'} + \bar{\alpha} \sum_{\mathbf{i} \notin I} \|\Delta^{\mathbf{m}(\mathbf{i})}[u_h](\mathbf{y})\|_V \right].$$

The proof is complete by taking the $L^p_\rho(\Gamma)$ norm on both sides of this last inequality and using the triangle inequality. \square

Remark 5.3.6. In the special case where $\mathcal{A}(u, v; \mathbf{y}) = \int_D a(\mathbf{y}) \nabla u \cdot \nabla v$ and $F(v; \mathbf{y}) = \int_D f v$, which corresponds to problem (5.5), the dual norm of the residual $\|R(u_h(\mathbf{y}); \mathbf{y})\|_{V'}$ can be estimated by

$$\|R(u_h(\mathbf{y}); \mathbf{y})\|_{V'} \leq C\eta(\mathbf{y}) \quad \text{with} \quad \eta(\mathbf{y}) = \left(\sum_{T \in \mathcal{T}_h} \eta_T(\mathbf{y})^2 \right)^{\frac{1}{2}}$$

with

$$\eta_T(\mathbf{y})^2 := h_T^2 \|f + \nabla \cdot (a(\mathbf{y}) \nabla u_h(\mathbf{y}))\|_{L^2(T)}^2 + \sum_{e \in \partial T} h_e \left\| \frac{1}{2} [a(\mathbf{y}) \frac{\partial u_h(\mathbf{y})}{\partial \mathbf{n}_e}]_{\mathbf{n}_e} \right\|_{L^2(e)}^2.$$

5.4 Adaptive algorithms

The error estimator deduced from Proposition 5.3.3 can be used to adaptively refine the mesh and increase the multi-index set. Such an adaptive strategy aims at reaching a given accuracy of the (FE and stochastic) error with computational cost as low as possible. The theory for mesh adaptation, often referred to as adaptive finite element method (AFEM), is well developed and studied. In particular, the convergence of some adaptive procedures has been provided in many different cases. The first result in this direction is the work by Dörfler [57], where the convergence of an adaptive algorithm for the Poisson equation is given. Over the past decades, much effort has been put in proving convergence of adaptive algorithms (with optimal rate) for various types of problems, see for instance [25, 42, 93, 114]. In the context of parametric/random PDEs, we mention the work by [46] where the convergence of an adaptive algorithm is given when the solution is approximated via a Taylor series. In [58, 59], where the random PDEs are solved with the Stochastic Galerkin FEM, the convergence is proved when the adaptation is performed in both physical and stochastic spaces. In this case, the extension of the results obtained for the AFEM in [42] is straightforward and strongly uses the so-called Galerkin orthogonality property. Finally, for the stochastic collocation method, we mention the paper [20] in which a (quasi-optimal) sparse grid method based on a *a priori/a posteriori* strategy is proposed and whose convergence is analysed in [94]. Moreover, an *a posteriori* sparse grid algorithm is given in [95]. So far, at least to our knowledge, there is no proof of convergence for adaptive stochastic collocation methods.

Here, we will use the *a posteriori* error estimate given in Proposition 5.3.3 to drive an adaptive procedure. We start by considering only stochastic space adaptation since mesh adaptation can be performed in a classical way. The error estimator ζ_I can be used to adaptively enrich the multi-index set I in order to reach a prescribed accuracy while minimizing the computational cost. The proposed adaptive procedure is given in Algorithm 4.

Algorithm 4 Adaptive algorithm (stochastic space adaptation)

Require: $\theta \in (0, 1)$ and $Tol > 0$

Ensure: multi-index set I such that $\zeta_I \leq Tol$

- 1: $I = \{\mathbf{1}\}$, $u_I = S_I[u_h]$, $\zeta_I = \zeta_{I,1}$
 - 2: **while** $\zeta_I > Tol$ **do**
 - 3: $J = \text{new_index}(\theta, I, \zeta_I)$ *select a subset of M_I satisfying (5.30)*
 - 4: $I \leftarrow I \cup J$ *update the multi-index set*
 - 5: $u_I = S_I[u_h]$ *compute the new sparse grid approximation*
 - 6: $\zeta_I = \sum_{\mathbf{i} \in M_I} \zeta_{I,\mathbf{i}}$ *compute the error estimator (5.18)*
 - 7: **end while**
-

It remains to define the routine `new_index` of Step 3, namely to define how we select the multi-index set $J \subset M_I$ to be added to the current set I . Following a so-called Dörfler marking, we choose to select J according to

$$\text{find } J \subset M_I : \sum_{\mathbf{i} \in J} \zeta_{I,\mathbf{i}} \geq \theta \sum_{\mathbf{i} \in M_I} \zeta_{I,\mathbf{i}} \quad \text{and} \quad I \cup J \text{ downward closed.} \quad (5.30)$$

We can think of several strategies to select J satisfying (5.30), keeping in mind that the goal is to minimize the computational cost. Since the set should remain downward closed at each iteration of the adaptive algorithm, we associate to each multi-index \mathbf{i} a set $A_{\mathbf{i}}$ which consists of all multi-indices that must also be included in I if \mathbf{i} is added to I so that I remains downward closed. Notice that $A_{\mathbf{i}} = \{\mathbf{i}\}$ if \mathbf{i} belongs to the reduced margin. Moreover, we can define a notion of profit for each multi-index $\mathbf{i} \in M_I$ as follows

$$P_{\mathbf{i}} := \frac{\sum_{\mathbf{j} \in A_{\mathbf{i}}} \zeta_{I,\mathbf{j}}}{\sum_{\mathbf{j} \in A_{\mathbf{i}}} W_{\mathbf{j}}} \quad (5.31)$$

taking into account all elements of $A_{\mathbf{i}}$. Here, we have denoted by $W_{\mathbf{i}}$ the *work* contribution of the multi-index \mathbf{i} , which can be defined by [95]

$$W_{\mathbf{i}} = \prod_{n=1}^N (m(i_n) - m(i_n - 1)). \quad (5.32)$$

In the case of nested sets of point, as considered here, it corresponds to the number of new points in Γ introduced if \mathbf{i} is added to I . We could also choose to set $W_{\mathbf{i}} = 1$ if we want to drive the adaptation only based on the error indicators. With these definitions of $A_{\mathbf{i}}$ and $P_{\mathbf{i}}$, we can formulate a possible version of the routine `new_index`.

Algorithm 5 `new_index`

Require: θ , I and ζ_I

Ensure: multi-index set $J \subset M_I$ satisfying (5.30)

```

1:  $J = \emptyset$ ,  $\rho = 0$ 
2: while  $\rho < \theta \zeta_I$  do
3:    $\mathbf{i} = \operatorname{argmax}_{\mathbf{i} \in M_I \setminus J} P_{\mathbf{i}}$ 
4:    $J \leftarrow J \cup A_{\mathbf{i}}$ 
5:    $\rho = \sum_{\mathbf{j} \in J} \zeta_{I,\mathbf{j}}$ 
6: end while
```

Remark 5.4.1. Notice that the set J returned by Algorithm 5 might not be the optimal set satisfying (5.30). Indeed, a better set could be obtained by re-computing at each iteration the profit $P_{\mathbf{i}}$ in (5.31) of the multi-indices $\mathbf{i} \in M_I \setminus (R_I \cup J)$ for which $A_{\mathbf{i}}$ contains a multi-index added at the previous iteration. For such multi-index \mathbf{i} , the set $A_{\mathbf{i}}$ has changed and thus the profit.

To summarize, we have to choose the following parameters:

- the value of the *Dörfler parameter* $\theta \in (0, 1)$,
- the value of $p \in [1, \infty]$ for the $L_p^p(\Gamma)$ norm,
- the definition of the work $W_{\mathbf{i}}$ by (5.32) or $W_{\mathbf{i}} = 1$ in (5.31).

Implementation

We give here some details about the computation of the error estimators $\zeta_{I,\mathbf{i}}$ defined in (5.18), with particular attention to the case $\mathbf{i} \in M_I \setminus R_I$.

We consider the case $p = \infty$. Since the images of the random variables Γ_n , $n = 1, \dots, N$, are bounded and u is smooth with respect to \mathbf{y} , the essential supremum norm can be replaced by the maximum norm. Of course, not all the points of Γ can be explored and we choose to approximate the maximum norm searching for the maximum over a given set $\Theta \subset \Gamma$ of finite cardinality. The error is therefore computed using

$$\begin{aligned} \|u - S_I[u_h]\|_{L^\infty_\rho(\Gamma;V)} &= \max_{\mathbf{y} \in \Gamma} |\|\nabla(u - S_I[u_h])(\mathbf{y})\|_{L^2(D)} \rho(\mathbf{y})| \\ &\approx \max_{\mathbf{y} \in \Theta} |\|\nabla(u - S_I[u_h])(\mathbf{y})\|_{L^2(D)} \rho(\mathbf{y})| \end{aligned}$$

which requires the solution of $|\Theta|$ PDEs to get the value of $u(\mathbf{y})$ for each $\mathbf{y} \in \Theta$. Notice that since the FE error will not be accounted for in the numerical results, all the computation can be done on the same spatial mesh. The computation of the error estimators $\zeta_{I,\mathbf{i}}$ can be done as follows. Let G be any downward closed multi-index set that does not contains \mathbf{i} and such that $G \cup \{\mathbf{i}\}$ is also downward closed. The error estimator for \mathbf{i} is then approximately

$$\begin{aligned} \zeta_{I,\mathbf{i}} &= \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h])\|_{L^\infty_\rho(\Gamma;L^2(D))} \\ &= \|S_{G \cup \{\mathbf{i}\}}[a \nabla S_I[u_h]] - S_G[a \nabla S_I[u_h]]\|_{L^\infty_\rho(\Gamma;L^2(D))} \\ &\approx \max_{\mathbf{y} \in \Theta} | \|S_{G \cup \{\mathbf{i}\}}[a \nabla S_I[u_h]](\mathbf{y}) - S_G[a \nabla S_I[u_h]](\mathbf{y})\|_{L^2(D)} \rho(\mathbf{y})|. \end{aligned} \quad (5.33)$$

The key-point here is that no PDE need to be solved to compute (5.33). This formula can be straightforwardly applied for all the multi-indices $\mathbf{i} \in R_I$ with $G = I$, since $G \cup \{\mathbf{i}\}$ is downward closed, but a special care is required for the elements in $M_I \setminus R_I$. The idea is to iteratively increase the multi-index set I to cover the full margin in such a way that it remains downward closed throughout the process. We proceed *layer by layer*, starting by adding the elements of the reduced margin R_I , as described in the pseudo-code of Algorithm 6.

Algorithm 6 Computation of $\zeta_{I,\mathbf{i}}$ for all $\mathbf{i} \in M_I$

Require: $I, S_I[u_h], a$

Ensure: $\zeta_{I,\mathbf{i}} \forall \mathbf{i} \in M_I$

```

1:  $G = I$ 
2: while  $G \neq I \cup M_I$  do
3:    $R = R_G \cap M_I$ 
4:   for  $\mathbf{i} \in R$  do
5:     compute  $\zeta_{I,\mathbf{i}}$  using (5.33)
6:      $G \leftarrow G \cup \{\mathbf{i}\}$ 
7:   end for
8: end while
```

Notice that R at line 3 is a subset of the neighbours of the element of the previous previous layer. Moreover, the order of selection of the elements of R in the *for* loop is irrelevant.

Remark 5.4.2. For the case $p \in [1, \infty)$, the $L_p^p(\Gamma)$ norm can be computed (either exactly or approximately) using a Gauss quadrature formula built upon $S_{I(l)}$ with level l high enough. Notice that the larger p and the larger the polynomial degree of the integrand, the larger the level l should be.

Simplified algorithm

Algorithm 4, based on a Dörfler marking, is designed in the spirit of AFEM. The idea for introducing such algorithm was to prove its convergence as it is done for example in [59] for the Stochastic Galerkin method. We have made several attempts in this direction, for instance to prove that the error estimator satisfies a certain contraction property or to use different markings as it is done in [93] to control the decrease of the *data oscillation*. Unfortunately, we have not been successful so far, mainly due to the lack of the so-called *Galerkin orthogonality* valid for both the physical and the stochastic spaces when using the SG-FEM. The proof of convergence of the proposed adaptive algorithm is thus still an open question.

In the numerical results of Section 5.5, we consider a *simplified version* of Algorithm 4. First of all, we allow the selection of elements of the reduced margin R_I only and not of the full margin M_I . This simplifies the definition of the profits, since we do not need to introduce the sets A_i . Indeed, we recall that if I is downward closed, then so is $I \cup \{\mathbf{i}\}$ for any multi-index $\mathbf{i} \in R_I$. The second modification is that we add only one multi-index at a time. More precisely, the adaptive algorithm that is used for the numerical experiments of Section 5.5 reads as follows.

Algorithm 7 Simplified adaptive algorithm (stochastic space adaptation)

Require: $Tol > 0$

Ensure: multi-index set I such that $\zeta_I \leq Tol$

- 1: $I = \{\mathbf{1}\}$, $u_I = S_I[u_h]$, $\zeta_I = \zeta_{I,1}$
 - 2: **while** $\zeta_I > Tol$ **do**
 - 3: $\mathbf{i} = \operatorname{argmax}_{\mathbf{i} \in R_I} P_{\mathbf{i}}$ select the multi-index with highest profit
 - 4: $I \leftarrow I \cup \{\mathbf{i}\}$ update the multi-index set
 - 5: $u_I = S_I[u_h]$ compute the new sparse grid approximation
 - 6: $\zeta_I = \sum_{\mathbf{i} \in M_I} \zeta_{I,\mathbf{i}}$ compute the error estimator (5.18)
 - 7: **end while**
-

Remark 5.4.3. The adaptive process of Algorithm 7 is driven only by the profit of the elements of the reduced margin R_I of the current set I . To reduce the computational cost, we could therefore compute $\zeta_{I,\mathbf{i}}$ for $\mathbf{i} \in R_I$ only. However, the global error estimator ζ_I would no longer be available and we have to define another stopping criterion for the algorithm. For example, we can prescribe a tolerance Tol on the highest profit, i.e. stop the adaptive procedure when $\max_{\mathbf{i} \in R_I} P_{\mathbf{i}} < Tol$.

5.5 Numerical results

We consider here numerical examples to test Algorithm 7. In all what follows, we choose m as in (5.13) and we use Clenshaw-Curtis nodes. The FE error is not accounted here. Moreover, we consider the case $p = \infty$ and we thus consider the error and estimator defined by respectively

$$\|u_h - S_I[u_h]\|_{L_p^\infty(\Gamma; H_0^1(D))} \quad \text{and} \quad \sum_{\mathbf{i} \in M_I} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla S_I[u_h])\|_{L_p^\infty(\Gamma; L^2(D))}.$$

Be aware that the initialization step is not counted in the number of iterations given below. Therefore, the cardinality of the set I at the k^{th} iteration is equal to $k + 1$.

First example

For this first example, we consider an inclusion problem with $N = 2$ random variables, similar to the one consider in [11] for $N = 8$. The physical domain, depicted in Figure 5.1-left, is the unit square $D = (0, 1)^2$. We identify three subdomains F , C_1 and C_2 , with F a square centred in the domain with side length equal to 0.2 and C_1 and C_2 two circular inclusions of radius 0.13. We define the random diffusion coefficient by

$$a(\mathbf{x}, \mathbf{Y}(\omega)) = a_0(\mathbf{x}) + \sum_{n=1}^2 \gamma_n \chi_n(\mathbf{x}) Y_n(\omega) \quad \text{with } a_0 = 1 \text{ and } Y_n \sim \mathcal{U}[-0.99, 0.99] \quad (5.34)$$

and we set the forcing term to $f(\mathbf{x}) = 100\chi_F(\mathbf{x})$, where χ_F and χ_n , $n = 1, 2$, denote the indicator function of each subdomain. The parameters γ_1 and γ_2 are used to introduce anisotropy in the problem, assigning more importance to one or another direction y_1 or y_2 .

For the numerical experiments of this first example, we have used the following setting. The FE mesh consists of 4961 vertices and 9696 triangles with minimal and maximal diameter h_T of about $7.367\text{e-}3$ and $2.854\text{e-}2$, respectively. Since we would like to test the efficiency of our error estimator, namely to see if it is a good control of the (stochastic) error, we compute the estimator $\zeta_{I, \mathbf{i}}$ for each multi-index \mathbf{i} of the margin M_I . We can therefore base the stopping criterion on the global estimator ζ_I , see Remark 5.4.3. We set the tolerance to $Tol = 10^{-6}$. Finally, we compute the $L_p^\infty(\Gamma)$ norm approximately using for Θ a 20×20 Cartesian grid of equispaced points in each direction.

Isotropic case

We start with the isotropic case $\gamma_1 = \gamma_2 = 1$ in (5.34). The mean and the standard deviation of the solution is given in Figure 5.2, while the evolution of the set I during the adaptive process is presented in Figure 5.3. The multi-index with the green cross indicates the selected element at the current iteration of Algorithm 7, i.e. the one with the highest profit that belongs to the reduced margin of the previous set.

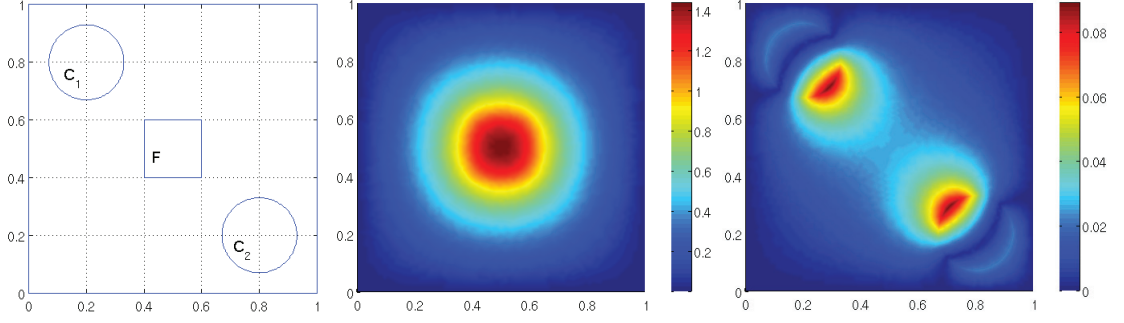


Figure 5.2: Geometry of the problem (left), expected value (middle) and standard deviation (right) of the solution for the case $\gamma_1 = \gamma_2 = 1$.

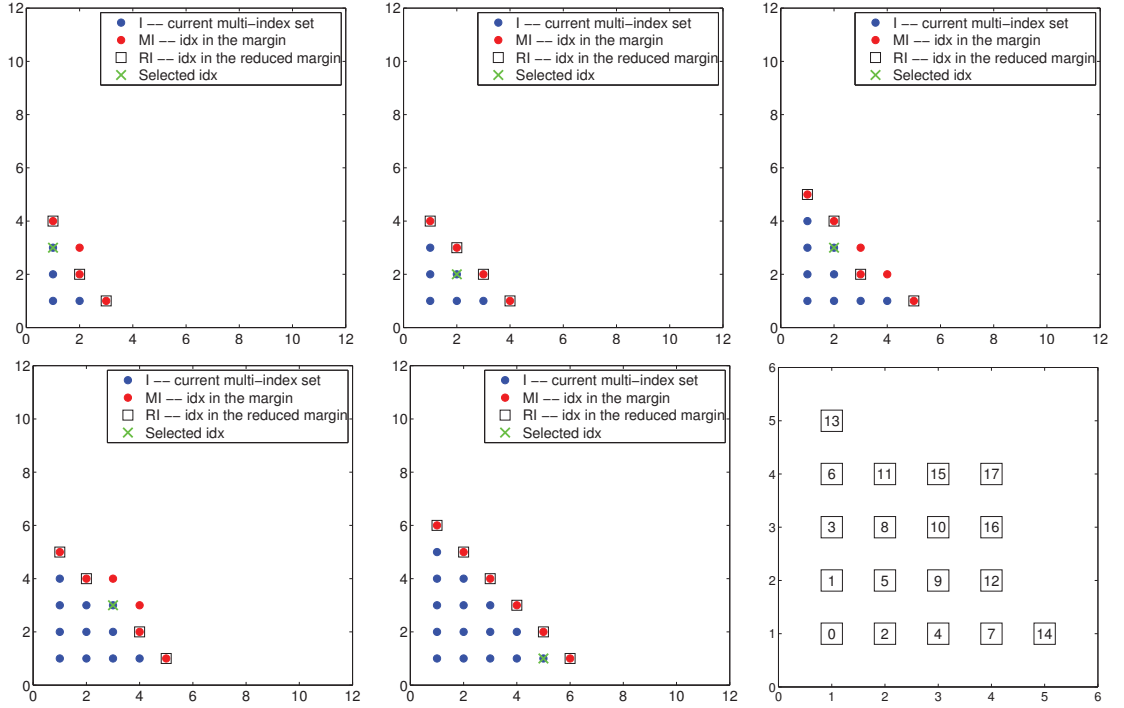


Figure 5.3: Evolution of I during the adaptive process for the case $\gamma_1 = \gamma_2 = 1$. From left to right and top to bottom: iterations 3,5,8,10,14 and order of selection of the multi-indices.

We can detect the isotropy of the problem by the *symmetrical* construction of the multi-index set. For instance, at iteration 11 the point (2,4) is added while (4,2) is selected at the next iteration. Moreover, we see that the estimator provides a good control of the error as shown in Figure 5.4, where the final multi-index set and the corresponding sparse grid are also given. It has been obtained after 17 iterations, yielding a sparse grids of 97 points and an error and an estimator of about $3.4649\text{e-}7$ and $8.1070\text{e-}7$, respectively. Finally, we mention that the highest profit of the elements of the reduced margin of this final stage is about $2.3702\text{e-}8$ and is achieved at (2,5).

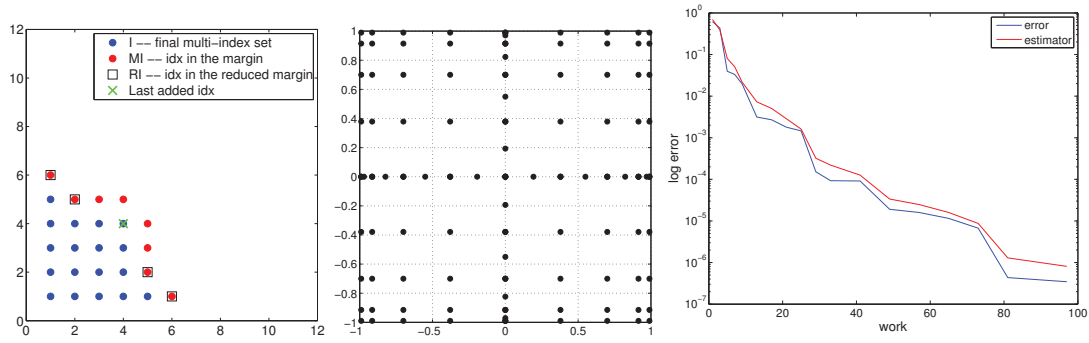


Figure 5.4: Final multi-index set I (left), final sparse grid (middle) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = \gamma_2 = 1$.

Anisotropic case

We now set different values for γ_1 and γ_2 in (5.34) to see if the adaptive algorithm is able to capture the anisotropy of the problem. We start with the trivial case $\gamma_1 = 1$ and $\gamma_2 = 0$, for which no point should be added in the second direction y_2 . This is indeed the result we get, as shown in Figure 5.5. At the end of the adaptive procedure, which requires 4 iterations, the sparse grid consists of 17 points and the error and estimator are about $1.4219\text{e-}10$ and $1.5276\text{e-}10$, respectively. The maximal profit among the elements of the reduced margin is $9.5472\text{e-}12$ and is attained at (6, 1).

Finally, we consider the case $\gamma_1 = 1$ and $\gamma_2 = 0.1$. We present in Figure 5.6 the set I at various steps of the adaptive construction. As expected, we can clearly identify a preferred direction, namely the horizontal direction which corresponds to y_1 .

The final situation, reached in 10 iterations, is given in Figure 5.7. In this case, there are 41 points in the sparse grid, the error and estimator are $6.8878\text{e-}8$ and $1.2500\text{e-}7$, respectively, and the maximal profit among the elements of the reduced margin is of $1.9995\text{e-}8$ at (3,3).

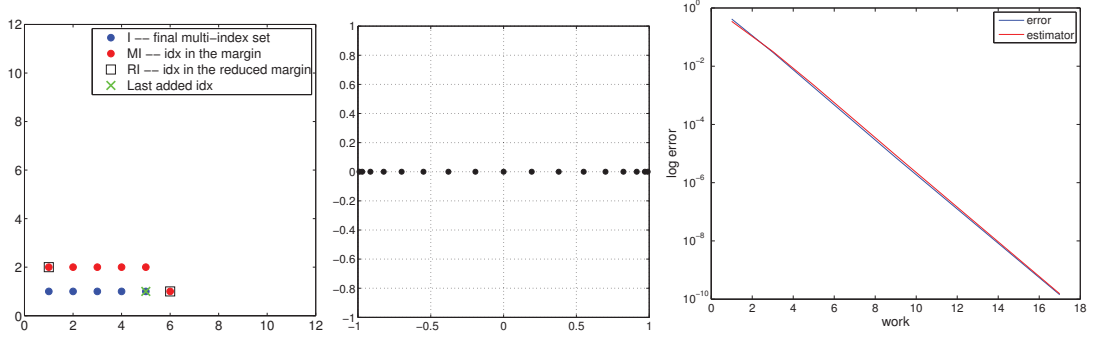


Figure 5.5: Final multi-index set I (left), final sparse grid (middle) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = 1$ and $\gamma_2 = 0$.

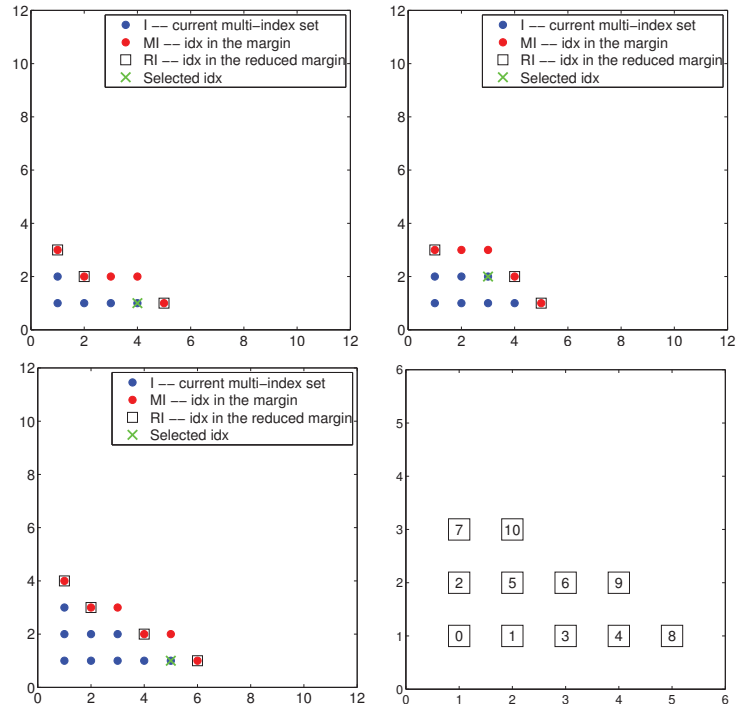


Figure 5.6: Evolution of the multi-index set I during the adaptive process for the case $\gamma_1 = 1$ and $\gamma_2 = 0.1$. From left to right and top to bottom: iterations 4,6,8 and order of selection of the multi-indices.

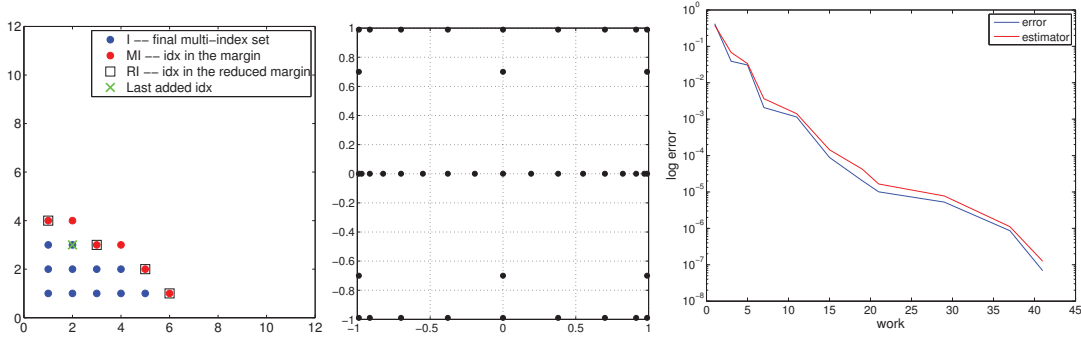


Figure 5.7: Final multi-index set I (left) and error and estimator with respect to the number of points in semi-logarithmic scale (right) for the case $\gamma_1 = 1$ and $\gamma_2 = 0.1$.

Anisotropic case $N = 8$

To conclude on this inclusion problem, we consider the case $N = 8$ as in [11] and we choose a similarly to (5.34) with $Y_n \sim \mathcal{U}[-0.99, 0.2]$ for $n = 1, \dots, 8$. The geometry is given in Figure 5.8-left, where the value of the coefficients γ_n , $n = 1, \dots, 8$, is also given. The FE mesh we are using contains 3805 vertices and 7416 triangles with minimal and maximal diameter h_T of about $1.0041e-2$ and $3.1153e-2$, respectively. Moreover, a set of 500 points randomly sampled from a multivariate uniform distribution is used for the approximation of the $L_p^\infty(\Gamma)$ norm. In Figure 5.8-right, we give the error and estimator for the 55 first iterations of Algorithm 7, after which the estimator is about $2.5102e-3$ and the sparse grid consists of 213 points in Γ . Moreover, the projection of the obtained multi-index set I over two directions, namely y_1 and y_4 , y_1 and y_5 and y_1 and y_7 , is presented in Figure 5.9.

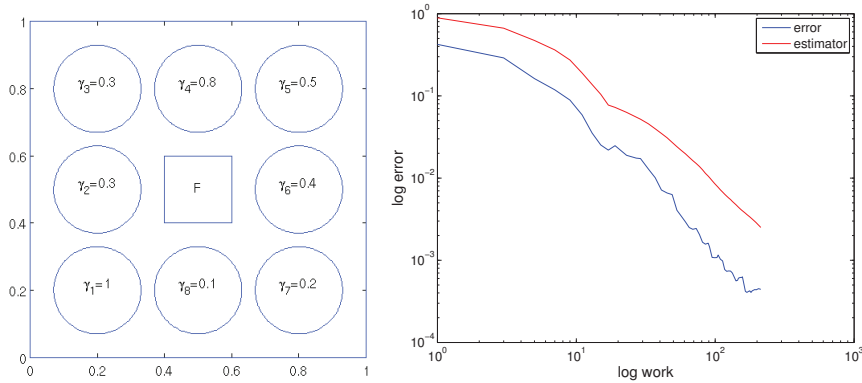


Figure 5.8: Geometry of the problem for $N = 8$ with indication of the coefficients γ_n , $n = 1, \dots, 8$ (left) and error and estimator with respect to the number of points in logarithmic scale for the 55 first iterations (right).

Even though the estimator still provides a reasonable control of the error, it is less efficient than for the case $N = 2$. We see several possible explanations for this behaviour and we give a

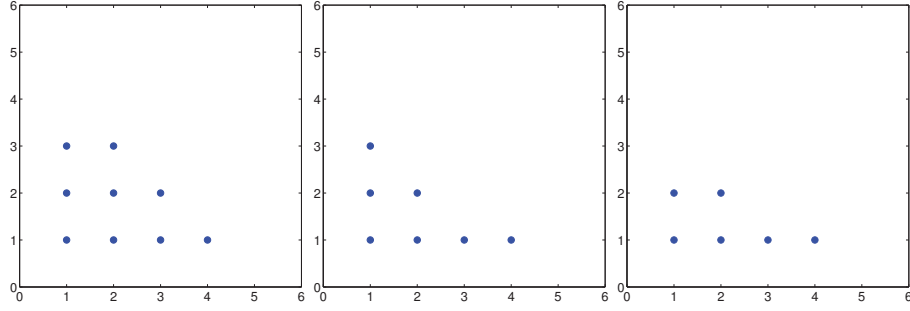


Figure 5.9: Projection of the multi-index set I obtained after 55 iterations on (y_1, y_4) (left), (y_1, y_5) (middle) and (y_1, y_7) (right).

non-exhaustive list below. First of all, we have not been able to prove that the error estimator provides a lower bound for the error. The difficulties arise, among other, from the lack of *Galerkin orthogonality* but also from the use of the triangle inequality to localize the estimator on each multi-index of the margin. Moreover, we are not taking into account the error due to the approximation of the $L_p^\infty(\Gamma)$ norm and further investigation should be made in this direction, namely trying to quantify this additional error and perform additional tests with other training sets Θ .

Second example

As a second numerical experiment, we take again the 2D example investigated in Section 1.7.2 of Chapter 1, namely we choose $f(\mathbf{x}) = 32(x_1(1 - x_1) + x_2(1 - x_2))$ and

$$a(\mathbf{x}, \mathbf{Y}(\omega)) = 1 + \sum_{n=1}^N \frac{\cos(2\pi n x_1) + \cos(2\pi n x_2)}{(\pi n)^2} Y_n(\omega) \quad \text{with} \quad Y_n \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$$

for $\mathbf{x} = (x_1, x_2) \in D$. We use a spatial mesh consisting of 2673 vertices and 5184 triangles with minimum and maximum diameter h_T of about 0.01 and 0.04, respectively. We set again the tolerance to $Tol = 10^{-6}$ in Algorithm 7 and the set Θ for the approximation of the $L_p^\infty(\Gamma)$ norm consists of 500 points in Γ randomly sampled from a multivariate uniform distribution. We consider the two cases $N = 3$ and $N = 5$.

The results for the case $N = 3$ are given in Figure 5.10. We plot the error and the estimator with respect to the work, i.e. number of points in the sparse grid. We also give the projection of the final multi-index set I over two directions, namely y_1 and y_3 . For this final state, obtained in 27 iterations, the error and the estimator are about $4.1493e-7$ and $9.1738e-7$, respectively, and the grid contains 141 points. Finally, we mention that the multi-index that has been in the last iteration to the final set I is $(4, 3, 1)$ and that the maximum profit among the elements of R_I is about $3.0159e-8$ and is reached at $(3, 2, 3)$.

The Figure 5.11 contains the results for the case $N = 5$. The final multi-index set I is projected

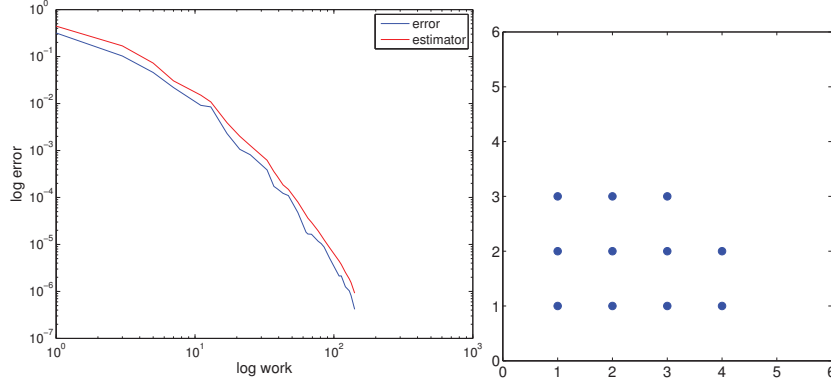


Figure 5.10: Error and estimator with respect to the number of points in logarithmic scale (left) and projection of the final multi-index set on (y_1, y_3) (right) for the case $N = 3$.

on y_1 and y_5 . The final grid has 469 points, for an error and estimator of about $2.2500e-6$ and $9.8095e-6$, respectively, and has been reached in 69 iterations. The last multi-index added to the set is $(4, 4, 1, 1, 1)$ and the maximum profit among the elements of the reduced margin of the final set is about $7.7365e-8$ at $(3, 2, 1, 2, 2)$.

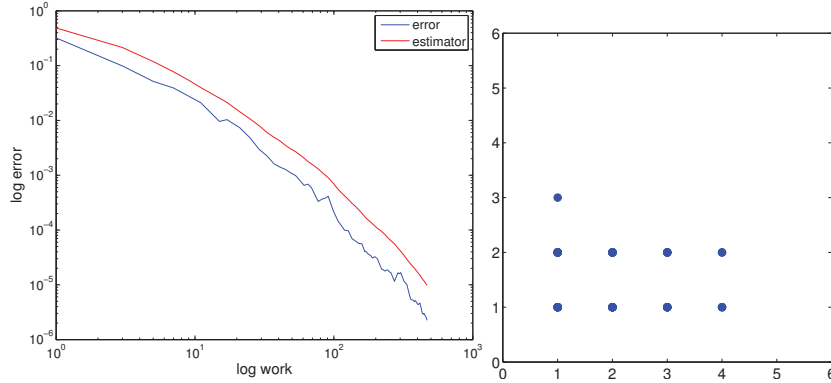


Figure 5.11: Error and estimator with respect to the number of points in logarithmic scale (left) and projection of the final multi-index set on (y_1, y_5) (right) for the case $N = 5$.

In both cases $N = 3$ and $N = 5$, the error estimator provides a good control of the error, the overestimation being slightly bigger for $N = 5$ than $N = 3$. Moreover, due to the decay of the a_n in n^{-2} , the random variables Y_n should have less and less influence as n increases. The adaptive algorithm is able to capture this feature, as seen for instance when projecting the obtained multi-index set over two different directions. From this experiment, together with the numerical results obtained for the inclusion problems, we see that the efficiency of the stochastic error estimator seems to be linked to the number of random variables. Further investigation should be made in this direction to determine whether this is indeed the case or if the reason is elsewhere, for instance the error due to the approximation of the $L_\rho^\infty(\Gamma)$ norm.

Conclusions

In this last chapter, we went out of the framework of small uncertainties considered in the previous chapters and in which a perturbation technique has been used for the stochastic space approximation. Here, we have considered the stochastic collocation method which is also appropriate for problems with a large amount of randomness but its use becomes challenging for problem in high dimensions. We have proposed a residual-based *a posteriori* error estimate that controls both the physical and stochastic space discretization. This estimate is valid under quite strong assumptions but that are often met in practice. First, we have assumed that the random diffusion coefficient depends in an affine way on a finite number of random variables, which is what we get for instance from a (truncated) Karhunen-Loève expansion of a random field. The second assumption is that the sparse grid operator is interpolatory, which requires the use of nested sequences of univariate nodes such as Clenshaw-Curtis or Leja nodes.

We have then proposed an adaptive sparse grid algorithm. The stochastic error estimator, which is localized on each element of the margin of the current multi-index set, is used to select the most profitable elements that should enter the set. The error estimator we have proposed presents the advantage to be computable without solving additional PDEs. However, it has the drawback that the profit needs to be recomputed at each iteration of the adaptive process since the residual depends on $S_I[u_h]$. We have made some numerical experiments to test the efficiency of a simple version of the adaptive algorithm. These are just preliminary yet promising results. They open the door to many improvements and prospects, including but not limited to

- quantify the error of approximation of the $L_\rho^\infty(\Gamma)$ norm using a finite number of (deterministic or random) points in Γ
- test different choices of family of points, such as Leja-sequence of points
- make a comparison with other methods, adaptive or not
- analyse the complexity of the proposed adaptive strategy
- prove the convergence of Algorithm 4
- take the FE error into account and do mesh refinement when the FE error dominates the stochastic one; take either the same mesh for all the collocation points or allow different refinements for the various points, see Remark 5.3.4
- consider the case of infinite number of random variables

5.A Miscellaneous results

We give here some preliminary results which might be useful to prove the convergence of Algorithm 4. In what follows, we will write I_k and I_{k+1} two successive multi-index sets produced

by the adaptive algorithm, that is $I_{k+1} = I_k \cup J_k$ with $J_k \subset M_{I_k}$ obtained using `new_index` and thus satisfying the Dörfler condition (5.30). Moreover, since we perform only stochastic space adaptation, we assume that there is no error due to FE approximation and the subscript h is no longer indicated in what follows. We write then $u_k = S_{I_k}[u]$ and $u_{k+1} = S_{I_{k+1}}[u]$ the sparse grid approximation corresponding to I_k and I_{k+1} , respectively.

First of all, since a depends affinely on the y_n , $n = 1, \dots, N$, we have that if $\mathbf{i} \in M_I$ then

$$\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla \Delta^{\mathbf{m}(\mathbf{j})}(u)) = 0 \quad \forall \mathbf{j} \in I \setminus \partial I. \quad (5.35)$$

Indeed, if $\mathbf{j} \in I \setminus \partial I$ then $\mathbf{j} + \mathbf{e}_n \in I$ for all $n = 1, \dots, N$.

For ease of notation, we will write $\|\cdot\|$ instead of $\|\cdot\|_{L^p_\rho(\Gamma; L^2(D))}$ in the sequel.

Proposition 5.A.1. (*Estimator reduction I*)

If $u_{k+1} = u_k$, then $\|\nabla(u - u_{k+1})\| = \|\nabla(u - u_k)\|$ but

$$\zeta_{I_{k+1}} < \zeta_{I_k}.$$

Proof. First of all, we split the margin of I_{k+1} into two disjoint parts as

$$M_{I_{k+1}} = (M_{I_k} \setminus J_k) \cup (M_{J_k} \setminus M_{I_k}).$$

Using the assumption $u_{k+1} = u_k$ we get then

$$\begin{aligned} \zeta_{I_{k+1}} &= \sum_{\mathbf{i} \in M_{I_{k+1}}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_{k+1})\| \\ &= \sum_{\mathbf{i} \in M_{I_{k+1}}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| \\ &= \sum_{\mathbf{i} \in M_{I_k} \setminus J_k} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| + \sum_{\mathbf{i} \in M_{J_k} \setminus M_{I_k}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| \\ &= \sum_{\mathbf{i} \in M_{I_k} \setminus J_k} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\|. \end{aligned}$$

For the last equality, we have used that $\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k) = 0$ for all $\mathbf{i} \in (M_{J_k} \setminus M_{I_k})$ thanks to (5.24). Indeed, if $\mathbf{i} \in (M_{J_k} \setminus M_{I_k})$ then $\mathbf{i} \notin (I_k \cup M_{I_k})$. Finally, we use the property of J_k in (5.30) to obtain

$$\begin{aligned} \zeta_{I_{k+1}} &= \sum_{\mathbf{i} \in M_{I_k}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| - \sum_{\mathbf{i} \in J_k} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| \\ &\leq (1 - \theta) \sum_{\mathbf{i} \in M_{I_k}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a \nabla u_k)\| \\ &= \kappa \zeta_{I_k}, \end{aligned}$$

with $\kappa = (1 - \theta) < 1$ for any choice of the Dörfler parameter $\theta \in (0, 1)$. □

One way to prove the convergence of Algorithm 4 is to prove a contraction property, for instance on the error, the estimator or some other quantity. The difficulty is therefore to first define the quantity on which we would like to prove a contraction property. We have tried to do it on the estimator, but, unfortunately, we have not been able yet to find a conclusion. So far, we have obtained the following relation

$$\begin{aligned}\zeta_{I_{k+1}} &\leq \sum_{\mathbf{i} \in M_{I_k} \setminus J_k} \|\Delta^{\mathbf{m}(\mathbf{i})}(a\nabla u_k)\| + \sum_{\mathbf{i} \in M_{I_{k+1}}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a\nabla(u_{k+1} - u_k))\| \\ &\leq (1 - \theta)\zeta_{I_k} + \sum_{\mathbf{i} \in M_{I_{k+1}}} \|\Delta^{\mathbf{m}(\mathbf{i})}(a\nabla(u_{k+1} - u_k))\|.\end{aligned}$$

Conclusions and perspectives

In this thesis, error analysis for PDEs with random input data has been performed on various problems with a focus on *a posteriori* error estimation.

The starting point was the well-studied elliptic diffusion model problem with random diffusion coefficient and affine dependence on the random variables. Assuming small amount of randomness in the model, characterized with the parameter ε , a perturbation technique was used expanding the exact random solution of this problem in powers of ε . Error estimation for the error between the exact solution and the finite element approximation of the truncated expansion has been established in great details, considering different measures of the error. Computing for instance only the first term in the expansion, which is deterministic, the *a posteriori* error estimate provides information about both sources of error, namely the physical space discretization and the uncertainty, and can be used to balance these two errors. Moreover, such error estimates are the basis for adaptive strategies designed to find an approximation of prescribed accuracy with computational cost as low as possible. Having *a posteriori* error estimate for the approximation of any order allows us to adaptively choose between mesh refinement and increase of the order of the expansion. The theoretical results have been validated and illustrated through many numerical experiments in one and two physical space dimensions. We are looking forward to perform numerical experiments on adaptive schemes of higher-order in ε . A proof of the lower bound for the *explicit* stochastic error estimator of the first order approximation, required to prove its efficiency, is still missing at the moment.

Next, steady-state nonlinear problems in random domains have been investigated. For such problems, the so-called *domain mapping method* has been used to transform the PDEs in random domains into PDEs on a fixed reference domain with random coefficients. All the analysis can then be made on this fixed reference domain and, from a numerical point of view, this method prevents the need of remeshing. Application to the one-dimensional viscous Burger's equation and the incompressible Navier-Stokes equations has been proposed. The well-posedness has been shown, under suitable conditions on the mapping and the input data, using a fixed-point theorem for existence and a variational argument for uniqueness. *A posteriori* error estimation has been proposed for a specific but rather general form of the mapping, again under the assumption of small perturbation. For the Navier-Stokes problem, two different estimates have been developed, each of them presenting advantages and draw-

Conclusions and perspectives

backs. Numerical results have been given for both problems. Possible extensions include the consideration of problems for which the mapping is not given analytically, numerical experiments on three-dimensional Navier-Stokes equations and analysis of the time-dependent Burgers and Navier-Stokes equations.

To extend the proposed methodology to other types of problems, a parabolic problem has been analysed next, namely the heat equation with random Robin boundary conditions. In addition to the perturbation technique and the finite element method for the stochastic and physical space approximations, respectively, an implicit time stepping scheme has been used for the time discretization. An *a posteriori* error estimate for the approximation of the first term in the expansion has been proposed and its efficiency has been investigated through two numerical examples. Application to problems of practical interest could be an interesting direction for a future work.

In the last part of this thesis, a residual-based *a posteriori* error estimate for the stochastic collocation finite element method has been proposed. The error estimator controlling the randomness in the problem has then been used to drive an adaptive sparse grid algorithm. Finally, promising preliminary numerical examples have been given that open the door to many thrilling perspectives, such as complexity analysis, comparison with other methods, combination with spatial mesh refinement or proof of convergence of adaptive scheme.

Bibliography

- [1] M. Ainsworth and J.T. Oden. A unified approach to a posteriori error estimation using element residual methods. *Numer. Math.*, 65(1):23–50, 1993.
- [2] M. Ainsworth and J.T. Oden. A posteriori error estimation in the finite element analysis. *Comput. Methods Appl. Mech. Engrg.*, 142(1-2):1–88, 1997.
- [3] M. Ainsworth and J.T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, 2000.
- [4] R.C. Almeida and J.T. Oden. Solution verification, goal-oriented adaptive methods for stochastic advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 199:2472–2486, 2010.
- [5] D.N. Arnold, F. Brezzi, and M. Fortin. A stable finite element for the Stokes equations. *Calcolo*, 21(4):337–344, 1984.
- [6] I. Babuška and P. Chatzipantelidis. On solving elliptic stochastic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 191(37-38):4093–4122, 2002.
- [7] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034, 2007.
- [8] I. Babuška and W.C. Rheinboldt. A posteriori error estimates for the finite element method. *Int. J. Numer. Methods Engrg.*, 12(10):1597–1615, 1978.
- [9] I. Babuška and W.C. Rheinboldt. A posteriori error analysis of finite element solutions for one-dimensional problems. *SIAM J. Numer. Anal.*, 18(3):565–589, 1981.
- [10] I. Babuška, R. Tempone, and G.E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825, 2004.
- [11] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. In S.J. Hesthaven and M.E. Rønquist, editors, *Spectral and High Order Methods for Partial*

Bibliography

- Differential Equations: Selected papers from the ICOSAHOM '09 conference, June 22-26, Trondheim, Norway*, volume 76 of *Lect. Notes Comput. Sci. Eng.*, pages 43–62. Springer, Berlin, 2011.
- [12] J.M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Arch. Rational Mech. Anal.*, 63(4):337–403, 1977.
- [13] W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2003.
- [14] R.E. Bank and R.K. Smith. A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.*, 30(4):921–935, 1993.
- [15] R.E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44(170):283–301, 1985.
- [16] S. Bartels, C. Carstensen, and G. Dolzmann. Inhomogeneous Dirichlet conditions in a priori and a posteriori finite element error analysis. *Numer. Math.*, 99(1):1–24, 2004.
- [17] A. Barth, C. Schwab C., and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.
- [18] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000.
- [19] H. Bateman. Some recent researches on the motion of fluids. *Mon. Weather Rev.*, 43(4):163–170, 1915.
- [20] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods. *Math. Models Methods Appl. Sci.*, 22(9):1250023–1–1250023–33, 2012.
- [21] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. Convergence of quasi-optimal stochastic Galerkin methods for a class of PDEs with random coefficients. *Comput. Math. Appl.*, 67(4):732–751, 2014.
- [22] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001.
- [23] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math.*, 85(4):579–608, 2000.
- [24] A. Bespalov, C.E. Powell, and D. Silvester. Energy norm a posteriori error estimation for parametric operator equations. *SIAM J. Sci. Comput.*, 36(2):A339–A363, 2014.
- [25] P. Binev, W. Dahmen, and R. DeVore. Adaptive Finite Element Methods with convergence rates. *Numer. Math.*, 97(2):219–268, 2004.

-
- [26] A. Bonito, R.A. DeVore, and R.H. Nochetto. Adaptive finite element methods for elliptic problems with discontinuous coefficients. *SIAM J. Numer. Anal.*, 51(6):3106–3134, 2013.
- [27] A. Bonito, I. Kyza, and R.H. Nochetto. Time-discrete higher-order ALE formulations: stability. *SIAM J. Numer. Anal.*, 51(1):577–604, 2013.
- [28] A.N. Boule. On the existence of the solution of Burgers’ equation for $n \leq 4$. *Internat. J. Math. & Math. Sci.*, 13(4):645–650, 1990.
- [29] M. Braack and P.B. Mucha. Directional do-nothing condition for the Navier-Stokes equations. *J. Comput. Math.*, 32(5):507–521, 2014.
- [30] C.A. Brebbia, J.C.F. Telles, and L.C. Wrobel. *Boundary Element Techniques: Theory and Applications in Engineering*. Springer, Berlin, 1984.
- [31] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods (3rd ed.)*, volume 15 of *Texts App. Math.* Springer, New York, 2008.
- [32] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO Anal. Numér.*, 8(R2):129–151, 1974.
- [33] C.M. Bryant, S. Prudhomme, and T. Wildey. A posteriori error control for partial differential equations with random data. ICES REPORT 13-08, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, 2013.
- [34] J.M. Burgers. A mathematical model illustrating the theory of turbulence. *Adv. Appl. Mech.*, 1:171–199, 1948.
- [35] T. Butler, C. Dawson, and T. Wildey. A posteriori error analysis of stochastic spectral methods. *SIAM J. Sci. Comput.*, 33(3):1267–1291, 2011.
- [36] A. Caboussat. *Analysis and numerical simulation of free surface flows*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2003.
- [37] D.G. Cacuci. *Sensitivity and Uncertainty Analysis, Volume 1: Theory*. Chapman & Hall/CRC, Boca Raton, 2003.
- [38] R.E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numer.*, 7:1–49, 1998.
- [39] C. Canuto and D. Franos. Numerical solution of partial differential equations in random domains: an application to wind engineering. *Commun. Comput. Phys.*, 5(2-4):515–531, 2009.
- [40] C. Canuto and T. Kozubek. A fictitious domain approach to the numerical solution of PDEs in stochastic domains. *Numer. Math.*, 107(2):257–293, 2007.
- [41] C. Carstensen and S. Funken. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput.*, 21(4):1465–1484, 2000.

Bibliography

- [42] J.M. Cascon, C. Kreuzer, R.H. Nochetto, and K.G. Siebert. Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal.*, 46(5):2524–2550, 2008.
- [43] J.E. Castrillon-Candas, F. Nobile, and R.F. Tempone. Analytic regularity and collocation approximation for PDEs with random domain deformations. *Comput. Math. Appl.*, 71(6):1173–1197, 2016.
- [44] J. Charrier. Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(1):216–246, 2012.
- [45] A. Chkifa. *Méthodes polynomiales parcimonieuses en grande dimension. Application aux EDP paramétriques*. PhD thesis, Laboratoire Jacques Louis Lions, 2014.
- [46] A. Chkifa, A. Cohen, R. DeVore, and C. Schwab. Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. *ESAIM: Math. Model. Numer. Anal.*, 47(01):253–280, 2013.
- [47] P.G. Ciarlet. *Mathematical Elasticity, Volume I: Three-Dimensional Elasticity*, volume 20 of *Stud. Math. Appl.* North-Holland, Amsterdam, 1988.
- [48] P.G. Ciarlet. Basic error estimates for elliptic problems. In P.G. Ciarlet and J.-L. Lions, editors, *Finite Element Methods (Part I)*, volume 2 of *Handb. Numer. Anal.*, pages 17–351. North-Holland, Amsterdam, 1991.
- [49] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.
- [50] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9(R2):77–84, 1975.
- [51] C.W. Clenshaw and A.R. Curtis. A method for numerical integration on an automatic computer. *Numer. Math.*, 2(1):197–205, 1960.
- [52] K.A. Cliffe, M.B. Giles, R. Scheichl, and A.L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Visual. Sci.*, 14(1):3–15, 2011.
- [53] M. Dauge. *Elliptic Boundary Value Problems on Corner Domains: Smoothness and Asymptotics of Solutions*, volume 1341 of *Lect. Notes Math.* Springer, Berlin, 1988.
- [54] J. Dick, F.Y. Kuo, and I.H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013.
- [55] J. Dick and F. Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, Cambridge, 2010.
- [56] J. Donea, A. Huerta, J.-Ph. Ponthot, and A. Rodríguez-Ferran. Arbitrary Lagrangian–Eulerian Methods. In *Encyclopedia of Computational Mechanics*, volume 1, chapter 14. John Wiley & Sons Ltd, 2004.

-
- [57] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.
 - [58] M. Eigel, C.J. Gittelsohn, C. Schwab, and E. Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.*, 270:247–269, 2014.
 - [59] M. Eigel, C.J. Gittelsohn, C. Schwab, and E. Zander. A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. *ESAIM: Math. Model. Numer. Anal.*, 49(5):1367–1398, 2015.
 - [60] H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite Elements and Fast Iterative Solvers: with application in incompressible fluid dynamics*. Numer. Math. Sci. Comput. Oxford University Press, 2005.
 - [61] A. Ern and J.-L. Guermond. *Theory and Practise of Finite Elements*, volume 159 of *Appl. Math. Sci.* Springer, New York, 2004.
 - [62] L.C. Evans. *Partial Differential Equations (2nd ed.)*, volume 19 of *Grad. Stud. Math.* American Mathematical Society, 2010.
 - [63] G.S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer Ser. Oper. Res. Financ. Eng. Springer, New York, 1996.
 - [64] P. Frauenfelder, C. Schwab, and R.A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2-5):205–228, 2005.
 - [65] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3):209–232, 1998.
 - [66] T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.
 - [67] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer, New York, 1991.
 - [68] M.B. Giles. Multi-level Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.
 - [69] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, volume 5 of *Springer Ser. Comput. Math.* Springer, Berlin, 1986.
 - [70] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *J. Comput. Phys.*, 230(10):3668–3694, 2011.
 - [71] C. Grandmont. Existence for three-dimensional steady state fluid-structure interaction problem. *J. Math. Fluid Mech.*, 4(1):76–94, 2002.
 - [72] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Society for Industrial and Applied Mathematics, 2011.

- [73] T. Grätsch and K.-J. Bathe. A posteriori error estimation techniques in practical finite element analysis. *Comput. Struct.*, 83(4):235–265, 2005.
- [74] D. Guignard, F. Nobile, and M. Picasso. A posteriori error estimation for elliptic partial differential equations with small uncertainties. *Numer. Methods Partial Differential Equations*, 32(1):175–212, 2016.
- [75] D. Guignard, F. Nobile, and M. Picasso. A posteriori error estimation for the steady Navier-Stokes equations in random domains. MATHICSE Technical Report 13.2016, Ecole Polytechnique Fédérale de Lausanne, 2016.
- [76] H. Harbrecht, M. Peters, and M. Siebenmorgen. Numerical solution of elliptic diffusion problems on random domains. *Preprint No. 2014-08*, 2014, revised on 06.02.2015.
- [77] H. Harbrecht, R. Schneider, and C. Schwab. Sparse second moment analysis for elliptic problems in stochastic domains. *Numer. Math.*, 109(3):385–414, 2008.
- [78] F. Hecht. New development in freefem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.
- [79] S. Heinrich. Multilevel Monte Carlo Methods. In S. Margenov, J. Waśniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing: Third International Conference, LSSC 2001 Sozopol, Bulgaria, June 6–10, 2001 Revised Papers*, volume 2179 of *Lecture Notes in Comput. Sci.*, pages 3624–3651. Springer, Berlin, 2001.
- [80] G. Jouvet. *Modélisation, analyse mathématique et simulation numérique de la dynamique des glaciers*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2010.
- [81] A.R. Khoei. *Extended Finite Element Method: Theory and Applications*. Wiley Ser. Comput. Mech. Wiley, 2015.
- [82] M. Kleiber and T.D. Hien. *The Stochastic Finite Element Method: Basic Perturbation Technique and Computer Implementation*. John Wiley & Sons Ltd, Chichester, 1992.
- [83] P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3):485–509, 1983.
- [84] O.A. Ladyzhenskaya. *The Boundary Value Problems of Mathematical Physics*, volume 49 of *Appl. Math. Sci.* Springer, New York, 1985.
- [85] R.J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts Appl. Math. Cambridge University Press, 2002.
- [86] G.R. Liu. *Meshfree Methods: Moving Beyond the Finite Element Method (2nd ed.)*. CRC Press, 2009.
- [87] M. Loève. *Probability Theory I (4th ed.)*, volume 45 of *Grad. Texts in Math.* Springer New York, 1977.

-
- [88] M. Loève. *Probability Theory II (4th ed.)*, volume 46 of *Grad. Texts in Math.* Springer New York, 1978.
- [89] G.J. Lord, C.E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge Texts Appl. Math. Cambridge University Press, 2014.
- [90] O.P. Le Maître and O.M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Sci. Comput. Springer Netherlands, 2010.
- [91] A. Manzoni. *Reduced Models for Optimal Control, Shape Optimization and Inverse Problems in Haemodynamics*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2012.
- [92] L. Mathelin and O. Le Maître. Dual-based a posteriori error estimate for stochastic finite element methods. *Comm. Appl. Math. and Comp. Sci.*, 2(1):83–115, 2007.
- [93] P. Morin, R.H. Nochetto, and K.G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488, 2000.
- [94] F. Nobile, L. Tamellini, and R. Tempone. Convergence of quasi-optimal sparse grid approximation of Hilbert-valued functions: application to random elliptic PDEs. *Numer. Math.*, 2015 (published online).
- [95] F. Nobile, L. Tamellini, F. Tesei, and R. Tempone. An adaptive sparse grid algorithm for elliptic PDEs with lognormal diffusion coefficient. In J. Garcke and D. Pflüger, editors, *Sparse Grids and Applications - Stuttgart 2014*, pages 191–220. Springer International Publishing, 2016.
- [96] F. Nobile, R. Tempone, and C.G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.
- [97] F. Nobile, R. Tempone, and C.G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.
- [98] F. Nobile and F. Tesei. A Multi Level Monte Carlo method with control variate for elliptic PDEs with log-normal coefficients. *Stoch. PDE: Anal. Comp.*, 3(3):398–444, 2015.
- [99] R.H. Nochetto, K.G. Siebert, and A. Veiser. Theory of adaptive finite element methods: An introduction. In R. DeVore and A. Kunoht, editors, *Multiscale, Nonlinear and Adaptive Approximation*, pages 409–542. Springer, Berlin, 2009.
- [100] J.T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Comput. Math. Appl.*, 41:735–756, 2001.

Bibliography

- [101] R.W. Ogden. *Non-linear Elastic Deformations*. Dover Civil and Mechanical Engineering. Dover Publications, 1997.
- [102] A.T. Patera and G. Rozza. *A Posteriori Error Estimation for Parametrized Partial Differential Equations*. Version 1.0, Copyright MIT 2006-2007, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2007.
- [103] M. Picasso. Adaptive finite elements for a linear parabolic problem. *Comput. Methods Appl. Mech. Engrg.*, 167(3-4):223–237, 1998.
- [104] M. Picasso, J. Rappaz, A. Reist, M. Funk, and H. Blatter. Numerical simulation of the motion of a two-dimensional glacier. *Int. J. Numer. Methods Engrg.*, 60(5):995–1009, 2004.
- [105] D.A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69 of *Math. Appl.* Springer, Berlin, 2012.
- [106] S. Prudhomme. *Adaptive control of error and stability of h-p approximations of the transient Navier-Stokes equations*. PhD thesis, The University of Texas at Austin, 1999.
- [107] S. Prudhomme, F. Nobile, L. Chamoin, , and J. T. Oden. Analysis of a subdomain-based error estimator for finite element approximations of elliptic problems. *Numer. Methods Partial Differential Equations*, 20(2):165–192, 2003.
- [108] A. Quarteroni and G. Rozza. Numerical solution of parametrized Navier-Stokes equations by reduced basis methods. *Numer. Methods Partial Differential Equations*, 23(4):923–948, 2007.
- [109] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23 of *Springer Ser. Comput. Math.* Springer, Berlin, 1994.
- [110] M. Schäfer, S. Turek (with support by F. Durst, E. Krause, and R. Rannacher). Benchmark computations of laminar flow around a cylinder. In E.H. Hirschel, editor, *Flow Simulation with High-Performance Computers II: DFG Priority Research Program Results 1993–1995.*, volume 48 of *Notes Numer. Fluid Mech.*, pages 547–566. Vieweg+Teubner Verlag, 1996.
- [111] C. Schwab and R.A. Todor. Karhunen-Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122, 2006.
- [112] G.D. Smith. *Numerical solution of partial differential equations: finite difference methods (3rd ed.)*. Oxford Appl. Math. Comput. Sci. Ser. Oxford University Press, 1985.
- [113] S.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Math. Dokl.*, 4:240–243, 1963. [Russian original in Dokl. Akad. Nauk SSSR, 148:1042–1045, 1963].

-
- [114] R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.
- [115] L. Tamellini. *Polynomial approximation of PDEs with stochastic coefficients*. PhD thesis, Politecnico di Milano, 2012.
- [116] R. Temam. *Navier-Stokes equations: Theory and Numerical Analysis*. North-Holland, 1977.
- [117] J.W. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods*, volume 22 of *Texts App. Math.* Springer, New York, 1995.
- [118] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, Chichester-Stuttgart, 1996.
- [119] R. Verfürth. *Adaptive Finite Element Methods*. Lecture Notes, Fakultät für Mathematik, Ruhr-Universität Bochum, Winter Term 2011/12, 2011.
- [120] S. Volkwein. *Mesh-independence of an augmented Lagrangian-SQP method in Hilbert spaces and control problems for the Burgers equation*. PhD thesis, TU Berlin, 1997.
- [121] X. Wang, S. Cen, and C. Li. Generalized Neumann expansion and its application in stochastic finite element methods. *Math. Probl. Eng.*, 2013, 2013.
- [122] G.W. Wasilkowski and H. Wozniakowski. Explicit cost bounds of algorithms for multivariate tensor product problems. *J. Complexity*, 11(1):1–56, 1995.
- [123] R.H.S. Winterton. Newton’s law of cooling. *Contemp. Phys.*, 40(3):205–212, 1999.
- [124] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.
- [125] D. Xiu and M. Tartakovsky. Numerical methods for differential equations in random domains. *SIAM J. Sci. Comput.*, 28(3):1167–1185, 2006.
- [126] D. Zhang and Z. Lu. A comparative study on uncertainty quantification for flow in randomly heterogeneous media using Monte Carlo simulations and conventional and KL-based moment-equation approaches. *SIAM J. Sci. Comput.*, 26(2):558–577, 2004.
- [127] D. Zhang and Z. Lu. An efficient, high-order perturbation approach for flow in random porous media via Karhunen-Loève and polynomial expansions. *J. Comput. Phys.*, 194(2):773–794, 2004.
- [128] O.C. Zienkiewicz and J.Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *Int. J. Numer. Methods Engrg.*, 24(2):337–357, 1987.

Curriculum Vitae

Personal Data

Name Diane Guignard
Date of birth January 14th, 1988
Nationality Swiss

Education

2012 – 2016 **PhD in Mathematics**

Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Thesis advisers: Prof. F. Nobile and Prof. M. Picasso.

2010 – 2012 **Master of Science in Applied Mathematics**

Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Master thesis at Caltech, Pasadena, USA under the supervision of Prof. T.Y. Hou.

2006 – 2010 **Bachelor of Science in Mathematics**

Ecole Polytechnique Fédérale de Lausanne.

Publications

1. D. Guignard, F. Nobile and M. Picasso. A posteriori error estimation for elliptic partial differential equations with small uncertainties. *Numer. Methods Partial Differential Equations*, 32(1): 175–212, 2016.
2. D. Guignard, F. Nobile and M. Picasso. A posteriori error estimation for the steady Navier-Stokes equations in random domains. MATHICSE Technical Report 13.2016 (submitted for publication).

Presentations

- 4th Workshop on Sparse Grids and Applications (Miami, Florida, USA, 4 October 2016)

Contributed talk: *A posteriori error estimate and adaptive sparse grid algorithm for random PDEs.*

- MATHICSE Retreat 2016 (Leysin, Switzerland, 28 June 2016)

Contributed talk: *A posteriori error estimation for PDEs with random input data.*

- ECCOMAS 2016 (Crète, Greece, 10 June 2016)

Talk in minisymposium: *A posteriori error estimate for the Navier-Stokes equations in random domains solved with a perturbation technique.*

- Colloque Numérique Suisse 2016 (Fribourg, Switzerland, 22 April 2016)

Talk: *A posteriori error estimation for the steady Navier-Stokes equations in random domains.*

- SIAM UQ 2016 (Lausanne, Switzerland, 8 April 2016)

Contributed talk: *A Posteriori Error Estimates for Navier-Stokes Equations with Small Uncertainties.*

- Reliable Methods of Mathematical Modeling (RMMM) 2015 (Zürich, Switzerland, 30 June 2015)

Talk in minisymposium: *A posteriori error estimation for elliptic partial differential equations with small uncertainties.*

- International Conference on Adaptive Modeling and Simulation (ADMOS) 2015 (Nantes, France, 9 June 2015)

Contributed talk: *A Posteriori Error Estimation for PDEs with Small Uncertainties.*

- Swiss Numerical Day 2015 (Geneva, Switzerland, 17 April 2015)

Poster: *A posteriori error estimation for partial differential equations with small uncertainties.*

- MATHICSE Retreat 2014 (Leysin, Switzerland, 12 June 2014)

Contributed talk: *A posteriori error estimation for PDEs with small uncertainties using perturbation methods.*

